CAS MONOGRAPH SERIES
NUMBER 13

PENALIZED REGRESSION AND LASSO CREDIBILITY

Thomas Holmes, FCAS
Mattia Casotto

CASUALTY ACTUARIAL SOCIETY

PENALIZED REGRESSION AND LASSO CREDIBILITY

Thomas Holmes, FCAS
Mattia Casotto

Casualty Actuarial Society
4350 North Fairfax Drive, Suite 250
Arlington, VA 22203
www.casact.org
(703) 276-3100

Penalized Regression and Lasso Credibility
By Thomas Holmes and Mattia Casotto

Library of Congress Cataloging-in-Publication Data
Penalized Regression and Lasso Credibility /Thomas Holmes and Mattia Casotto
ISBN (print edition) 978-1-7333294-3-9
ISBN (electronic edition) 978-1-7333294-4-6
1. Actuarial science. 2. Classification ratemaking. 3. Insurance—mathematical models
I. Holmes, Thomas. II. Casotto, Mattia

Table of Contents

CAS Monograph Editorial Board

Acknowledgments

The success of this monograph owes much to the invaluable contributions of several individuals and organizations. We are particularly grateful to Arun Kadavankandy, Giovanni Frigeri, and Waldemar Schulgin, whose thoughtful feedback and contributions of visualizations made the content more engaging and less reliant on dense notation. We would also like to extend our thanks to the insurance software company Akur8, whose support was instrumental in providing the time needed to develop and share the methodologies detailed in this book.

Special appreciation goes to Brandon Smith, who expertly guided us in tailoring this monograph to meet the specific needs of an actuarial audience. We are also thankful for the thorough and insightful feedback from anonymous reviewers, as it was crucial in enhancing the work from initial drafts to its current form. Lastly, we would like to recognize the Casualty Actuarial Society for the platform to present our findings, making this endeavor possible.

Thomas Holmes would like to thank his wife, Liz, and his friends and family for their kind support and patience as they heard the words "We're almost done with the paper" for at least a year. He would also like to thank his soon-to-be-born daughter for providing strong motivation to finish the monograph before her arrival.

Mattia Casotto would like to extend his heartfelt gratitude to his wife, Maria, whose insightful suggestions and project management skills were instrumental in navigating the ups and downs of writing the monograph. Her support and infinite patience provided the motivation necessary to bring this work to completion.

About the Authors

Thomas Holmes is Akur8's Chief Actuary for the US region and received his FCAS in 2019. He has experience with actuarial modeling for personal and commercial insurance, and is a frequent presenter at CAS events and Akur8 Academy webinars. Additionally, he volunteers with the CAS on predictive modeling topics and performs industry outreach to share actuarial modeling methodologies and best practices. Thomas holds music degrees from the University of Michigan and Ohio University, and enjoys playing the piano and writing music in his spare time.

Mattia Casotto is Akur8's US Head of Product and Principal Scientist. He is the co-author of various works such as the research papers "Derivative Lasso" and "Credibility and Penalized Regression." With more than nine years of experience in predictive modeling in insurance, Mattia was one of the original team members who started the pricing software Akur8. He holds two master's degrees, one in Mathematics and one in Quantitative Finance, and has a passion for transparent machine learning and music.

Introduction

Predictive modeling has a number of operational applications in the insurance industry, and actuaries have access to a generous tool kit of modeling techniques to best address the various use cases. Among those modeling techniques, generalized linear models (GLMs) are a common choice for frequency, severity, and pure premium loss modeling.

The unpenalized GLM approach comes with one well-documented shortcoming: while minimum bias and univariate techniques can incorporate credibility in the calculation of indicated rating relativities, there has been no statistically straightforward, consistent way of incorporating actuarial credibility into a GLM. The Casualty Actuarial and Statistical (C) Task Force described this shortcoming as follows: "GLMs effectively assume that the underlying datasets are 100% credible, no matter their size. If some segments have little data, the resulting uncertainty would not be reflected in the GLM parameter estimates themselves (although it might be reflected in the standard errors, confidence intervals, etc.)" (2020, 4).

GLM output can warn a user of instability in a parameter estimate through a wide standard error, but it does not adjust the coefficient to take the large volatility into account. Instead, the practitioner may perform ad hoc adjustments to consider the lack of credibility or volatility in a specific segment. Post-modeling adjustments performed in this necessarily univariate manner may result in a suboptimal final rating plan.

Fortunately, this issue can be addressed by applying an enhanced version of a GLM: penalized regression. Penalized regression has similarities with credibility procedures as documented in Miller (2015) and Casotto, Banterle, and Beraud-Sudreau (2020). In this monograph we review GLMs and then introduce penalized regression and its connections to credibility. We describe the motivation for the technique as well as why lasso is our preferred penalization for actuarial analysis. Furthermore, we show how lasso penalization can be used as **lasso credibility** through a new use of the offset. We are not creating a new form of modeling from scratch, but rather combining existing tools to create an actuarially sound credibility procedure.

Using penalized regression for credibility has significant implications in actuarial analysis. With lasso credibility, the amount of data needed for predictive modeling is reduced. This means actuaries can use smaller data sets that might be insufficient for a stable GLM. Also, when we use lasso penalization as lasso credibility, the usage must now align with the guidelines of Actuarial Standard of Practice No. 25, *Credibility Procedures* (hereafter referred to as ASOP 25). We explain how the guidance in ASOP 25 applies to penalized regression as a credibility procedure.

This perspective shift to credibility should affect the way actuaries interpret and evaluate lasso credibility modeling results. For example, consider the question of whether one should include a specific factor, or variable, in a model. Modelers using a GLM may rely on a *p*-value analysis, which can provide a measure of whether the data is compatible with the absence of such a factor (the null hypothesis) (Wasserstein and Lazar 2016).

We show that whereas *p*-values answer the question of **significance**—*Is this significantly different than zero (or a more extreme value)?*—lasso penalized regression answers a question of **credibility**—*How much credibility, if any, should we give to this coefficient?* Lasso credibility answers a similar question—*How much credibility, if any, should we give to this coefficient's deviation from our complement?* By using lasso penalization or lasso credibility, an actuary can simultaneously evaluate the significance and magnitude of a coefficient, while an unpenalized GLM's *p*-values will evaluate only significance. Users of lasso credibility will need to let go of the idea of *p*-values and embrace a credibility interpretation of coefficients to correctly apply and evaluate the methodology. Furthermore, as we will see, the tuning of the penalty parameter (which acts as a credibility parameter) is both simpler and more robust than the examination of *p*-values.

The main body of the paper is written to provide a reader with the intuition behind penalized regression as a credibility procedure and practical guidance on how to implement lasso credibility. We provide a minimum necessary background for these approaches, reserving statistical proofs and rigorous defense of the concepts for a series of appendices. We hope that this structure allows the communication of these ideas to a broad audience without a lack of precision or loss of statistical rigor.

We supplement the guidance with a case study comparing lasso credibility, penalized regression, and traditional GLM models on data sets of varying size. The case study shows that lasso credibility can outperform both GLM and penalized regression when the model is informed by an adequate complement of credibility. This section has accompanying code on the CAS GitHub,[1] and we highly encourage readers to pull the code and run it alongside the case study. Additionally, we provide optional exercises to familiarize yourself with the behavior of lasso credibility in alternate scenarios.

[1] https://github.com/casact/mg-credibility

1. A Review of GLMs

> Generalized linear models (GLMs) are a means of modeling the relationship between a variable whose outcome we wish to predict and one or more explanatory variables.
>
> —Goldburd et al. (2016), *Generalized Linear Models for Insurance Rating*

In this chapter, we review the basics of GLMs to provide the minimum background necessary to introduce penalized regression and lasso credibility. First, we explore the linearity of GLMs. That foundation will help us demonstrate how to incorporate a complement of credibility into the GLM framework. Then, we describe the link function and give an example of how link functions allow a modeler to easily use a GLM's output as a multiplicative rating table. We revisit this example later in the paper to show how to implement the output of a lasso credibility model as an adjustment to an existing set of rating tables. Finally, we discuss the full credibility assumption of GLMs and how it allows for the use of p-values to evaluate coefficients. This prepares us for a discussion on how penalized regression does not assume full credibility, and therefore p-values are not appropriate in penalized regression and should be replaced by another type of model evaluation.

1.1. Definitions and Terminology

We begin with a short introduction to GLMs. For a comprehensive introduction, we refer the reader to Goldburd et al. (2016).

A GLM consists of three elements:[2]

1. A target variable Y, a random variable following a probability distribution from the exponential family, which is in turn defined by a selected variance function and dispersion parameter.
2. A linear predictor $\eta = X\beta$, where X is the design matrix and β is the coefficient vector.
3. A monotonic link function g such that $E(Y) = \mu = g^{-1}(\eta)$.

These elements have established connections to common insurance concepts.

Y represents the random variable that models the risk and y_i represents the actual observed risk for row i. Depending on the nature of the risk, one makes different statistical

[2] The GLM definition is taken from Casualty Actuarial and Statistical (C) Task Force (2020) with some minor changes.

assumptions. For example, the actuary may model accident frequency via the Poisson probability distribution and accident severity via a gamma distribution.

The matrix X is such that its rows X_i contain the information about each record and any covariate relevant for predicting the considered risk. Typically, X_i represents a unit of risk specified by the modeler for each row i. For example, one exposure may represent a year of observation or a single policy. The columns of the matrix provide a numerical representation of the available information on the risk covariates. The values of the coefficients β define how the covariates are linearly combined to estimate the risk. Using X_i and β, we can represent the linear combination of the covariates as $\eta_i = X_i\beta$.

Finally, μ_i represents the expected risk estimate for each row i. The linear combination of the features $X_i\beta$ are related to the expected risk via the link function $\mu_i = g^{-1}(X_i\beta)$. In this notation, the intercept β_0 is implicit. The specific choice of the link function depends on the target probability distribution chosen during modeling. For example, for Poisson and gamma distributions, the preferred link function is the logarithm, which gives a multiplicative model.

The process of building (or fitting) a GLM requires the specification of the target variable Y and its statistical assumptions together with the covariates X. The output of the fitting procedure is a set of coefficients β that maximizes the likelihood of observing Y with expected mean μ_i and the assumed target distribution given the data X. This process of fitting coefficients based on the observed likelihood is at the core of why "GLMs effectively assume that the underlying datasets are 100% credible, no matter their size." We now explore these elements a bit more deeply.

1.2. The Linear Predictor

GLMs are called generalized *linear* models because the relationship between the predictor variables X and the expected risk μ_i is determined by the **linear** predictor η. The formula for this relationship is

$$\eta = \beta_0 + \beta_1 X_1 \ldots \beta_n X_n = \beta_0 + X\beta. \tag{1.1}$$

For a change in each individual characteristic X_i (holding all other X constant), there is a linear change in the value of η. For instance, for each integer increase in X_1, the linear predictor η increases by the value of β_1. This is not to say that the prediction is always linear with respect to the underlying risk characteristic. For example, when using the predictor "age of vehicle squared," the relationship will be linear with respect to "age of vehicle squared," and therefore quadratic with respect to the risk characteristic "age of vehicle." It is quite common to include multiple polynomial terms (linear, squared, cubed) in a model for a single risk characteristic. The creation of these variable transformations, often referred to as **feature engineering,** is an essential part of fitting a GLM. The creation and inclusion of polynomial terms is an example of feature engineering for a continuous variable. To encode categorical variables, one can introduce

dummy variables that take the values of either 1 or 0 to represent the presence or absence of a certain predictor value. This is referred to as one-hot encoding in the machine learning literature.

1.3. Distributions and Link Functions

The link function is the relationship between the linear predictor η in Equation 1.1 and the predicted value μ. If we use no link function, then we are using η to directly predict μ. By using the log link function, we can predict the log of μ instead:

$$ln(\mu) = \beta_0 + \beta_1 X_1 \ldots \beta_n X_n,$$

or equivalently

$$\begin{aligned}\mu &= \exp(\beta_0 + \beta_1 X_1 \ldots \beta_n X_n) \\ &= \exp(\beta_0) \times \exp(\beta_1 X_1) \times \ldots \times \exp(\beta_n X_n).\end{aligned}$$

By using this link function, the modeled components are now combined multiplicatively to create a predicted expected value. This is ideal for actuarial pricing, as many rating plans are a combination of multiplicative rating tables. Additionally, the link function allows for the use of different error distributions.

Since the linear predictor η can potentially take any value, the correct choice of the link function is key in GLM modeling. The inverse of the link function determines the expected mean μ—hence the link must be chosen such that η is mapped to the correct range of values. For example, for the gamma and Poisson distributions, the expected mean must be positive. Hence the log link is appropriate for those distributions as the inverse of the log link is the exponential function, which is always positive. To note another example, the mean of a binomial (logistic) variable must be between 0 and 1, and hence the logit is one of the appropriate link functions for the Bernoulli distribution, as its inverse, the logistic function, maps all values in the range from 0 to 1.

Table 1.1 shows some commonly used distributions and link functions for actuarial models. Other link functions may be used for some of these distributions, but the ones listed are the most common in actuarial applications (Goldburd et al. 2016).

Table 1.1. Commonly Used Distributions in Actuarial Modeling

Model Type	Distribution	Link	Inverse Link
Frequency	Poisson, negative binomial	Log	Exp
Severity	Gamma, inverse Gaussian	Log	Exp
Pure premium	Tweedie	Log	Exp
Propensity, retention, conversion	Bernoulli	Logit	Logistic

1.4. The Offset

When building a model, we may want to consider the effect of risk characteristics without coming up with a prediction for them. Deductibles, for example, are best priced through a loss elimination ratio analysis rather than a GLM (Goldburd et al. 2016). Instead of modeling or ignoring these risk characteristics, they can be included as an **offset** in our GLM. The offset term is an additional item in our linear equation. Consider a GLM with a logarithmic link. The formula for the offset is given by

$$ln(\mu) = \beta_0 + \beta_1 X_1 \ldots \beta_n X_n + \text{offset}.$$

Assuming we are offsetting a deductible characteristic, the offset would be a column in our data set representing the coefficient for the surcharge or discount at the record's deductible level. A GLM will then directly include this coefficient in its prediction of μ when fitting the optimal values of β. Multiple risk characteristics—e.g., deductible factors, increased limit factors, territory relativities, etc.—can be included in a single offset term.

1.5. Table-Based Output: An Example

Let's create a two-variable pricing model for home insurance as an example. We will use a Tweedie distribution and a log link to model pure premium directly. The first variable is the presence of a fire extinguisher, encoded as 1 without a fire extinguisher or 0 with a fire extinguisher. This 1 or 0 value would be represented by X_1. The second variable will be age of home, encoded as the integers 0–10. The appropriate integer per record would be represented as X_2:

$$\hat{\mu} = \exp(\beta_0) \times \exp(\beta_1 X_1) \times \exp(\beta_2 X_2).$$

Let's assume that our model fit these convenient values:

$$\beta_0 = \log_e(100) \approx 4.605$$

$$\beta_1 = \log_e(1.2) \approx 0.182$$

$$\beta_2 = \log_e(1.01) \approx 0.01$$

The predicted pure premium would then be calculated as follows:

$$\begin{aligned}\text{Pure premium} &= \exp(\beta_0) \times \exp(\beta_1 X_1) \times \exp(\beta_2 X_2) \\ &= \exp(4.6057) \times \exp(0.182 \times X_1) \times \exp(0.01 \times X_2) \\ &= 100 \times 1.2^{X_1} \times 1.01^{X_2}.\end{aligned}$$

We can represent this model in the following rating tables:
Base rate: exp(4.6057) = 100.

Fire Extinguishers	Factor
No	$\exp(0.182 \times 1) = 1.200$
Yes	$\exp(0.182 \times 0) = 1.000$

Age of Home	Factor
0	$\exp(0.01 \times 0) = 1.000$
1	$\exp(0.01 \times 1) = 1.010$
2	$\exp(0.01 \times 2) = 1.020$
3	$\exp(0.01 \times 3) = 1.030$
4	$\exp(0.01 \times 4) = 1.041$
5	$\exp(0.01 \times 5) = 1.051$
6	$\exp(0.01 \times 6) = 1.062$
7	$\exp(0.01 \times 7) = 1.072$
8	$\exp(0.01 \times 8) = 1.083$
9	$\exp(0.01 \times 9) = 1.094$
10	$\exp(0.01 \times 10) = 1.105$

This easy translation from beta coefficients to rating tables is one of the many reasons that GLMs have been used in actuarial pricing for quite some time.

1.6. Likelihood Optimization: Full Credibility Assumption

The full credibility assumption of GLMs is related to the optimization process used when fitting the model's β parameters. The procedure for computing the GLM parameters β is via the maximization of the log-likelihood (or equivalently, minimization of the negative of the log-likelihood) of observing y_i under the assumption that they follow the chosen error distribution with mean being μ_i.

$$\begin{aligned}\hat{\beta}_{\text{GLM}} &= \underset{\beta}{\operatorname{argmax}}\, \text{LogLikelihood}\left(y, X, \beta\right) \\ &= \underset{\beta}{\operatorname{argmin}} - \text{LogLikelihood}\left(y, X, \beta\right). \end{aligned} \tag{1.2}$$

The maximization of likelihood alone will always treat the data as fully credible, and will not consider volatility in the estimates of β_p or the significance of the improvement each β_p might add to the overall model. In short, **GLM estimates are unstable on segments with low exposures**.

Section A.1 provides a motivation and proof for this statement. The implications of this behavior can be shown using our two-variable example model.

Assume that the likelihood maximization process determines that a surcharge of 20% is the "most likely" estimate of the true surcharge. The implications of this result may vary wildly depending on the data. We examine three scenarios where a single-variable GLM would output such a surcharge as the "most likely" estimate.

Scenario 1: Credible and sound estimate

Category	Exposures	Average Loss
With fire extinguisher	1,000,000	100
Without fire extinguisher	1,000,000	120

In this scenario, a β representing a 20% surcharge will be output as the most likely value of the surcharge. The estimate assigns full credibility to both categories in the data. An actuary would likely be confident in implementing this surcharge.

Scenario 2: Midway estimate

Category	Exposures	Average Loss
With fire extinguisher	1,000,000	100
Without fire extinguisher	5,000	120

In this scenario, a β representing a 20% surcharge will be output as the most likely value of the surcharge. The estimate assigns full credibility to both categories in the data. An actuary may believe that there is some signal to the true surcharge but may not be fully confident in the point estimate.

Scenario 3: Noncredible estimate

Category	Exposures	Average Loss
With fire extinguisher	1,000,000	100
Without fire extinguisher	10	120

In the third scenario, a β representing a 20% surcharge will be output as the most likely value of the true surcharge. The estimate, again, assigns full credibility to both categories in the data. But whereas a 20% surcharge is the "most likely" result given our limited data, we would not be fully confident in the estimate.

In a univariate analysis, an actuary might use a credibility procedure to determine the appropriate surcharge. However, in a multivariate setting, the question remains open because unpenalized GLMs do not incorporate credibility.

Although the modeler has no control over the **estimate** of the model, it is still possible to decide whether a factor needs to be included at all in the model.

That evaluation is binary due to the structure of a GLM: should we include this variable at full credibility, or should we exclude it entirely? The most common method of answering this question is the evaluation of *p*-values.

1.6.1. P-Values

In statistical hypothesis testing, *p*-values are a commonly used tool to accept or reject a null hypothesis against an alternative hypothesis. The aim of statistical hypothesis testing is to decide whether the data provides sufficient evidence against the null hypothesis, in which case this hypothesis is rejected in favor of the alternative hypothesis. *P*-values control the error of rejecting the null hypothesis when the null hypothesis is true and the experiment is repeated an infinite number of times. In GLM modeling we can apply this principle and use *p*-values to decide whether a given coefficient is significant. The null hypothesis being tested is then that a given coefficient's true value is zero or more extreme, i.e., the coefficient is not significant. Thus, the ***p*-value** of a coefficient represents the probability that a coefficient result at least as extreme as the estimate could have happened assuming that the null hypothesis (often assuming one parameter β_j is 0 or more extreme) is true.

When a coefficient's *p*-value is below 0.05, this means that there is less than a 5% chance that the observed results could have happened if the true value of the coefficient was zero and the experiment was repeated on different data an infinite number of times. In other words, the probability that the coefficient is due purely to randomness in the data is less than 5%.

A *p*-value of 0.05 is commonly used as a threshold for coefficient significance. When a coefficient's *p*-value is equal to or greater than the selected threshold, the coefficient is deemed **insignificant** and we cannot reject the null hypothesis that the true coefficient is zero. In this case, the coefficient is usually removed from a model. When a coefficient's *p*-value is less than 0.05, we reject the null hypothesis that the coefficient is zero and the coefficient is considered **significant**. When a coefficient is significant, it is included in the model and given full credibility.

P-values and significance testing have several limitations:

1. Significance testing is a binary test that answers only the question "Is this coefficient likely not zero or more extreme?"
2. Although GLM output provides tools such as confidence intervals to evaluate coefficient stability, it does not provide statistical guidance on how to make corresponding adjustments. For example, when using a *p*-value threshold of 0.05, how should an actuary treat a coefficient with a *p*-value of 0.047? How much should an actuary trust the coefficient of a variable accepted by the actuarial and regulatory community as a predictor of loss that has a reasonable value and a *p*-value of 0.06? Such decisions are purely judgmental.
3. The traditional 0.05 level of significance is arbitrary. Numerous authors say that it is not an appropriate threshold for many studies (Wasserstein and Lazar 2016). Other scholars suggest that significance testing should be removed altogether.

4. Significance testing is iterative due to its post hoc application. The addition or removal of coefficients may affect the significance of other coefficients.
5. Nonbinary adjustments for questionable *p*-values must be made after modeling. *These actuarial selections are frequently made on a univariate basis and are therefore often suboptimal and contrary to the multivariate nature of a GLM's structure.*

Additional misconceptions and limitations are detailed in Greenland et al. (2016). Despite these downsides, *p*-values are widely used to evaluate GLM coefficients because they are a convenient and simple metric with which to perform significance testing. However, as we will later see, *p*-value significance testing is inappropriate (and in some cases not even possible) when using penalized regression. Penalized regression's ability to evaluate and adjust coefficients during the modeling process eliminates the need for post hoc significance testing.

1.6.2. Lack of Credibility in GLM Estimates

A consequence of the lack of credibility considerations during the fitting process is that a modeler must perform **post hoc** procedures on the coefficients of a model if one or more of the coefficients are unreasonable.

There are two kinds of post hoc analyses:

1. An analysis informing a subsequent model iteration
2. An analysis informing selections from final modeled coefficients

When the first post hoc analysis is incorporated back into the model, it must be a binary application. Either the coefficient should be included and receive full credibility or it should be excluded and receive no credibility—GLMs do not have another option. As described earlier, *p*-value significance testing is one appropriate methodology to arrive at this binary recommendation.

A common post hoc analysis that informs selections from final modeled coefficients is the credibility procedure. Unfortunately, such a credibility procedure must necessarily be done on a variable-by-variable basis as we are examining and adjusting one coefficient at a time. As we pointed out before, these adjustments may result in a suboptimal final model as they do not reflect the multivariate structure of the GLM.

To summarize, a coefficient can either (a) be assigned partial credibility during a post hoc univariate analysis or (b) receive full credibility or be removed during the multivariate fitting process. As we will see, penalization solves this dilemma by allowing a coefficient to be assigned partial credibility in a multivariate fitting procedure.

2. A Brief Review of Credibility

> Credibility, simply put, is the weighting together of different estimates to come up with a combined estimate.
>
> —*Foundations of Casualty Actuarial Science*

In the context of ratemaking, credibility provides a framework with which to combine an estimate based on observed experience (observed losses, frequencies, or loss ratios), subject to volatility, with a more stable yet less individualized estimate—the complement of credibility. This combination aims to improve on both estimates to create better predictions of future values.

The estimates are blended together via the credibility factor, normally referred to as Z, a factor between 0 and 1 that will give more or less weight to the observed experienced or the complement of credibility:

$$\text{Estimate} = Z \times \text{Observed Experience} + \left(1 - Z\right) \times \text{Complement of Credibility}.$$

Two main types of credibility are found in the literature: **classical** and **Bühlmann.** Even if they differ in terms of the underlying hypothesis and formulation of the factor Z (see Table 2.1), they share the same basic credibility property: the credibility factor increases with the number of observations n (i.e., the exposure). In that sense, unlike simple GLMs, the credibility framework enables a user to incorporate information on the number of observations directly into the estimates.

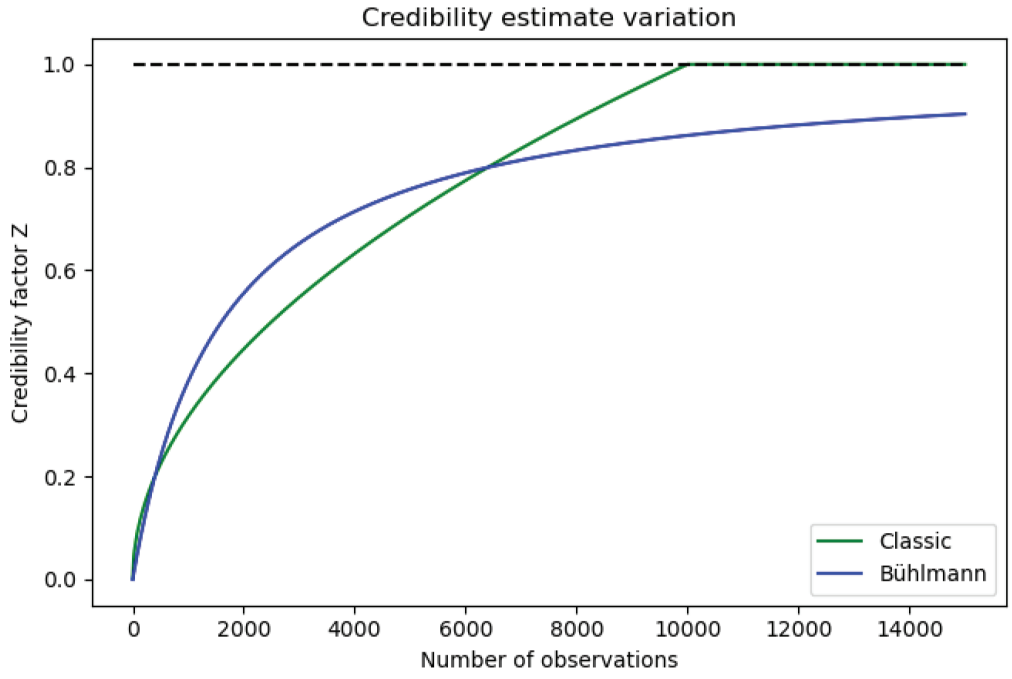
2.1. Incorporation of Credibility into GLM Estimates

There are, of course, simplistic ways to adjust GLM estimates with these traditional credibility methodologies, but they all share the drawbacks highlighted in Klinker (2011): "Some actuaries have been known to apply an ad hoc credibility adjustment to coefficients output by a GLM. In some cases this even produces results similar to those arrived at by more statistically rigorous methods. If so, then what is so wrong with the ad hoc credibility adjustment of GLM output? . . . This gets back to the old issue that a sequence of steps, each optimal individually, may not be optimal in the aggregate" (1–2).

Table 2.1. Main Parameters of Classical and Bühlmann Credibility

Classical Credibility	Bühlmann Credibility
$Z = \min\left(\sqrt{\frac{N}{N_{full}}}, 1\right)$ (2.1)	$Z = \frac{n}{n+k}$ (2.2)
Additional parameters:	Additional parameters:
$N_{full} = N_{full}(K, P)$ — number observations to reach full credibility	k — ratio of $\sigma^2_{PV}/\tau^2_{HM}$, with
P — probability that the observations are within estimated risk	σ^2_{PV} — expected process variance (within class variance)
K — tolerance to error, as % of risk	τ^2_{HM} — variance of hypothetical means (between class variance)

Figure 2.1. Evolution of the credibility factor *Z* for a given estimate *j* as a function of the number of observations *n*. The *Z* for the classical credibility is computed using Equation 2.1 with *N* = 15,000. Bühlmann credibility uses *k* = 1,600 in Equation 2.2.



If the simplistic ways are insufficient, then how can we best incorporate credibility into model fitting? To obtain a statistically rigorous multivariate modeling technique that can incorporate credibility, at least three necessary properties must be satisfied:

1. The estimation of the parameters shall not rely on maximizing the log-likelihood (or variance) alone: any technique with this property will inevitably assign 100% credibility to the data.
2. Estimates will be shrunk toward the complement of credibility GLMsand the amount of shrinkage will depend on the number of observations.
3. To consider correlations, the "credibility weighting" of the coefficients must be a part of the fitting procedure, not a post-processing step on top of a GLM.

Penalized regression satisfies these three criteria when applied in a specific manner. Before we describe the application of penalized regression as credibility, we introduce penalized regression as a modeling technique and contrast it with a GLM.

3. Penalized Regression

We introduce penalized regression first via the general penalization formula so that we can then focus on the three most popular penalization methods: lasso, ridge, and elastic net. After highlighting the differences with and similarities to unpenalized GLMs, we describe guidelines to use when selecting and analyzing the result of a penalized model. Finally, we explain the rationale behind using the **lasso** penalty for actuarial analysis owing to its unique ability to create a **sparse** and parsimonious model.

3.1. Types of Penalized Regression

In Section 1.6, we saw that GLM estimates always attribute 100% credibility to the data, regardless of the underlying exposure. The reason lies in maximizing the likelihood formula, which targets the goodness of the fit alone as described in Section A.1.

Penalized regression slightly modifies the likelihood formula by adding a **penalty** term to the GLM optimization, thereby adding a credibility component to the cost function that the unpenalized GLM was missing. Penalized regression **jointly** optimizes the trade-off between

- goodness of fit (likelihood) and
- prior assumptions on the shape of the coefficients (penalty).

When we design the penalty, we can design it to favor models with desirable properties, such as a low number of parameters. This effectively adds a credibility component to the cost function and regularizes the likelihood.

The following is the general formula for penalized regression:[3]

$$\begin{aligned}\hat{\beta} &= \underset{\beta}{\operatorname{argmin}} - \text{LogLikelihood}\left(y, X, \beta\right) + \lambda\ \text{Penalty}\left(\beta\right) \\ &= \underset{\beta}{\operatorname{argmin}}\ \text{NLL}\left(y, X, \beta\right) + \lambda\ \text{Penalty}\left(\beta\right). \end{aligned} \tag{3.1}$$

In the formula, $\lambda \geq 0$ is a positive number, referred to as the **penalty parameter.** The penalty parameter λ plays the role of a dial that assigns more or less importance

[3] Different formulations of the penalized regression can be found in the literature and in open-source solvers. For example, often the negative log-likelihood (NLL) is normalized by the number of observations n. In all cases, the formulas are equivalent after a reparameterization of the parameter (for example $\lambda \to \frac{\lambda}{n}$).

to the goodness of fit of the model in the training database or to the prior structure of the coefficients.

Penalized regression is a well-established technique in machine learning with a massive amount of accompanying research and literature. This section gives only a very basic introduction to the topic, and we refer readers to Hastie, Tibshirani, and Friedman (2009) and van Wieringen (2015) for a more in-depth (and mathematically dense) treatment of the subject. Wüthrich and Merz's (2023) book and the final chapter of Goldburd et al. (2016) discuss penalized regression from an actuarial perspective.

There are three main, established types of penalties:

- The **ridge** penalty is given by the sum of squares (or l2 squared norm) of β, that is,

$$\text{Ridge}(\beta) = \frac{1}{2}\sum_{j=1}^{p} \beta_j^2 = \frac{1}{2}|\beta|_2^2.$$

- The **lasso** penalty is given by the sum of the absolute values (or l1 norm) of β, that is,

$$\text{Lasso}(\beta) = \sum_{j=1}^{p} |\beta_j| = |\beta|_1.$$

- The **elastic net** penalty is a blend of both the ridge and the lasso penalties, linearly combined via a user-defined parameter $0 < \alpha < 1$:

$$\text{Elastic Net}_\alpha(\beta) = \frac{1-\alpha}{2}\sum_j \beta_j^2 + \alpha\sum_j |\beta_j|.$$

Penalized regression techniques such as the lasso and the ridge are at their core very similar to GLMs in their mathematical specification (Goldburd et al. 2016) in that they preserve the same underlying structure:

- The estimated parameter is related to a **linear combination** of the explanatory variables by a **link function** (Section 1.2).
- Observations are assumed to follow a **distribution** around the estimated parameter (Section 1.3).
- The **table-based output** (Section 1.5) and the variable parametrizations are preserved.

The primary difference is given by the choice of the penalty (lasso or ridge) and the value of the penalty parameter λ. Those choices will determine the difference between the estimates of the parameters β of a penalized model and those of the unpenalized GLM.

3.1.1. Role of Penalty Parameter λ

The parameter lambda $\lambda \geq 0$ is of fundamental importance in the penalized regression framework.

When $\lambda = 0$, the penalty term is removed and the resulting **model is an unpenalized GLM.** In this case, the coefficients can be quite noisy as they fully react to the data. When λ is sufficiently large, the penalty term in Equation 3.1 gains increasingly more

importance. Since the aforementioned penalties are designed to regularize or shrink the coefficients toward zero, a large penalty term leads to a solution whose coefficients will be either zero or a negligibly small number depending on the type of penalization.

For a fixed parameter λ, as the β coefficients move away from zero, the value of the penalty term increases.

In these scenarios, the exact behavior of the coefficients is decided by the type of penalization, as we'll detail in the next sections.

The penalty parameter is sensitive to the parameterization of the feature matrix X: for a given value of λ, the result will differ if the same quantity is, for example, expressed as miles or in kilometers. For this reason (among others), it is best practice to automatically standardize the features[4] before solving Equation 3.1.

Most statistical software supporting penalized regression, both proprietary and open source (glmnet in R or scikit-learn in Python), provide ways to automate the different steps in the computation: automatically standardize the coefficients, fit the penalized model, and return the unstandardized coefficients.

3.1.2. Ridge Regression

The formula for ridge regression adds the ridge penalty term to the GLM likelihood optimization:

$$\hat{\beta}_{\text{Ridge}} = \underset{\beta}{\operatorname{argmin}} \operatorname{NLL}\left(y, X, \beta\right) + \lambda\left(\frac{1}{2}\sum\nolimits_i \beta_i^2\right).$$

The formulation of the ridge penalty can be traced back to Hoerl (1962). The wide adoption of ridge regression is a consequence of its ability to provide stable estimates with highly correlated variables. GLMs are known to become unstable in the presence of highly correlated variables due to aliasing (see Section C.2). The ridge penalty term provides protection against the coefficients "blowing up" as they might in a GLM.

Figure 3.1 shows the differences in the GLM output coefficients between various ridge models with a varying λ.

We observe that the greater the penalty, the more the coefficient tends to be shrunk toward zero, compared with its unpenalized GLM counterpart ($\lambda = 0$). Furthermore, the amount of shrinkage depends on the underlying amount of exposure: the lower the exposure, the higher the magnitude of the shrinkage as the penalty changes. This property matches the behavior of Bühlmann's credibility method. In Appendix A we prove that under some underlying hypotheses, ridge is a multivariate transposition of the Bühlmann credibility, where Bühlmann's K has a one-to-one correspondence with ridge's λ parameter.

We can get a more holistic representation of the relationship between the penalty and the models' parameters by visualizing the **coefficient path**. Figure 3.2 shows the

[4] Standardization involves rescaling the features (or columns) of the data set X so that they have a mean of 0 and a standard deviation of 1.

Figure 3.1. Comparison of an unpenalized GLM and a GLM with various, increasing levels of ridge penalty. Simple GLM corresponds to the fit with $\lambda = 0$.

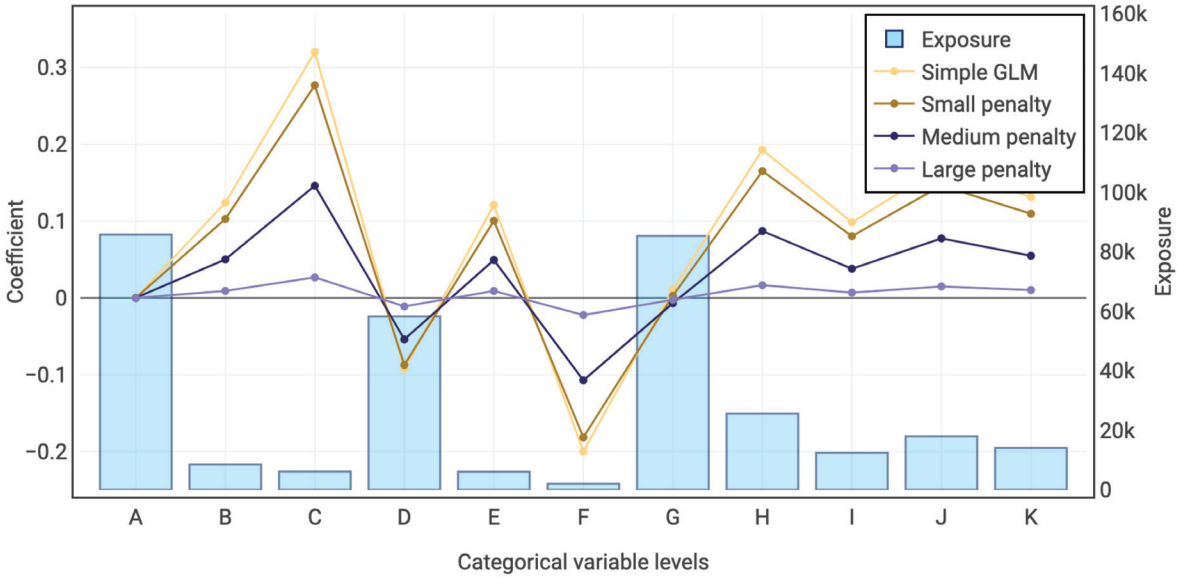
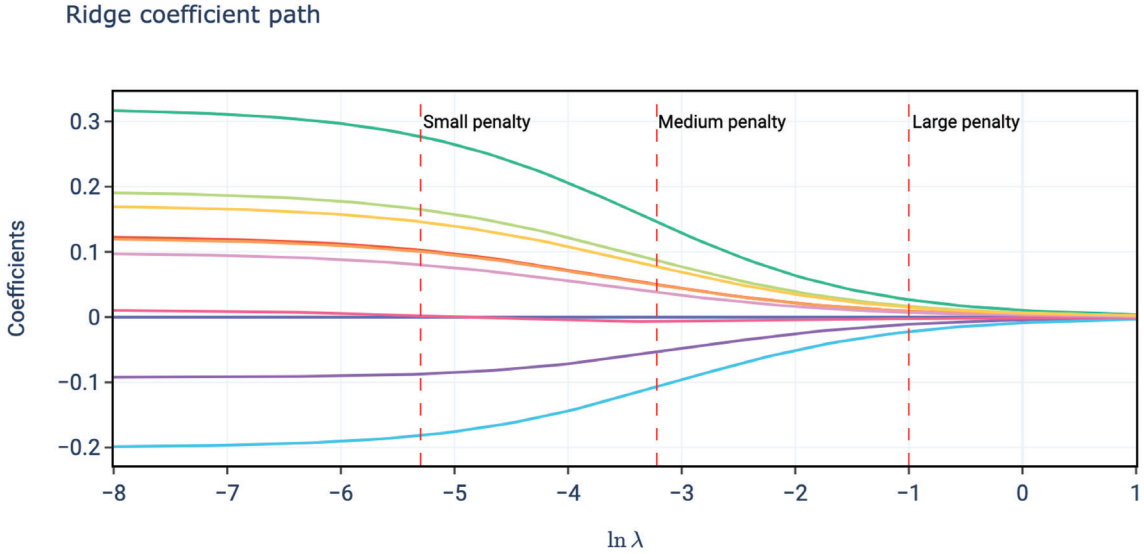


Figure 3.2. Coefficient path plot of Figure 3.1, obtained by computing solution for a wide range of penalty parameters. Dotted lines represent the values of λ penalty used to represent Figure 3.1. At very high values of the penalty (on the right), the coefficients are heavily shrunk and close to zero. Conversely, for very low values of the penalty, on the left, the coefficients are materially equal to the unpenalized GLM estimates.



movement of standardized coefficients with varying λ for ridge regression. A high penalty term will shrink all coefficients to a negligibly small value but will not reduce them directly to zero. As the penalty term decreases, all standardized coefficients increase gradually.

3.1.3. Lasso

The formula for the **lasso** regression adds the lasso penalty term to the GLM likelihood optimization:

$$\hat{\beta}_{\text{Lasso}} = \underset{\beta}{\operatorname{argmin}} \operatorname{NLL}\left(y, X, \beta\right) + \lambda\left(\sum_j \left|\beta_j\right|\right).$$

Introduced by Tibshirani (1996), the lasso achieves **sparsity**—i.e., the ability to set coefficients that are nonsignificant exactly to zero—as part of the fitting procedure. This means that lasso can automate both variable and factor selection and estimation. The sparsity property of the lasso is at the root of its wide success in various applications.

Figure 3.3 compares an unpenalized GLM with lasso GLMs using increasing levels of λ. Similar to ridge, the lasso's parameters tend to be shrunk toward zero. Furthermore, the lower the exposure, the higher the amount of shrinkage. The main difference is that for certain values of the parameter β_{Lasso}, the parameters β_j are set to zero, causing the rating plan coefficient to shrink to 1.0 after the application of the log link ($\exp(0) = 1$).

The lasso's ability to set coefficients to zero and thereby perform **variable selection** is shown through the coefficient path (Figure 3.4).

Figure 3.3. Comparison of an Unpenalized GLM and a GLM with Various, Increasing Levels of Lasso Penalty

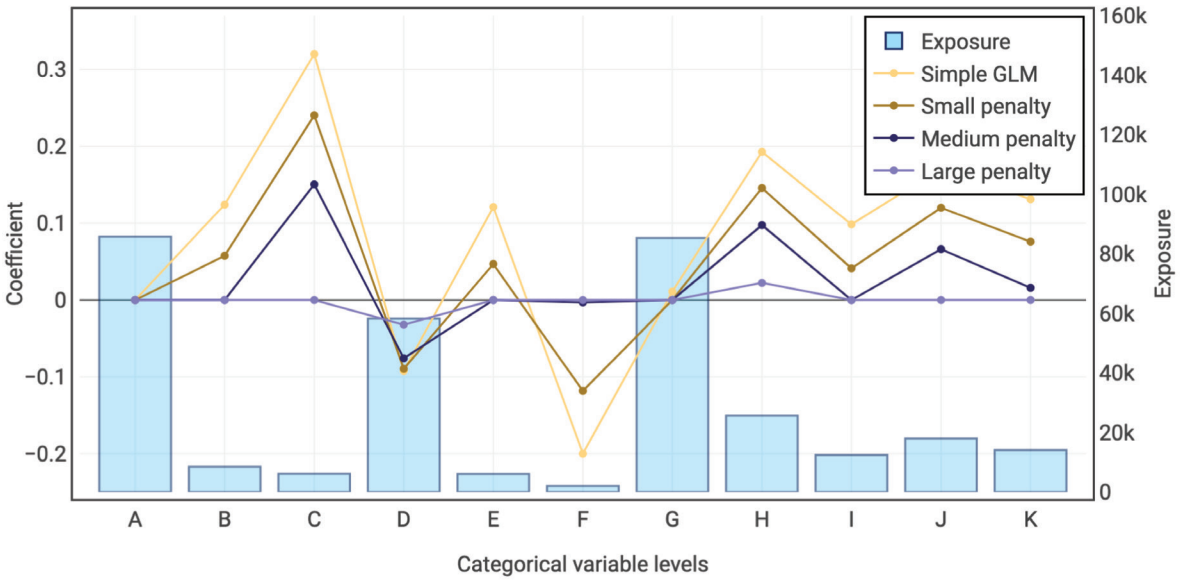
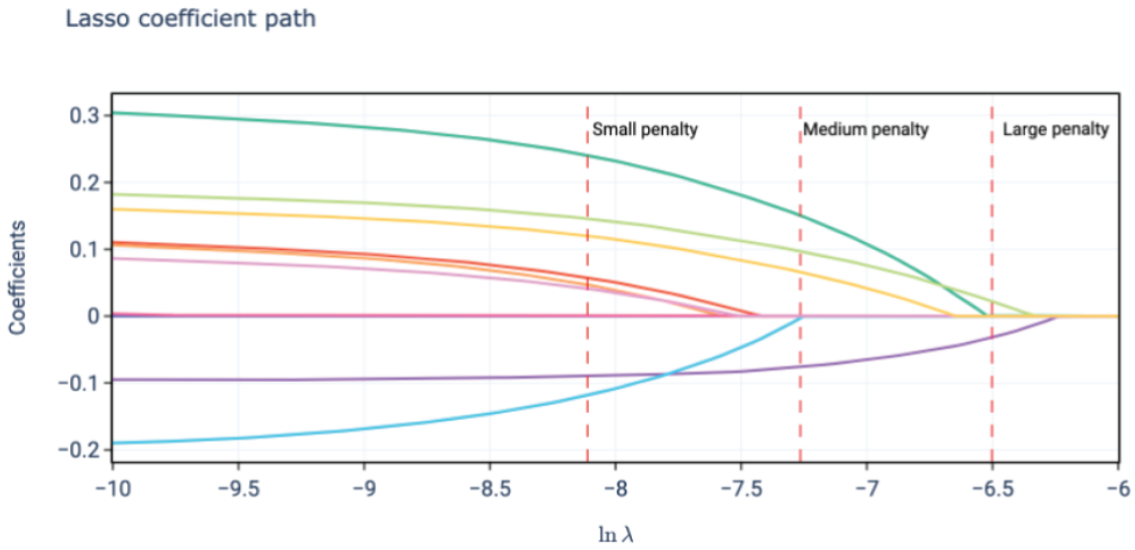


Figure 3.4. Coefficient path plot of Figure 3.3, obtained by computing solutions for a wide range of penalty parameters. Dotted lines represent the values of λ penalty used to represent Figure 3.1. Coefficients are completely removed from the model using lasso penalization with a sufficiently large penalty term.



In lasso penalization, a sufficiently high penalty term will set all coefficients exactly to zero. As the penalty decreases, coefficients will overcome the penalty and are introduced into the model at different times.

The interested reader can investigate the ability of lasso to achieve sparsity:

- Section A.3 shows how sparsity arises from a Bayesian perspective (priors).
- Appendix D illustrates an **optimization** perspective. In particular, it explains with simple arguments why introducing the absolute value function $|\beta|$, which is non-differentiable, leads to sparsity and variable selection.

Elastic Net

In the presence of strong collinearities among the variables X, the lasso and ridge may behave differently:

- Ridge will include all collinear variables and attribute a similar parameter β_j to each one of them.
- Lasso will likely select one of the correlated variables and set the others to zero.

Depending on the use case, a modeler could benefit from including both the ridge and lasso penalties in the same model. The flexibility of the penalized framework allows one to combine both penalties. This approach is known as **elastic net**, whose penalty is a convex combination of both the lasso and ridge penalties:

$$\hat{\beta}_{\text{Elastic Net},\alpha} = \underset{\beta}{\operatorname{argmin}}\, \text{NLL}\left(y, X, \beta\right) + \lambda\left(\frac{1-\alpha}{2}\sum_j \beta_j^2 + \alpha\sum_j \left|\beta_j\right|\right).$$

Figure 3.5 compares an unpenalized GLM with GLMs that incorporate various levels of elastic net penalty. Figure 3.6 shows the coefficient path for the elastic net, showcasing its ability to combine both the ridge regression and lasso behavior.

The elastic net requires the modeler to define an additional hyperparameter α on top of the parameter λ. In general, the choice of the parameter α depends on the nature of the data, and the methodology to compute this parameter is outside of the scope of the monograph.

Figure 3.5. Comparison of an Unpenalized GLM and One with Various Increasing Levels of Elastic Net Penalty

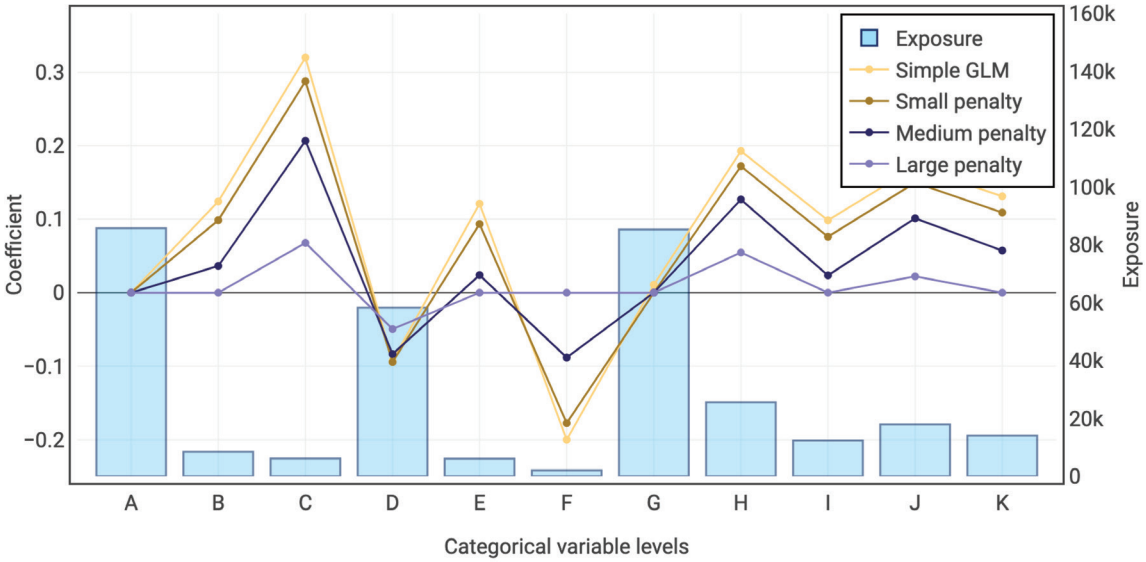
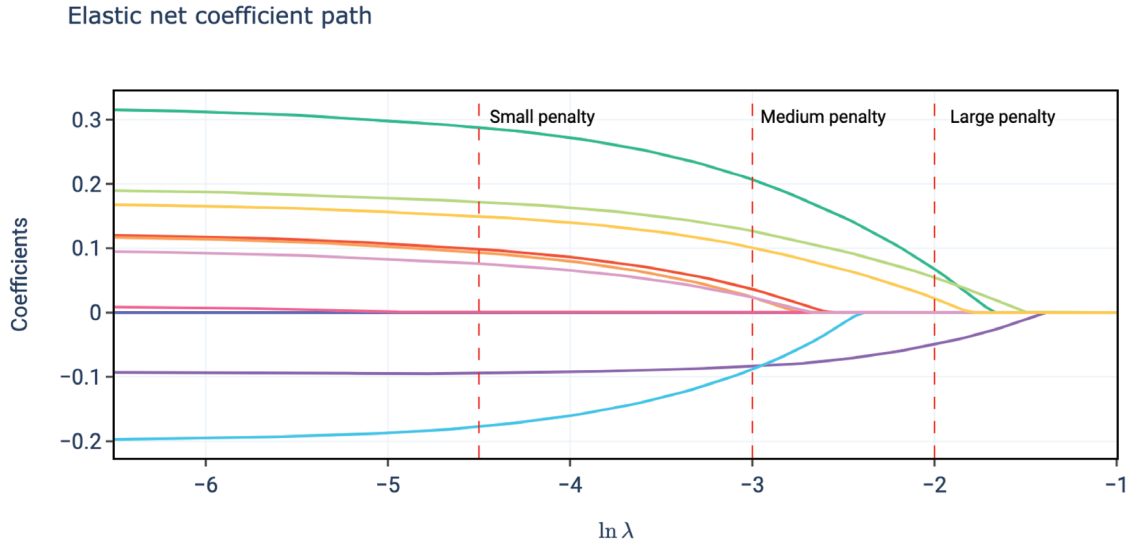


Figure 3.6. Coefficient path plot of Figure 3.5, obtained by computing solutions for a wide range of penalty parameters. The elastic net coefficient path is a blend of both the lasso and ridge coefficient paths, allowing for both variable selection and shrinkage of the coefficients.



3.2. Lasso is Recommended for Actuarial Applications

> A sparse statistical model is one having only a small number of nonzero parameters or weights. It represents a classic case of "less is more": a sparse model can be much easier to estimate and interpret than a dense model. In this age of big data, the number of features measured on a person or object can be large, and might be larger than the number of observations. The sparsity assumption allows us to tackle such problems and extract useful and reproducible patterns from big datasets.
>
> —Hastie, Tibshirani, and Wainwright (2015), *Statistical Learning with Sparsity*

As shown in Figure 3.4, for certain values of the parameter λ, some coefficients are shrunk exactly equal to zero. Since the solution of a lasso can set some of the coefficients exactly to zero, lasso is a natively sparse model.

On the other hand, the ridge penalty, as shown in Figure 3.2, does not have the ability to set coefficients directly to zero—hence it is not a natively sparse model. We won't detail all the purely statistical benefits of sparsity here as literature already exists on the subject, such as Hastie, Tibshirani, and Wainwright (2015).

In addition to statistical benefits, sparsity is valued for its actuarial benefits:

- A sparse actuarial pricing model will be more stable over time than a dense model. Avoiding constant changes in pricing characteristics is valued by insurers, regulators, and customers.
- A sparse actuarial pricing model will be simpler than a dense model. Interpretability is valued by internal stakeholders as well as regulators and policyholders.
- A sparse modeling technique automatically sets a statistical materiality standard. This clarifies the boundary between actuarial and statistical judgment during modeling and during model review.

Additionally, lasso penalization exhibits a desirable responsiveness to significant coefficients. Factor curves can be concave in lasso penalization as shown in Figure 3.4, and this allows a lasso model to be more immediately responsive to signal than ridge or elastic net. Once a variable has passed the threshold of materiality, its coefficient may grow quickly. Ridge penalization instead reacts slowly as all coefficient paths grow slowly at first and only quickly increase as the penalty term gets quite low.

We therefore recommend the use of lasso penalization for most actuarial applications. As we will see later, lasso penalization is ideal for the application of penalized regression as a credibility procedure.

3.3. Selecting the Penalty Parameter

The preceding sections emphasized the penalized regression framework's ability to incorporate credibility within a GLM through the introduction of both a penalty structure (ridge or lasso) and a penalty parameter λ. This penalty parameter clearly plays a critical role in determining the final estimates of the model, as we saw previously in Figure 3.1 and Figure 3.3.

While this section focuses on describing the standard methodologies selecting the penalty parameter in a penalized regression, it is important to note that the final decision on the penalty parameter should not be based solely on data-driven considerations. Actuarial judgment is also crucial and may be reflected in the choice of a slightly higher penalty parameter. This way, the final estimates can give more weight to the complement of credibility or a prior assumption.

Since the incorporation of actuarial judgment in the estimates can be better described under a more practical use case, we refer to Chapter 6 as a supplement to the methodology outlined here.

A "correct" value of the penalty parameter λ cannot be found via an explicit, analytical formula. This differs from other parameters used in actuarial methods, and the reason can be traced to the multivariate nature of the penalized regression.[5]

The lack of standard formulas is not in itself a limitation, as it allows a practitioner to approach the selection of the penalty parameter from a different, and in a sense, more practical perspective.

One of the desired behaviors of a model estimate is to generalize well to unseen data, and hence it is appropriate to choose statistical quality of fit (as measured by deviance or Gini) as the criteria to select the most appropriate penalty parameter. The standard procedure for choosing the penalty parameter consists of computing the generalization performance of a range of penalty values, and then selecting the value that has the best generalized predictive power.

The generalization performance is usually approximated via cross-validation (Hastie, Tibshirani, and Friedman 2009), which is a general procedure that allows us to "simulate" the behavior of a model with previously unseen data.

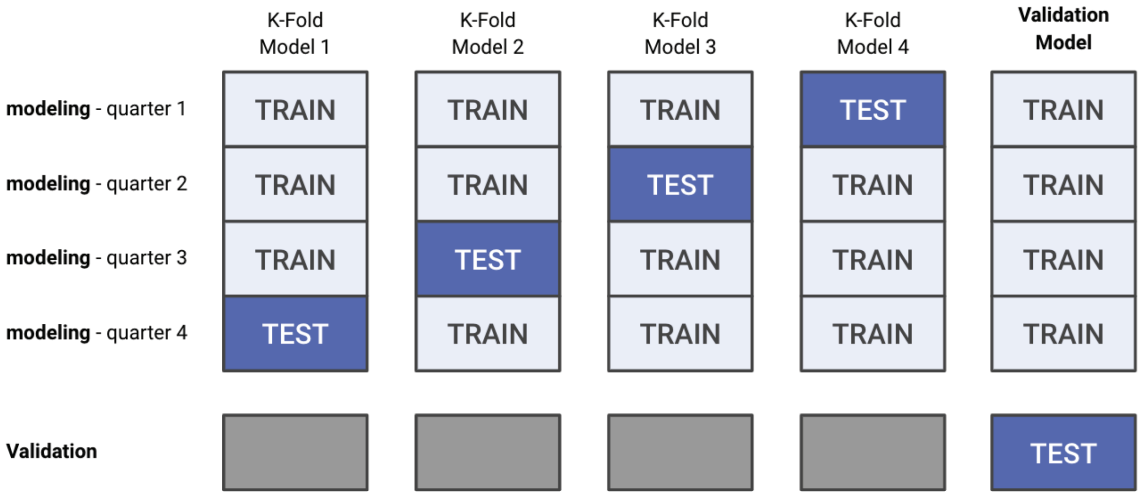
Figure 3.7 illustrates the procedure by which to evaluate, select, and validate the choice of a penalty parameter. We start by dividing the data into two sets: the modeling set and the validation set. The modeling set is used to build the model, and the validation set is used to assess the final model's performance.

We perform cross-validation on the training set, which involves dividing it into four (or any number of) folds and using each fold in turn as the test set while the rest is used for training. For each trained cross-validation model, we calculate performance metrics such as deviance, Gini, and pseudo-R^2 on the corresponding testing fold. These metrics are then combined to estimate the penalty parameter's performance on unseen data. Most solvers incorporate cross-validation routines, returning the penalty parameter value with the best average metric across all folds.

---

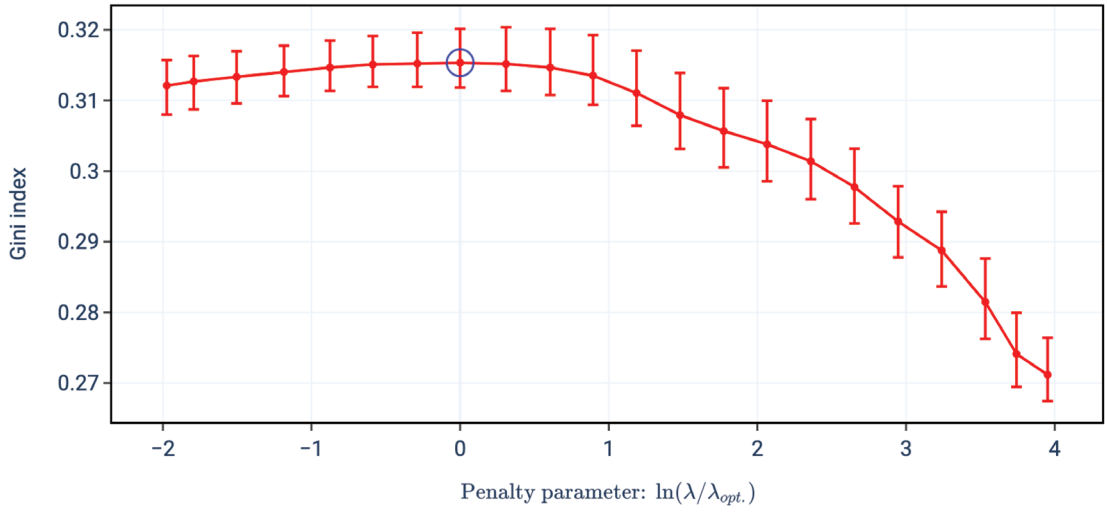
[5] This differs from standard credibility procedures, which do require the preliminary computation of some data-driven quantity of interest—for example, the K parameter in Bühlmann credibility (Table 2.1). This quantity is estimated via standard formulas (Bühlmann and Gisler 2005), which in particular require the estimates of average and variances for each of the individual classes. This implies that the quality of the estimator of K decreases as the number of individual classes are considered. When considering a multivariate model, as in penalized regression, the amount of observations sharing the exact same characteristics increases so significantly that these explicit formulas cannot be applied.

Figure 3.7. The Recommended Setup for 4-Fold Cross-Validation is a 20%/20%/20%/20% Split for Training and Reserving the Remaining 20% as a True Holdout Set



When selecting an optimal cross-validation penalty parameter, it is crucial to examine the overall results of the cross-validation process. Figure 3.8 demonstrates the evolution of cross-validation outcomes as the penalty parameter increases. Each red point signifies the mean error for a given penalty, while the error bars indicate the metric's variation within each score. Notably, even though the optimal penalty parameter λ_{opt} is identified, penalties within a (log) distance between −1 and 1 exhibit comparably similar results when accounting for score variation.

Figure 3.8. The model performance as measured by Gini increases as lambda increases up to a point, and then begins to decrease. Error bars represent the range of Gini calculated in cross-validation.



Given these observations, the cross-validation procedure should not unilaterally dictate the penalty parameter selection. Instead, it provides **a range of viable penalties** that actuaries can scrutinize by analyzing the model's coefficient values, among other factors such as overall reasonability of the model and other actuarial considerations. In this way, although the cross-validation process assists in informing the penalty parameter choice, the final value should always be assessed for actuarial appropriateness. One frequent application of actuarial judgment is to select a marginally higher penalty value, leading to estimates closer to the selected complement of credibility.

After selecting the final penalty parameter value and finalizing input variables, the modeler should validate the model on the holdout validation set, which has not been used during the cross-validation routine, as seen in the "Validation model" column of Figure 3.7. When comparing a proposed model to a current model, double lift charts built on the full k-fold training set will not be helpful as an overfit model will seem to outperform a properly fit model. A true holdout set allows candidate models to be compared fairly on data neither model has seen before.

3.4. Lasso and Variable Transformations

There are various ways to include a variable in a GLM model. The GLM monograph by Goldburd et al. (2016) splits the nature of predictor variables into two groups: **categorical** and **continuous.**

We analyze the impact of the lasso penalty for each of these variable types and additionally provide a specific discussion of **ordinal** and **control** variables.

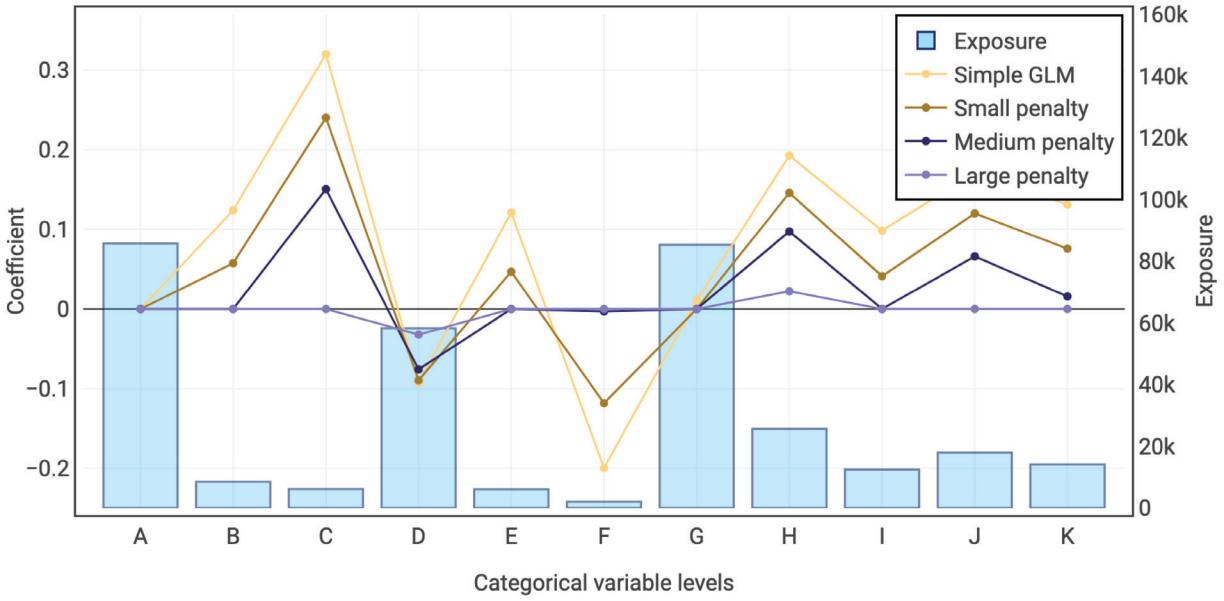
3.4.1. Categorical Variables

A categorical variable takes on one of two or more possible values, thereby assiging each risk to a "category." Each of these values (levels) is modeled independently—i.e., it has a dedicated coefficient β_j. In a GLM (penalized or unpenalized), a categorical variable is represented by collections of β_j, each representing the impact of each category with respect to an arbitrary fixed level, called the "base level." Including or excluding a specific coefficient β_j determines whether such a level is deemed significant by the modeler.

In Figure 3.9, we repurpose Figure 3.3 to illustrate the impact of the penalty on categorical variables.

Depending on the strength of the penalty, the lasso sets some coefficients to zero, thus providing an adaptive grouping of those less significant levels with the base. For the other selected levels, the value of the coefficient provides a "credibility-weighted" deviation from the base level. It is worth noting that unlike in a GLM, the predictions from penalized regression will change if a different base level is selected. In a GLM, the predictions will be the same but the confidence intervals and p-values will be different. Readers are encouraged to see how much this choice matters in the case study (Chapter 7) by changing the selected base level for various categorical variables.

Figure 3.9. Comparison of an Unpenalized GLM and a GLM with Various, Increasing Levels of Lasso Penalty



3.4.2. Continuous Variables

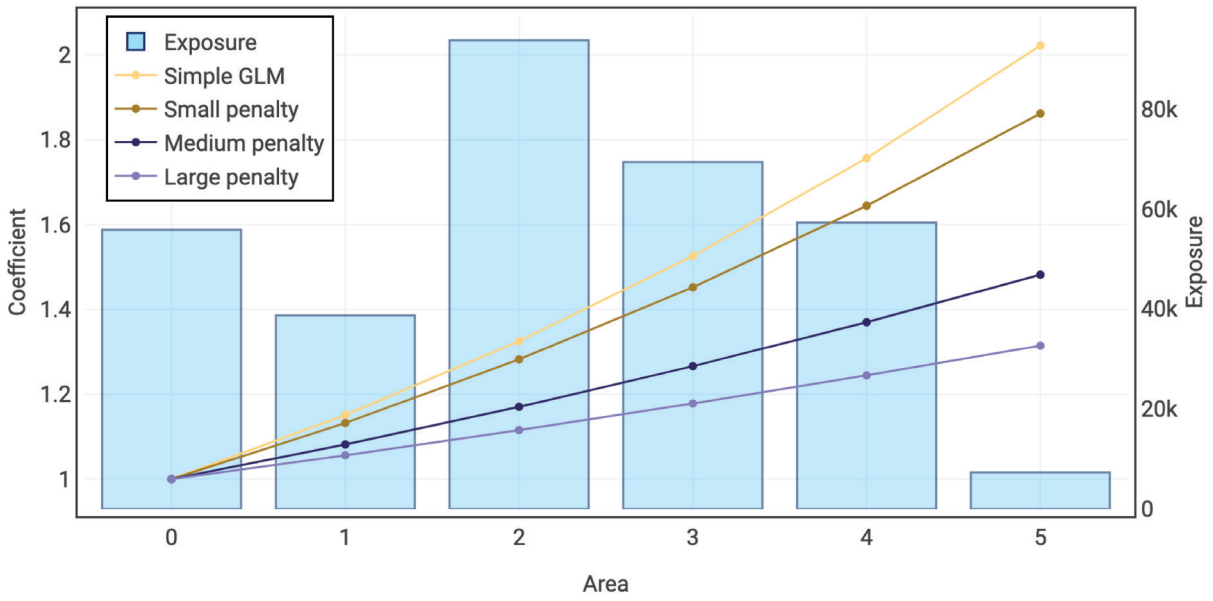
A continuous variable is a numeric variable that represents a measurement on a continuous scale. In a GLM, continuous variables can be represented by one or multiple coefficients β according to the nature of the variables and the modeling decisions.

We start by evaluating the impact of the lasso penalty for a **linear** representation of a continuous variable: a single coefficient β is associated to a variable via the relationship β_x. The β value represents the **slope** of the linear impact of the variable in the model. Since the lasso penalty either shrinks or sets to zero the coefficient, the impact of the penalty will correspond in a slope reduction or removal of the variable depending on the strength of the effect as seen in Figure 3.10.

Whereas categorical variables identify a **change in level**, continuous variables identify a **change in slope**. Continuous variables may be appropriate for lasso penalization, but we will see that continuous variables are quite difficult to use in lasso credibility (defined in Chapter 5).

To make matters more challenging, in practical applications, the linear representation may be overly simplistic due to the nonlinearities naturally arising in insurance data. Examples could be either age of home or age of driver. Those variables may be represented by multiple coefficients β_j, each mapped to the parameter of some nonlinear curves. One such example is a third-degree polynomial encoding, where for a given variable x three coefficients β_1, β_2, β_3 will represent respectively a linear, a parabolic, and a cubic function. Under such a representation (or **feature engineering**) the interpretation of each individual coefficient is less intuitive: the modeler can determine the

Figure 3.10. Lasso Fit for Various Penalty Values λ for a Continuous Variable



appropriateness of the factor only by plotting the joint effect of the coefficient to the variable as illustrated in Figure 3.11.

Since the direct interpretation of each individual coefficient to the model is not clear, for such feature engineering sparsity is less necessary. Furthermore, polynomial feature transformations give rise to correlated predictors (x, x^2, x^3). One caution when using lasso for variable selection is that the presence of many highly correlated predictors will produce suboptimal results due to the staggered entrance of such predictors (Hastie, Tibshirani, and Friedman 2009). Using lasso to determine the optimal combination of feature transformation is not recommended.

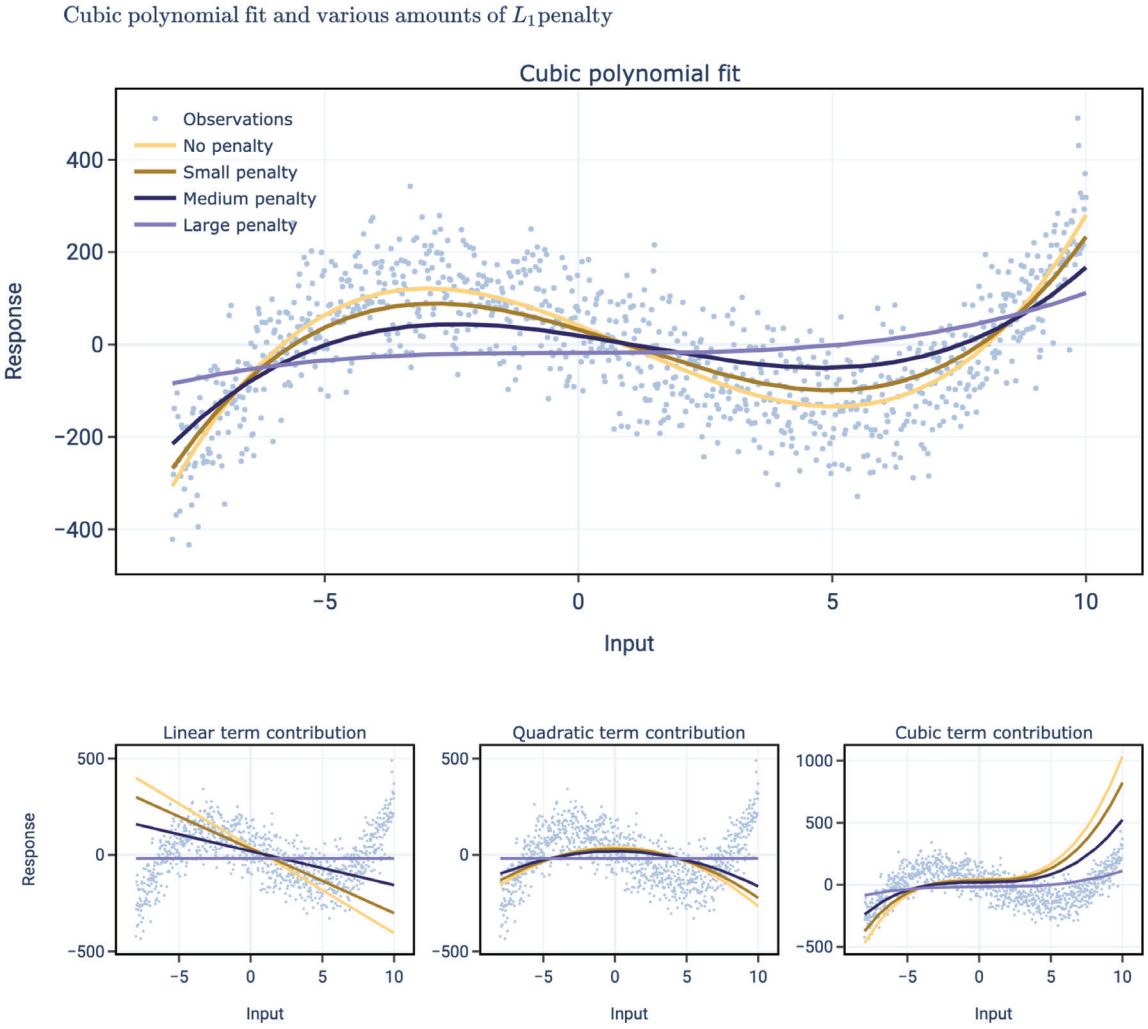
The modeler wishing to extensively combine polynomial terms or implement other complex feature transformations such as linear or cubic splines may find it beneficial to include some ridge penalty via the elastic net penalized regression. On the other hand, the lasso penalty allows us to model nonlinearities in a much more efficient manner than unpenalized GLMs via the **ordinal treatment.**

3.4.3. Ordinal Variables

An ordinal treatment of a variable relies not on the numeric values of the variable but instead on stepwise indicators. Under such a treatment, when a coefficient is zero, it results in the grouping of two consecutive levels. Such a representation is extremely powerful and allows the lasso penalty to automatically detect nonlinear effects.

The idea of blending an ordinal treatment with lasso penalized regression was originally proposed by Tibshirani et al. (2005) in their fused lasso paper. From that paper, several variations of the lasso penalty have been explored and discovered in various fields. In the actuarial field, this methodology is explored from two different perspectives.

Figure 3.11. Lasso fit for various penalty values λ for a third-degree polynomial fit $\beta_1 x + \beta_2 x^2 + \beta_3 x^3$ on a continuous variable. The top of the illustration represents the cumulated effect of the polynomial curve with different degrees of penalization, which results in an overall shrinkage of the curve. The bottom of the illustration provides the plot of the change of the individual polynomial function $\beta_i x^i$ to the varying degrees of penalty.

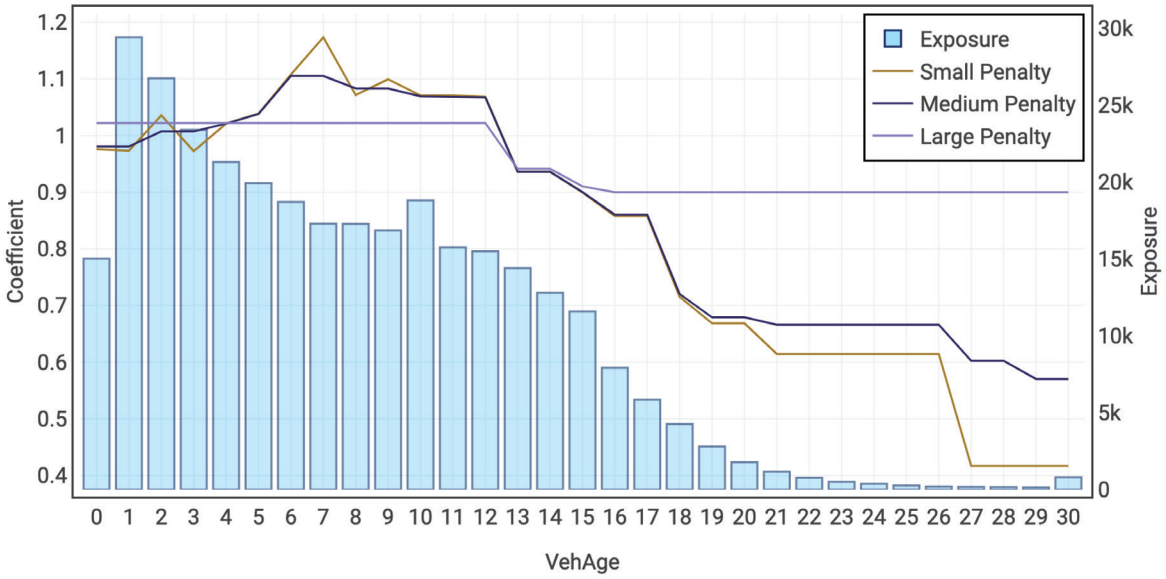


In the accurate GLM (AGLM) method of Fujita et al. (2020), variables are transformed and encoded through step or binary transformations. In the derivative lasso of Casotto and Holmes (2023), the difference between these consecutive numeric levels is penalized directly. The results in both cases coincide.

Figure 3.12 represents the resulting effects of varying levels of lasso penalty for ordinal variables.

Ordinal variables also help in the selection of an actuarially sound penalty term. We can start with the statistically optimal penalty term and then increase it slightly until any unintuitive behaviors are penalized out of the model. In Figure 3.12, suspicious

Figure 3.12. Lasso Fit for Various Penalty Values λ for an Ordinal Variable



reversals present in the small penalty scenario are subsequently canceled in the medium- and large-penalty scenarios. The medium penalty can be considered both statistically and actuarially appropriate.

3.4.4. Control Variables

Control variables like "year" and "state" are often used in loss models to account for (control for) the signal from such variables so that they do not flow into other risk characteristics. In GLMs, such variables are often left in the model regardless of significance. In lasso regression, it may happen that some levels of these control variables will be removed from the model by penalization.

Whether or not to apply a penalty term to such special factors is a modeling and actuarial decision, and both options can be motivated by different arguments. In most cases, allowing control variables to be fitted and penalized with other variables is appropriate. By fitting them at the same time, the model has the opportunity to allocate signal appropriately between control variables and potentially correlated predictor variables. Additionally, if a control variable is removed from a model through penalization, it is unlikely that the limited signal will have a material effect on correlated predictors due to the same penalization.

If a modeler wants to ensure that a control variable soaks up all the signal that it possibly can (at the risk of taking signal away from predictor variables), it may be appropriate to apply a stepwise modeling approach. A modeler could first fit the control variables and optionally a few key predictors with a low or absent penalty term. Then, they can offset the coefficients for those control variables when fitting their desired model with an appropriate penalty term. Using the stepwise approach, the modeler is deciding to give less of a penalty to their control variables.

3.5. Lasso for Variable Selection

One can also use lasso penalization for variable selection because of its ability to set coefficients directly to zero while maintaining a material coefficient for other variables. A modeler can start with a lambda that is sufficiently high to remove all variables from the model. Then, the modeler can gradually decrease the penalty until variables begin to enter the model. In our earlier example (Figure 3.4), this method may be appropriate for variable selection. This approach is not possible with ridge, as the descent of all coefficients is much more uniform, and coefficients are never set directly to zero.

One caution when using lasso for variable selection is that the presence of many highly correlated predictors will produce suboptimal results due to the staggered entrance of these highly correlated predictors. Such a highly correlated collection may arise when including a wide variety of transformations of the same variable in a model (e.g., $x \rightarrow x^2$, x^3, $\log(x)$, . . .) as seen in Section 3.4.3. Using lasso to determine the optimal combination of feature transformation is not recommended. Instead, some amount of variable pruning is necessary, and we recommend actuarially selecting between highly correlated predictor variables before using lasso for variable selection.

4. The Bias–Variance Trade-Off

4.1. Introducing the Bias–Variance Trade-Off

We have explained what penalized regression is, but we have yet to demonstrate how it improves upon an unpenalized GLM. The core of this argument is intuitively supported by our cross-validation lambda grid search (Figure 3.8). At the leftmost point, the performance of a GLM is displayed. As we increase the penalty, the performance increases up to a maximum, and then decreases. This is not by chance, nor is it a cherry-picked example from the data. Both in practice and theory, this happens consistently: a model with (the right amount of) penalization will outperform a standard GLM. This fact is determined by what is known as the **bias–variance** trade-off.

The bias–variance trade-off is a very generic and general concept in machine learning. Bias here must not be confused with "biased" model in the context of protected classes as described in Mosley and Wenman (2022). It may be helpful to also remember that mean squared error (MSE) decomposes directly to bias and variance: $\text{MSE} = \text{Bias}^2 + \text{Variance}$. When the bias–variance trade-off is improved, the MSE decreases.

We introduce the concept of the bias–variance trade-off from the perspective of a GLM, and demonstrate how variable selection is often performed to maximize this trade-off. The trade-off may be viewed analogously as **underfitting versus overfitting**:

- A model with **high bias** is often described as **underfit.**
- A model with **high variance** is often described as **overfit**.

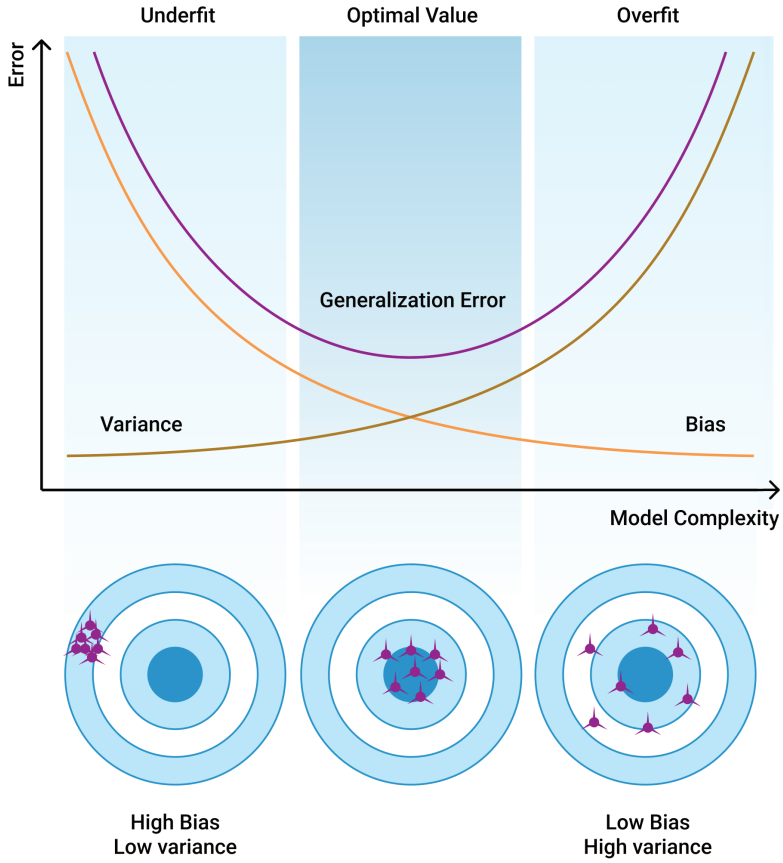
We then show how **penalized regression reduces variance through coefficient shrinkage** rather than manual coefficient removal. Finally, we draw connections between the bias–variance trade-off and both penalized regression and credibility. These connections will help lay the groundwork for the treatment of penalized regression as an actuarially sound credibility procedure.

4.2. Defining the Bias–Variance Trade-Off

Statistical theory shows that the error of **any** model on unseen data is the sum of two kinds of errors: the **bias** and the **variance** as described in Figure 4.1.

- **Bias** represents the error between the model we are building and the "real" model. Assuming we have access to infinite data, the bias will measure how the structure of the model we are building will replicate the "real" underlying model.

Figure 4.1. In a hypothetical scenario where we have access to an almost infinite amount of data, we would favor a very complex model such as the one on the right with plenty of variables (low bias). However, in a realistic scenario with limited data points, such a model will poorly generalize due to the instability of the parameters (high variance). Conversely, if we were to choose a too simplistic model, we could obtain a very stable model having low variance but very high bias. In practice we would always favor a right trade-off between bias and variance.



- **Variance** represents the error of building the model on the specific data set that we used to fit the model versus other, richer data sets. This component becomes more important the smaller the data set and the "noisier" the effect. Unfortunately, small noisy data is quite common when dealing with insurance data.

Minimizing the generalization error is desirable when building models, and ideally one will look to minimize both the bias and variance error components **simultaneously**. Unfortunately, the bias–variance trade-off highlights a harsh truth: minimizing one of either model bias or variance will irremediably increase the other. Finding the model with the minimal error requires finding the **optimal bias–variance trade-off**.

4.3. Bias–Variance Trade-Off: A GLM Perspective

As a reminder, **variance** represents the error of building the model on a specific data set used to fit the model versus other, richer data sets. Imagine you are building a two-variable model on your own imperfect modeling data set. If you magically had twice the amount of data, the modeled coefficients would be more accurate and the predictions would have less error. Double it again, and this improvement continues. The error being reduced due to modeling on a richer data set is **error introduced by variance**.

Unfortunately, we cannot magically double our data. Some modeled coefficients may be unreliable, and a typical solution is to remove those coefficients from the model. You may decide to remove one variable and now you are left with a one-variable model. No matter how much data is added to this one-variable GLM, the coefficient for our removed variable is fixed at 0 and its error will never increase or decrease. The error between 0 and the coefficient's true value is the **error introduced by bias**.

Did we make the right decision? That depends. If we leave the variable in our model, how much error in our predicted estimates is introduced by variance? If we remove the variable, how much error is introduced by the bias of setting this coefficient directly to zero? Before describing how the trade-off can be maximized through penalization, let's explore an example of the trade-off in an unpenalized GLM.

Let's think back to our homeowners example GLM in Section 1.6. In the second scenario, the fire extinguisher variable certainly has an impact on the target we are modeling. However, the amount of underlying exposures is so limited that the resulting GLM estimate will have a high variance.

In Table 4.1, model 1 is our original model and model 2 is a new model fitted without the fire extinguisher indicator.

Table 4.1. Sample GLM Results for Models Fit with and without the Fire Extinguisher Variable

Model Output	Model 1	Model 2
exp(Intercept)	100	105
$\exp(\beta_{\text{no fire extinguisher}})$	1.2	NA
$\exp(\beta_{\text{age of home}})$	1.01	1.01
Average overall prediction	110	110

NA = not applicable.

Risk Description	Model 1 Prediction	Model 2 Prediction
No extinguisher, home age 0	120	105
With extinguisher, home age 0	100	105

By removing the fire extinguisher variable, we have certainly reduced the variance of our model. Remember that the **variance** represents the error of building the model on the specific data set we used to fit the model versus a different or more robust data set. When we remove this variable, the factor for not having a fire extinguisher will be the same on any data set that we use: 1.0. Certainly we will not see an increased or decreased error from the fire extinguisher variable on a different data set. Now, the question is whether the removal of this variable and corresponding reduction in variance introduces too much **bias** into our model. Are we potentially moving too far away from the "real" model by removing this variable?

If we include too many variables in our model, it will have a high variance and be **overfit**. To take an extreme example, imagine that only a single policy does not have a fire extinguisher. The extreme resulting relativity would be quite overfit to the training data and would perform poorly on the holdout data. It is intuitive that adding more and more information on policies without fire extinguishers would decrease the error of this model.

If we do not include enough variables in our model, it will be biased and **underfit.** Underfit models are excluding relevant information that could make for a more accurate prediction. If we remove the fire extinguisher variable, we may be excluding relevant information. The decision about whether to include a variable is made through an evaluation of the bias–variance trade-off.

4.4. Evaluating the Bias–Variance Trade-Off

We evaluate the bias–variance trade-off in a GLM after model fitting using various statistics. The traditional metrics are the Akaike information criterion (AIC) and the Bayesian information criterion (BIC):

$$\text{AIC}(\beta) = 2\text{NLL}(\beta) + 2\{\#\text{ of degrees of freedom}\}.$$

$$\text{BIC}(\beta) = 2\text{NLL}(\beta) + \log(n_{obs})\{\#\text{ of degrees of freedom}\}.$$

Both metrics are "penalized" measures of fit, as they are both the sum of two terms:

- The "likelihood" term, expressing the quality of fit within the training set
- The "degrees of freedom" term, which expresses the **complexity** of the model in terms of number of parameters

Both the AIC and the BIC are metrics that penalize the goodness of fit based on the number of coefficients in the model. There is a hurdle in the measure of fit that each coefficient must overcome to be included in the model.

If our fire extinguisher variable overcame this improvement and model 1 had a better AIC than model 2, model 1 would be superior and model 2 would be considered **biased** and **underfit**. If model 2 had a better AIC, this is a sign that the additional variable is not adding sufficient power and model 1 may have too much **variance** and be **overfit**.

It is impossible to improve both bias and variance simultaneously. When building a traditional GLM, the tools available to the modeler to maximize the bias–variance trade-off are the addition and removal of variables. A modeler must make use of post hoc penalized measures of fit like the AIC and the BIC to determine the bias–variance trade-off of a variable in a GLM.

Conversely, it is possible to find the bias–variance trade-off that best generalizes on unseen data. In fact, we demonstrated this optimization process earlier by the selection of a lambda penalty parameter through cross-validation. Rather than using a post hoc penalized goodness-of-fit statistic to evaluate the bias–variance trade-off of variable inclusion and exclusion, lasso penalized regression uses shrinkage to apply an optimal bias to coefficients during the fitting process. If the coefficient that maximizes this trade-off is zero, lasso penalization will remove the coefficient from the model completely. The use of penalization within the model fitting process removes the need for post hoc penalized metrics.

Section C.3 further develops how the generalization error can be measured. In particular, it shows how cross-validation can be seen as a better alternative to AIC and BIC, and further details the connections between AIC and BIC and penalized regression.

In more traditional actuarial analysis, this bias–variance trade-off is often addressed through credibility procedures.

4.5. Bias–Variance Trade-Off and Credibility

Let's think conceptually about the bias–variance trade-off and credibility using the fire extinguisher variable from our earlier example model. In the example, our data was showing an indicated factor of 1.2. If our data is thin, we might combine this with a selected complement of credibility of 1.1 using classical or Bühlmann credibility:

$$\text{Credibility-weighted factor} = 1.2 \times Z + 1.1 \times \left(1 - Z\right).$$

By weighing these together, the resulting estimate between 1.2 and 1.1 will have less variance than 1.2 alone if we have selected a stable, reasonably uncorrelated complement.

However, we are now introducing a bias to our models. Our selected complement of credibility is not based on the data, and therefore imposes a potential source of error outside of the data. The Z used in credibility is calculated to maximize the bias–variance trade-off by reducing the variance of a partially credible estimate through the introduction of an informed bias to the estimate.

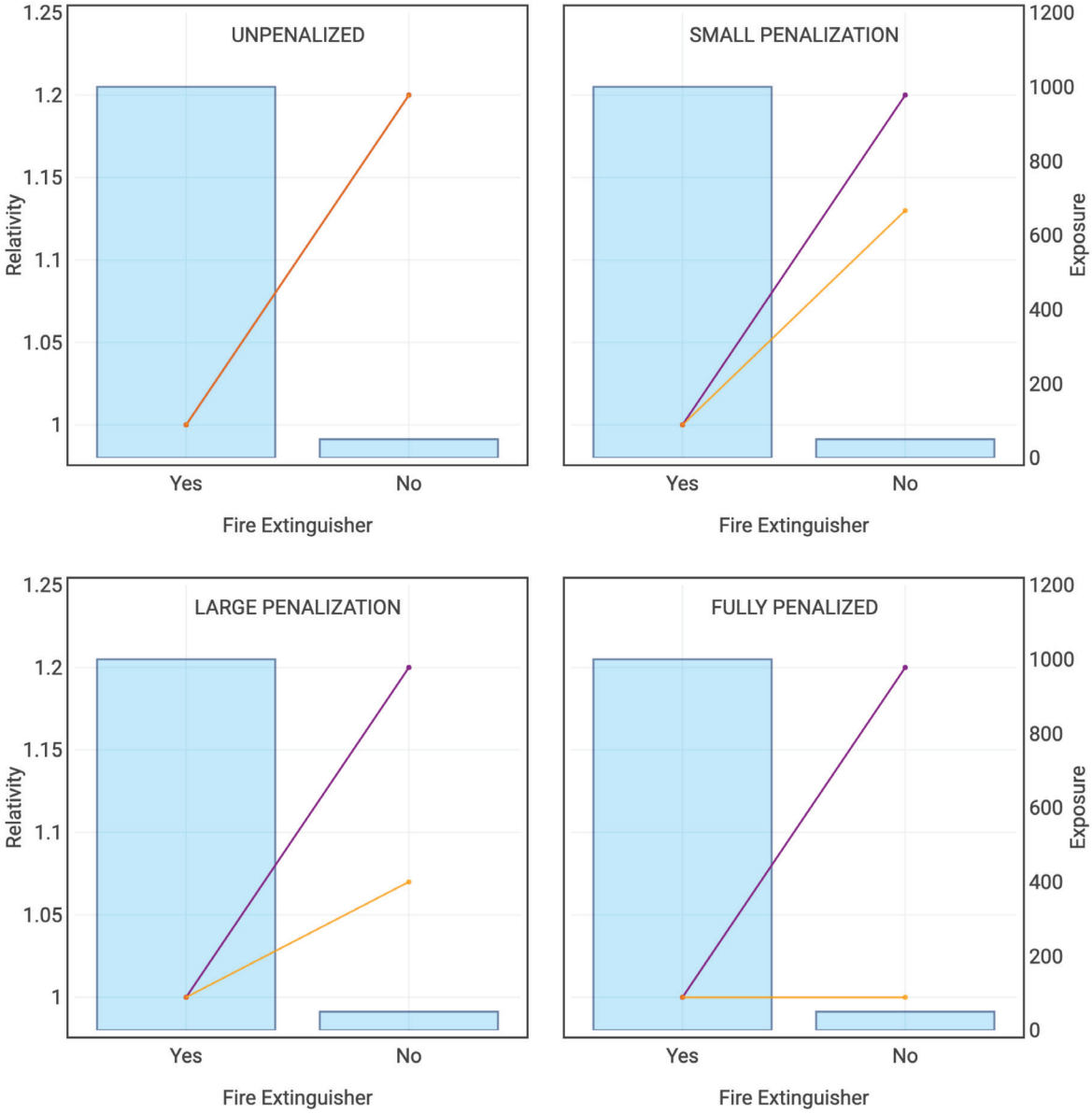
The shrinkage introduced by penalized regression similarly introduces a bias to reduce the variance of estimated predictive values. The difference between this example and traditional penalized regression is that when using credibility, we are biasing our estimate toward a selected complement instead of a null coefficient. We now explore how penalized regression acts as a credibility procedure through this introduction of bias. This introduction lays the foundation for our later discussion on the implementation of lasso credibility.

4.6. Penalized Regression and Credibility

Let's again return to our fire extinguisher–only model. A GLM that has been fitted using only the fire extinguisher variable will produce a coefficient corresponding exactly to the pure premium relativity of the two categories in the data. Owing to the full credibility assumption of GLMs, the coefficient will correspond exactly to this relativity no matter how few exposures are in the "No" category. This is the unpenalized graph in Figure 4.2.

Now, we are going to apply a lasso penalty to our model. By introducing a sufficiently large penalty term, our output coefficient for "No" will shrink all the way to

Figure 4.2. The experienced (purple) and the indicated relativity (yellow) are shown across different values of penalization λ for the fire extinguisher variable. The experienced and indicated relativities completely overlap without penalization.



zero and be removed from the model. This extreme application of bias is effectively assigning no credibility to the data. This is represented by the fully penalized graph in Figure 4.2.

When lambda is between these two extreme values, the coefficient will be somewhere between the GLM estimate and zero.

The bias that penalization introduces can be restated in the traditional credibility equation. We can represent every predicted value of β_i using a credibility value of $0 < Z < 1$ by weighing a fully credible GLM estimate β_{GLM} with the noncredible coefficient estimate of 0. Note that Z is arbitrary and cannot be directly calculated from the selected penalty parameter:

$$\beta_i = \beta_{GLM} \times (Z) + 0 \times (1 - Z).$$

This similarity is not a coincidence. In Section A.2, we prove that **Bühlmann credibility and ridge penalization are in fact equivalent in a special case**. The mathematical relationship between penalization and credibility has been explored in Miller (2015) and Casotto, Banterle, and Beraud-Sudreau (2020), and we include a detailed explanation of the relationship between penalized regression and Bayesian statistics in Appendix A. Additionally, Appendix B contains a robust defense of penalized regression as an actuarial credibility procedure through the lens of ASOP 25. Rather than simply being "credibility-like," we suggest that penalized regression, when used properly, can be applied as an actuarially sound credibility procedure in the form of **lasso credibility**.

4.7. Conclusion: Benefits of Lasso Penalization

Penalized regression applies a penalty term to the size of coefficients during the maximization of likelihood. The penalty value λ adjusts coefficients for their credibility and volatility. Selecting the penalization factor allows one to find the optimal generalization error of the bias–variance trade-off. Therefore, it is not necessary to perform post hoc significance testing to ensure that all coefficients are valid. To the extent that a variable is not beneficial to the fitting process, it is given partial credibility and shrunk toward zero **during** the fitting process. Lasso penalization will fully remove noncredible coefficients from the model **during** the fitting process.

This evaluation of coefficients during the fitting process directly addresses the issues with *p*-value significance testing discussed in Section 1.6.1.

Lasso penalization

- answers the question "Is this coefficient credibly not zero, and how much can we trust this coefficient?" as opposed to just "Is this coefficient likely not zero or more extreme?";
- provides a sound estimate of how much a coefficient or effect can be trusted—rather than a post hoc univariate adjustment for *p*-values close to 0.05, penalized regression automatically shrinks such coefficients by optimizing the bias–variance trade-off in a multivariate analysis (estimates are developed **jointly** considering correlations, and not on a one-by-one basis, as in GLM or *p*-values);

- does not use an arbitrary threshold of significance (e.g., 5%) but instead removes coefficients that do not overcome the penalty parameter (this penalty parameter is tuned and adjusted for each model individually based on sound procedures);
- does not require the post hoc removal of variables and refitting, as it will automatically remove noncredible variables during the fitting process; and
- does not require the post hoc adjustment of variables based on significance testing, as those variables are adjusted on a **multivariate basis** during the fitting process based on their credibility.

Penalized regression takes the null hypothesis in *p*-value significance testing and instead uses a version of this hypothesis as the complement in a credibility procedure. **Whereas variable evaluation in GLMs is a binary, often univariate, post hoc process, variable evaluation in penalized regression is a continuous, multivariate process that occurs during model fitting**.

Now that we have introduced penalized regression as a credibility procedure using a significance test's null hypothesis as a complement, we are ready to change this complement to something more actuarially appropriate through **lasso credibility**.

5. Lasso Credibility

We have identified penalized regression as a credibility procedure where the complement of credibility is a null coefficient $\beta = 0$. This complement of credibility coincides with the null hypothesis in *p*-value testing. The next logical step is to enhance this procedure by using a more appropriate complement of credibility than $\beta = 0$ for all β.

Implementing such a credibility procedure has three requirements: an **offset** representing the complement of credibility, an **ordinal** or **categorical** treatment of all variables, and the **use of penalized regression** as the credibility procedure. This methodology is best applied through lasso penalization instead of ridge or elastic net, and we will refer to it as **lasso credibility**.

5.1. The Offset: Applying a Complement in Lasso Credibility

In actuarial modeling, the offset has traditionally been reserved for the application of weights as well as deductibles, limits, or other rating relativities that are best selected outside of a GLM. The reader can find a comprehensive overview of the applications of offsetting in Yan et al. (2009). When applied to GLMs and traditional penalized regression, an offset is normally included without a corresponding predictor variable.

The mathematical definition of an offset is a fixed column of coefficients that contributes to the linear component $X\beta$ of a GLM or penalized regression:

$$\eta = g(\mu) = \beta_0 + X\beta + \text{offset}.$$

Consider the example in Section 4.5, where we wanted to incorporate a complement of credibility of 1.1 to the fire extinguisher variable.

Lasso credibility implements such a complement of credibility via a new application of the offset. In lasso credibility (as opposed to GLMs or traditional penalized regression) it is necessary to **include both an offset as well as the predictor variable for the same rating characteristic**.

The model's prediction can be stated as

$$\begin{aligned}\log(\mu) = \log(\text{Prediction}) &= \beta_0 + \text{offset} + \beta_1 X_1 + \beta_2 X_2 \\ &= \beta_0 + \beta_{1\,\text{offset}} X_1 + \beta_{2\,\text{offset}} X_2 + \beta_1 X_1 + \beta_2 X_2 \\ &= \beta_0 + (\beta_{1\,\text{offset}} + \beta_1) X_1 + (\beta_{2\,\text{offset}} + \beta_1) X_2. \qquad (5.1)\end{aligned}$$

We see that the coefficients of the model can be decomposed into two components:

- The **fixed** component $\beta_{j,\text{offset}}$ fully determined by the modeler's assumption
- The **variable** component β_j, determined by the data and the methodology (GLM or penalized GLM).

Fitting this model as a GLM with or without an offset will yield the exact same predictions because of the GLM's assumption of full credibility (see Section A.1). If the offset is not included, the GLM will output the best β_j to maximize the likelihood in the optimization process. We call this $\beta_{j,glm}$. If an offset is included, the GLM will output a β_j such that

$$\beta_{j,\text{offset}} + \beta_j = \beta_{j,\text{glm}}$$

and the likelihood is again maximized. The accompanying code for Chapter 7's case study contains an example of a GLM with and without offsets producing the exact same predictions.

5.2. Ordinal Variables

Ordinal variables, like categorical variables, represent **magnitude**; continuous variables, on the other hand, represent **slope.** By representing every variable in the model as ordinal or categorical, all of the resulting coefficients represent a **magnitude** of change. That magnitude is consistent with traditional credibility approaches, and therefore actuarial judgment and the considerations of ASOP 25 are easy to apply. Additionally, by using an ordinal treatment of variables, lasso credibility can also automatically identify deviations from the selected complement of credibility without additional feature engineering.

By treating a continuous variable as a stepwise ordinal variable (one step for every driver age, for example), a lasso credibility model has the ability to identify the **magnitude** of a credible difference from the complement at any step. It is by chaining these ordinal steps together that lasso credibility can fit complex and unknown deviations from a complement of credibility. Those steps without a credible difference are automatically removed when using lasso penalization, while steps with a credible difference are included. It is recommended that ordinal steps be sufficiently granular to not inappropriately pregroup levels of a given characteristic.

5.3. Lasso Credibility as a Credibility-Weighted GLM

We now focus on how penalized regression can be leveraged to find solutions that are trade-offs between the complement of credibility (here β_{offset}) and the observed data (in this case β_{glm}).

Suppose that we are building a lasso regression model, with penalty level λ, with the following prediction structure:

$$\text{Prediction} = \exp\left(\beta_0 + \left(\beta_{1,\text{offset}} + \beta_1\right)x_1 + \left(\beta_{2,\text{offset}} + \beta_2\right)x_2 + \cdots\right).$$

When $\lambda = 0$ (hence no penalization is used), the offset and modeled coefficients will sum to the unpenalized GLM coefficient as the coefficient will be given full credibility:

$$\begin{aligned}\text{Unpenalized coefficient} &= \beta_{j,\text{offset}} + \beta_j \\ &= \beta_{j,\text{glm}}\end{aligned}$$

When λ is sufficiently high ($\lambda >> 0$), all variable components β_j will be removed from the model and our indicated coefficient will be $\beta_{j,\text{offset}}$:

$$\begin{aligned}\text{Fully penalized coefficient} &= \beta_{j,\text{offset}} + \beta_j \\ &= \beta_{j,\text{offset}} + 0 \\ &= \beta_{j,\text{offset}}\end{aligned}$$

As demonstrated in Section 3.1.3, the coefficient of a penalized regression model moves from a GLM coefficient to zero as the penalty term moves from zero to a sufficiently high number. In lasso credibility, we cannot separate the modeled coefficient from its accompanying offset. The graph in Figure 5.1 shows the paths of the combined offset and modeled coefficients in lasso credibility. Rather than shrinking to zero, our coefficient estimates collapse to our offset complement of credibility.

Therefore, for a general value of λ:[6]

$$\beta_{j,\text{offset}} \le \beta_{j,\text{offset}} + \beta_{j,\text{lasso}} \le \beta_{j,\text{glm}} \quad \text{if} \quad \beta_{j,\text{offset}} \le \beta_{j,\text{glm}}$$

$$\beta_{j,\text{offset}} \ge \beta_{j,\text{offset}} + \beta_{j,\text{lasso}} \ge \beta_{j,\text{glm}} \quad \text{if} \quad \beta_{j,\text{offset}} \ge \beta_{j,\text{glm}}$$

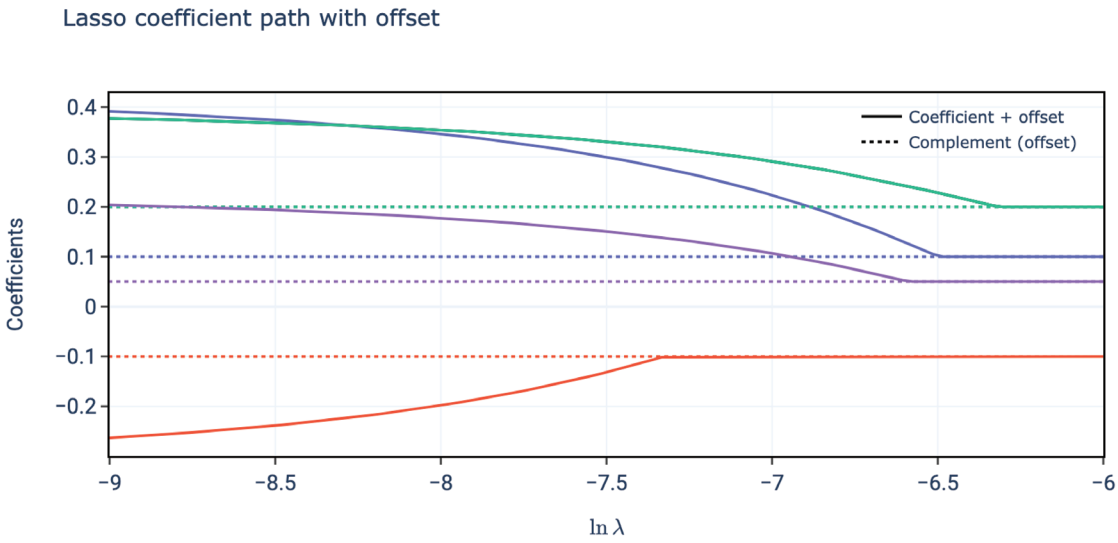
It follows that lasso credibility behaves as we would expect from a credibility procedure and there exists a Z such that

$$\text{Coefficient}_j = \left(\beta_{j,\text{offset}} + \beta_j\right) = Z \times \left(\beta_{j,\text{glm}}\right) + \left(1 - Z\right) \times \left(\beta_{j,\text{offset}}\right)$$

In this section, we have provided the intuition behind and practical effect of the relationship between penalized regression and credibility. In Appendix A we formalize this intuition and prove that penalized regression is a sound mathematical credibility procedure. Both Bühlmann credibility and penalized regression can be seen as examples of Bayesian estimation. Under a Bayesian interpretation, the connection between the two methodologies is evident. For the scope of this introduction, the reader can take

[6] The inequality is presented for illustrative purposes, and although it will hold in many situations, there are cases where it will not. In general, it will not hold when the coefficient will change sign through the coefficient path, which happens under high correlation. A comprehensive description of how the coefficient may change sign with varying λ can be found in Efron et al. (2004).

Figure 5.1. When Coefficients Collapse to Zero, the True Effect of the Variable Collapses to the Offset



for granted that penalized regression is a sound framework for applying credibility in a multivariate setting.

5.4. Terminology and ASOP 25

The terms in bold below are defined generally in ASOP 25's Section 2, "Definitions"—the definitions we provide in this list are meant to clarify how we apply the terms in the monograph going forward:

- **Risk characteristics:** the characteristics of the risk represented by our predictor variables
- **Risk classification system:** relativities assigned to predictor variables by the model
- **Subject experience:** experienced relativities in the modeling data set
- **Relevant experience:** the selected offset relativities *or* a 1.0 relativity
- **Credibility procedure:** lasso credibility (penalized regression)

Appendix B contains a more extensive description of penalized regression as a credibility procedure through the lens of ASOP 25. Possible complements of credibility will fulfill the requirements of ASOP 25's Section 3.3, "Selection of Relevant Experience." Additional guidance can be found in Boor (1996). Here are some examples of complements of credibility when using lasso credibility for a pure premium loss model:

- A 1.0 relativity (lasso credibility's default assumption)
- A prior loss model's relativities
- A countrywide loss model that includes the data being modeled as a small subset
- An existing rating plan
- Competitor relativities
- Industry relativities

If all variables use the default assumption, we would refer to the model as lasso penalized regression instead of lasso credibility. We can use this 1.0 assumption in lasso credibility by including no complement where there is no prior knowledge of a variable's effect on risk and including an appropriate complement for other variables where such knowledge exists.

5.5. Selecting and Evaluating a Penalty Parameter in Lasso Credibility

In lasso credibility, one can select a penalty parameter using the methodology described in Section 3.3. An actuary can reframe this process as testing the level of credibility that generalizes the best test statistic across the data set for all β. As before, cross-validation will select a single λ value, and the variation in optimal values of λ between model folds allows a modeler to make a credibility selection that deviates from this point estimate. This expert judgment applied when selecting a penalty parameter in lasso credibility is supported both through cross-validation statistical support as well as the considerations in ASOP 25's Section 3.4.

Here are some examples of an appropriate judgmental increase of the penalty parameter :

- If the complement of credibility consists of the current factors, the increase of the penalty parameter can be used as a methodology to select between current and indicated to mitigate policyholder impacts.
- If the model factors are showing instability, the penalty can be increased to provide additional stability. This is most common on small data sets or when using techniques like derivative lasso (Casotto and Holmes 2023) and AGLM (Fujita et al. 2020) that use penalization to create ordinal variable factor curves.
- If modeling data is adjusted for trend or incurred but not reported (IBNR) claims, or if case reserves are based on generic estimates, the volatility of the data may be understated. In this case, an increase of the penalty parameter may be appropriate to avoid overestimating the credibility of the modeling data.
- The increase is warranted if the modeler believes the selection produces a more actuarially appropriate model given all information available.

While cross-validation produces a statistical range of appropriate penalty terms both higher and lower than the best estimate, it is generally actuarial best practice to select only values of lambda higher than the point estimate. Such a higher estimate will be more conservative and is similar to selecting "between current and indicated" when using traditional actuarial methodologies. It is uncommon but supportable to decrease the penalty parameter using actuarial judgment. This may be appropriate in the following situations:

- When pricing a line of business where the complement is known to have some deficiencies

- When there is a known change of significant magnitude from the selected complement of credibility
- If the relevant experience is from an older or more out of date source and the actuary wants to give greater weight to the more recent subject experience

For additional considerations in adjusting the penalty parameter, refer to ASOP 25's Section 3.4, "Professional Judgment."

5.6. Calculating Indicated Rates in Lasso Credibility

Let's return again to our two-variable homeowners model. In this example, we are refitting the two-variable homeowners model using an old model's factor table output as a complement of credibility. We will rearrange the equation for our model predictions as we did in Equation 5.1:

$$\begin{aligned}\text{Prediction} &= \exp\left(\beta_0 + \log\left(\text{offset factors}\right) + \beta_1 X_1 + \beta_2 X_2\right)\\ &= \exp\left(\beta_0 + \beta_{1\,\text{offset}} X_1 + \beta_{2\,\text{offset}} X_2 + \beta_1 X_1 + \beta_2 X_2\right)\\ &= \exp\left(\beta_0 + \left(\beta_{1\,\text{offset}} + \beta_1\right) X_1 + \left(\beta_{2\,\text{offset}} + \beta_2\right) X_2\right)\end{aligned}$$

As a reminder, these are the coefficients from our prior model that we are using as a complement of credibility:

$$\beta_{1\,\text{offset}} = 0.182$$

$$\beta_{2\,\text{offset}} = 0.01$$

For now, we will ignore the recommendation that ordinal variables be used in place of continuous variables in lasso credibility. Let's say our new model fits with these convenient coefficients:

$$\beta_0 = 4.6057$$

$$\beta_1 = -0.087$$

$$\beta_2 = 0.01$$

While β_1 and β_2 are the coefficients output from the model, the indicated coefficients will be $(\beta_{1\,\text{offset}} + \beta_1)$ and $(\beta_{2\,\text{offset}} + \beta_2)$. The corresponding factor for not having a fire extinguisher would then be $\exp((\beta_{1\,\text{offset}} + \beta_1)X_1)$. If the coefficient β_1 was penalized out of the model, the model is giving full credibility to the complement and the indicated coefficient would be $\beta_{1\,\text{offset}}$. In this case, the indicated relativity would be identical to our prior model.

Our new indicated rating tables would be calculated as follows:

Base rate: $\exp(4.6057) = 100$.

Fire Extinguishers	Factor
No	$\exp((0.182 - 0.087) \times 1) = 1.100$
Yes	$\exp((0.182 - 0.087) \times 0) = 1.000$

Age of Home	Factor
0	$\exp((0.01 + 0.01) \times 0) = 1.000$
1	$\exp((0.01 + 0.01) \times 1) = 1.020$
2	$\exp((0.01 + 0.01) \times 2) = 1.040$
3	$\exp((0.01 + 0.01) \times 3) = 1.061$
4	$\exp((0.01 + 0.01) \times 4) = 1.082$
5	$\exp((0.01 + 0.01) \times 5) = 1.104$
6	$\exp((0.01 + 0.01) \times 6) = 1.126$
7	$\exp((0.01 + 0.01) \times 7) = 1.149$
8	$\exp((0.01 + 0.01) \times 8) = 1.172$
9	$\exp((0.01 + 0.01) \times 9) = 1.195$
10	$\exp((0.01 + 0.01) \times 10) = 1.219$

This example's calculations are simple because our current and prior models had an identical parameterization. In practice, the prior model's parameterization may not be known or the complement may be based on post-modeling selected rates. Lasso credibility does not require the knowledge of prior parameterization as the decomposition of the linear equation generalizes in all cases when using **ordinal** variables.

When the link is logarithmic, each coefficient β_j has an equivalent factor representation as $\exp(\beta_j)$. In this case, it is more intuitive to think about the indicated relativities in terms of **factors** than **coefficients.** For each item in a rating table, we will calculate the new indicated relativity by multiplying the offset relativity by the corresponding factor $\exp(\beta_j X_j)$ from the model. By Equation 5.1:

$$\begin{aligned}\text{Prediction} &= \exp\left(\beta_0 + \text{offset} + \beta_1 X_1 + \beta_2 X_2\right) \\ &= \text{intercept} \times \text{offset factor} \times \text{fire extinguisher factor} \\ &\quad \times \text{home age factor} \\ &= \text{intercept} \times \text{extinguisher offset} \times \text{home age offset} \\ &\quad \times \text{fire extinguisher factor} \times \text{home age factor}\end{aligned}$$

where

$$\text{Indicated fire extinguisher factor} = \text{extinguisher offset} \times \text{extinguisher indicated factor;}$$

and

$$\text{Indicated home age factor} = \text{home age offset} \times \text{home age indicated factor.}$$

This generalization has important implications for variable transformations and the building of a lasso credibility model. In penalized regression or a GLM, we are modeling the relationship of variables to a target. However, in lasso credibility, we are instead modeling the credible differences between a complement of credibility and a target. This difference is part of why in the next chapter we recommend applying an ordinal treatment to any variable previously considered continuous.

5.7. Lasso Credibility Conclusions

Lasso credibility is a powerful technique that can be used to enhance penalized regression by using the offset to implement a complement of credibility. As we will see in Chapter 7's case study, this enhancement allows lasso credibility to be used on data sets that are too small to build either a GLM or lasso penalized regression model. Where a GLM would be unstable and lasso penalized regression would shrink too many variables to zero, lasso credibility can reflect credible signal where available and shrink volatile experience instead toward an appropriate complement relativity.

The statistical components of this methodology—lasso penalized regression, ordinal variables, and the offset—are not new. What we have introduced in this section is simply a shift in perspective to make full use of the credibility-based nature of penalized regression. As actuaries explore additional data science methodologies, we expect that similar perspective shifts will allow other known tools to be used in innovative and actuarially sound ways.

6. Lasso Penalized Regression and Lasso Credibility Model Diagnostics

The modeler must review lasso penalized regression and lasso credibility models in a different manner than they would a GLM. Lasso models provide no *p*-values, and the evaluation of variables shifts from significance to credibility. Additionally, lasso credibility models require review of the complement of credibility as the complement can greatly influence the indicated coefficients or receive full credibility. New visualizations such as relativity plots and a switch from continuous to ordinal variables will greatly aid model building and model review.

6.1. Review of the Lambda Penalty Parameter

GLM validation relies heavily on the evaluation of *p*-values and standard errors, but lasso penalization does not provide such statistics. Methods that exist to create approximate *p*-values have not been largely successful (Casella et al. 2010) and are not recommended for model review as they can be misleading. Fortunately, as we described earlier, lasso penalization removes the need for post hoc significance testing through the application of partial credibility during the fitting process. Therefore, review of lasso credibility models should move from post hoc significance analysis to an evaluation of the lambda penalty term used to assign credibility.

- GLM review focuses on **significance**: Are we sure that a coefficient should be included?
- Lasso review focuses on **credibility**: How much can we trust the subject experience, if at all?

Table 6.1 compares treatment of variables in a GLM and treatment in lasso.

A properly selected penalty term automatically accounts for all considerations reviewed by *p*-value significance analysis through the application of credibility. The following questions are sufficient when reviewing the selected penalty term for appropriateness:

- Was the lambda parameter selected via cross-validation or another robust methodology?
- If there was an adjustment to the lambda penalty parameter, was the adjustment favoring a more robust model? Why was this adjustment made?
- Is the behavior of variables intuitive with the selected lambda penalty parameter?

Table 6.1. A Comparison of Variable Behavior in GLM and Lasso Credibility by Statistical Importance

Variable Importance	GLM	Lasso Credibility
Low	Subjective decision rule, i.e., remove when p-value > 0.05.	Automatically set to 0.0.
High	Full credibility is assigned to the observed relativity.	Credibility is assigned **more** to the **observed** than the complement of credibility.
Medium	Full credibility is assigned to the observed relativity during model fitting. Manual adjustments may be appropriate for p-values near to 0.05.	Credibility is assigned **more** to the **complement** of credibility than the observed experience.

The lasso penalty term provides a data-driven, uniform, and efficient method of addressing partial credibility. Lasso penalization eliminates the need for post hoc significance testing, hence significantly reducing the effort for model reviews. In the validation of lasso credibility models, a review of the complement of credibility is necessary.

6.2. Review of the Complement of Credibility

The review of a complement of credibility can be initially guided by the considerations in ASOP 25 and existing actuarial literature. We will not spend time on these traditional review criteria (i.e., similarities and differences between the subject and relevant experience, etc.) but will instead discuss considerations unique to lasso credibility and where a reviewer may want to focus their review depending on what is most material to their scope.

Lasso credibility is a multivariate procedure, and therefore we should consider correlations between risk characteristics that are being offset. Penalized regression natively assigns signal among correlated predictors, but using an offset can hinder the model's ability to detect shifts in signal between them. A reviewer should also pay special attention to correlated risk characteristics if their complements come from multiple sources. This combination of complements may provide an overestimated or underestimated prior assumption.

In addition to reviewing a complement of credibility by itself, modeling results will provide information on the appropriateness of the complement of credibility. To compare the complement of credibility to model output, we recommend the use of **relativity plots**. These plots will help to identify segments where a reviewer may want to spend additional time evaluating both the complement of credibility and the modeled results.

6.3. Relativity Plots

Traditional univariate charts are insufficient for lasso credibility model review. The indicated relativities from a lasso credibility model are deviations from a complement of credibility, and therefore it is essential to combine those relativities with the offset to see the true indicated coefficient. We recommend the use of relativity plots, which may contain the following items:

- The complement relativity (offset relativity)
- The indicated relativity (offset combined with modeled relativity)
- The observed relativity (optional)
- Exposures

See Figure 7.17 for an example of a relativity plot.

If one uses several transformations for a single variable, one needs to combine all such transformations to create the final indicated relativity. This combined relativity allows a reviewer to see how and where the model has indicated credible differences from the complement. By overlaying the observed values, we can see how reactive the model is to the experience in the data. Although lasso credibility is likelihood based rather than exposure based, benefit to likelihood correlates highly with amount of exposure. Including exposure bars allows a reviewer to see a representation of the amount of experience supporting a deviation from the complement of credibility. The combination of these four above elements makes relativity plots a helpful tool for lasso credibility model review.

6.3.1. Using Relativity Plots to Guide Model Review

The review of relativity plots focuses on the following question: Are the deviations from the complement of credibility stable and intuitive across all variables?

The question is not trivial as what is considered "stable" and "intuitive" varies between and within models. We first discuss how such definitions differ when the complement is receiving full, partial, or no credibility. Then we provide guidance on how to review stability and intuitiveness differently in categorical, ordinal, and continuous variables.

6.3.2. Full Credibility in the Complement

A reviewer could start by examining coefficients where the complement has received full credibility. Here are two scenarios where lasso credibility will assign full credibility to the complement:

1. When there is sufficient exposure to model but the experience is too similar to the complement of credibility to deviate
2. When there is insufficient data in a segment to pass the threshold of lasso credibility

The first scenario is ideal because the segment is credible and our model has penalized the modeled coefficient to zero. We can be confident that the estimate is reasonable

given the high credibility, and it is not recommended for a reviewer to spend significant time scrutinizing this coefficient.

In the second scenario, the complement is likely to receive full credibility—even if it is unreasonable. Even though this segment may not be a material portion of the book today, an inappropriate complement could expose an insurer to adverse selection, prohibit growth, or place an unreasonable burden on policyholders. Large deviations between observed and predicted (offset) are expected for these smaller segments and should not be used to justify that the complement is unreasonable, and similarly, small deviations should not be used to justify that the complement is reasonable. Instead, a reviewer should rely purely on traditional considerations when reviewing the complement when little data is available.

6.3.3. Partial Credibility in the Complement

A more common situation is that the complement will receive partial credibility. The predicted relativity will be somewhere between the observed relativity and complement relativity. Two common scenarios are as follows:

1. Where a specific segment has a medium amount of data that passes the threshold of significance but not enough to deviate significantly.
2. Where the entire data set is smaller than necessary to achieve full credibility.

Small differences between the complement and indicated relativities are most often an ideal result in a lasso credibility model. In a GLM, small coefficients are often scrutinized for low significance as it may be hard to reject the null hypothesis that the true coefficient is zero. Small coefficients should be far more accepted in lasso credibility as they have overcome the lasso credibility standard and reflect the assignment of partial credibility between the complement and experienced relativities. When deviations are proportional to the data in a segment and reasonable given the complement experienced relativities, the assigned credibility is likely appropriate.

We recommend an observer focus more of their time on segments with medium or small amounts of data that have large differences between the complement and indicated relativities. Large differences across many coefficients could indicate that the penalty term is too small and needs to be increased. Large differences in single categories or unexpected areas of an ordinal variable could suggest the presence of an outlier in the data that may need adjustment. When reviewing large deviations, remember that the same likelihood-based credibility standard is applied to all variables equally. It is possible that a large deviation is supportable based on the data—especially if that deviation is actuarially intuitive and other coefficients are behaving appropriately.

When the entire data set is smaller than necessary to achieve full credibility, a review of coefficient stability and intuitiveness is extremely important. If many coefficients have actuarially unintuitive deviations, that is a sign that the selected penalty term may be too small. Be careful not to reject unintuitive deviations outright as sometimes the real world does not behave as expected.

It is also helpful to think of this situation as relativities being pulled toward a complement or pulled toward the data's experience. We want this pull to be balanced toward the truth, and not over- or underreactive. A penalty term that is too small will potentially overshoot the true relativities while a penalty term that is too large will not sufficiently move toward the true relativities. Small data sets create the trickiest situation to review because both the complement of credibility and selected penalty term will have a large impact on model output. Lasso credibility is still worth it, however, and we show that it provides the highest benefit over other methodologies when modeling on smaller data sets in the case study (Section 7.6.2).

6.3.4. Limited or No Credibility in the Complement

Sometimes, our complement is receiving limited or no credibility because the segment has a large amount of data. For example, a large data set with a binary characteristic split 50/50 in data will see limited effect from all but the worst of complements. Large deviations from the complement should still be investigated, but a change of complement of credibility is unlikely to be material to the model output. If a segment is receiving full credibility with limited experience, it may be a sign that the penalty parameter is too low. If a segment is receiving full credibility with large experience, a reviewer can be confident in the estimates being produced without extensive review.

6.4. Review by Variable Type

When building a lasso credibility model, the modeler is seeking to identify the credible difference between the complement and the signal in the data. This is problematic, as often the location of that difference is unknown at the beginning of the modeling process! For categorical variables, we will see that a modeler has the freedom to use more granular categorical variables in lasso credibility than in a GLM or penalized regression. Continuous variables do not receive a similar benefit and are best substituted with an ordinal treatment of variables in lasso credibility.

6.4.1. Categorical Variables

Categorical variables are the most intuitive to review in lasso credibility.

In a GLM, categorical levels that are insignificant may be grouped with each other or with the base level to obtain a stable coefficient. In lasso credibility, we recommend avoiding the grouping of categories and instead maintaining categories that are at least as granular as the selected complement of credibility. When a level is insignificant, it will be collapsed to the complement. This assignment of full credibility to the complement removes the need to group a noncredible category with another category.

Each of the categorical levels can be evaluated using similar criteria to classical or Bühlmann credibility with the difference that any change to the penalty term will result in changes across all variables. The most important item to review in categorical relativity plots is whether the deviation is directionally appropriate and whether the magnitude of the change is properly responsive.

6.4.2. Continuous Variables

Continuous variables are difficult to conceptualize in lasso credibility as it is the **slope** that is penalized and not a categorical **magnitude**. Additionally, variable transformations no longer reflect the shape of the overall curve, but rather the shape of the **difference** from the curve of the complement of credibility. As we said earlier, the shape of that deviation is usually unknown and is in fact part of what a lasso credibility model is meant to discover.

Because of this, great care is needed when using continuous variables in lasso credibility. The example in Chapter 7 uses continuous variable transformations that are known to reflect the true distribution of the data, but that will never be the case in practice. Continuous variables are used later only so that our GLM and lasso credibility models both start with the same information and can be more easily compared. In practice, we **highly** discourage the use of continuous variables in lasso credibility.

Review should focus first on the ability of the continuous transformation to accurately capture potential differences from the complement. For example, if a linear variable alone is included, the only difference the model will identify is a change in overall slope and not a hinge or change in slope after a certain value.

Second, review can focus on the potential extrapolation of continuous variables to levels where there is minimal credibility. One should investigate whether an indicated change continues to grow in magnitude in the tails of a continuous variable—especially if the experience in the data does not strongly support such a change. This could be a sign of unintuitive extrapolation, and the variable will need to be adjusted.

Again, we use continuous variables in our case study in Chapter 7 only because the true form of the deviation from the complement of credibility is known. Similarly, we would recommend the use of continuous variables in lasso credibility only if one knows that a change in slope will properly capture the deviation from a complement of credibility. We cannot provide an example of this situation in a real application.

6.4.3. Ordinal Variables

On the other hand, ordinal variables are easy to review in lasso credibility and can be quite useful in determining an appropriate penalty term. If an ordinal variable contains reversals (up for age 20, down for 21, up for 22, etc.), the selected penalty term is likely too low. A common method of selecting a credibility standard is to start with the indicated penalty term and gradually increase that term until all unintuitive reversals are removed from the indicated relativities. In most situations, this results in both an actuarially and statistically sound model.

The tails of ordinal variables also have the benefit of not extrapolating beyond what is indicated by the data as an uncapped continuous variable will extrapolate. At the same time, ordinal variables may not extrapolate where a modeler believes it is justified. Ordinal variables force a modeler to choose and justify an extrapolation as opposed to allowing a selected variable transformation to provide "support" for a relativity beyond what may be statistically justified. For this reason, judgmental adjustments and

extrapolation at the tails of ordinal variables may be appropriate when using both lasso credibility and traditional lasso penalization.

6.4.4. Control Variables

Whether or not to apply a complement for control variables is a judgmental decision, and both options can be motivated by different arguments.

It is reasonable to include a best estimate complement of credibility for control variables when using lasso credibility. For example, each state's prior overall rate relativity could be used as an offset in a loss model. With this complement, coefficients for individual state categories will correct for any significant difference between the estimate and experienced relativity. Determining this complement is difficult for a factor such as "year," which can include effects from trend, development, and changes in legal environment.

It is also reasonable not to include offsets for control variables. As with traditional penalized regression, if the model views the effect of a control variable as insignificant, that insignificant signal is unlikely to have a material effect on other variables. Although a modeler should continue to use control variables in lasso credibility, it is up to them to decide whether using a corresponding complement of credibility is material and necessary.

6.5. Model Validation Conclusions

By moving away from the full credibility assumption of GLMs, lasso credibility model validation changes from a review of significance to a review of credibility. The change is motivated first by the benefits of penalized regression through the bias–variance trade-off and second by the ability to introduce bias toward a selected complement of credibility. The change in focus greatly simplifies model review.

The statistical and actuarial review of a lasso credibility model is quite straightforward. If the complement is appropriate and all variables are treated as categorical or sufficiently granular ordinal variables, the only item left to review is the penalty parameter. If the parameter was initially selected in a statistically sound manner (cross-validation) and adjusted in an actuarially sound manner (adjusted higher to produce actuarially sound relativities as reviewed in relativity plots), we suggest that the model is a proper statistical application of lasso credibility. Review can focus mostly on post-modeling selections and any application of further actuarial judgment.

Given the simplicity of the ideal application of lasso credibility, it is difficult to build a deficient model without some clear red flags in the test statistics or relativity plots. Questionable behavior of a lasso credibility model should be further evaluated through a combination of penalized regression and traditional credibility expertise.

The next chapter consists of a case study that provides examples of both good and poor lasso credibility models. We hope that these examples can help readers become more comfortable with the processes of building and evaluating lasso credibility models.

7. Case Study

From the previous chapters, we hope the reader has gained a basic understanding of the application of lasso penalized regression as an actuarial credibility procedure. In this chapter we offer a practical application to demonstrate how modeling methods may evolve with lasso credibility. To do so, we replicate a **pricing model refresh** project that is common in the United States. The case study will walk through a model refresh process using generalized linear modeling and lasso credibility in parallel. We identify the differences between the approaches and point out key concepts of lasso credibility along the way.

The case study is not intended to prove that lasso credibility is better than other methodologies in all cases. As the data is simulated, one could select a seed to create favorable or unfavorable model comparisons. Instead, we intend the case study to demonstrate how lasso credibility truly acts as a multivariate credibility procedure with the benefits of both traditional credibility procedures and multivariate penalized regression. Deviations from this methodology are expected based on an insurer's unique situation, and guidance for such deviations should come from both materials on penalized regression and ASOP 25 credibility procedures.

Our hope is that through this monograph, the accompanying case study, and ASOP 25, the reader can gain the necessary understanding and skills to apply lasso credibility effectively in practice.

Accompanying code is provided on the CAS GitHub,[7] together with additional exercises for the reader. One need not do the exercises to understand the basic concepts of lasso credibility, but we highly encourage readers to investigate items that they feel could be relevant to their existing practices. Actuaries working on their coding skills may find such items to be practical exercises for self-improvement, and we encourage code contribution via pull requests to enhance the accompanying code.

7.1. Countrywide Modeling and State Refits

A model refresh involves creating one or more models to develop new rating relativities for an insurer's rating plan. It is common practice for U.S. insurers to first fit a model to a large data set, including all relevant data available for the line. For example, an insurer might build a model using all claims data from the last five years' experience

[7] https://github.com/casact/mg-credibility

for a given line of business, across all states. Such a model is referred to as a **countrywide model**.

The countrywide model may be quite robust if a large amount of data is available. The individual states, however, have unique regulations and varying behavior of risk characteristics. Because of those differences, an insurer may want to adjust the robust countrywide model for each state based on its available data. The process could result in the creation of 51 different rating relativities—one for each state and Washington, D.C. For simplicity, we use three main categories to represent the states:

- **Small state.** Such a state offers the modeler insufficient data with which to build a stable GLM, and therefore the modeler has no choice but to **adopt the countrywide model**.
- **Medium state.** A state in this category has potentially sufficient data for individual modeling but the modeler **usually adopts the countrywide model with post-modeling adjustments** for specific state characteristics. Building a model for a medium state may be relatively more time-consuming due to data variability, and the decision between a new model or adjusting the countrywide model is often based on a cost–benefit analysis. Post hoc univariate adjustment of a model, as mentioned earlier, is often suboptimal versus multivariate approaches.
- **Large state.** Such a state has sufficient credibility for the modeler to rely on the state's own experience to **build a GLM** if desired.

The full credibility nature of a GLM does not allow a modeler to easily blend countrywide and state-specific experience. Instead, the modeler must choose between adopting the countrywide model, refitting a GLM from scratch, or performing manual ad hoc adjustments. On the other hand, lasso credibility allows a modeler to blend the state's experience with a complement of credibility provided by the countrywide model during the modeling process. One can expect the credibility assigned to the state-specific experience to vary with the state's size, being higher for larger states and lower for smaller ones.

7.2. Case Study Summary

The case study starts by building a countrywide model from scratch as we assume that there is no complement of credibility that this insurer trusts more than its own data for this model. In our example, the countrywide data set is quite credible and the differences between lasso penalization and GLM are minimal.

Differences begin to appear when looking at the **large state level**, where we will fit a GLM, a lasso penalization model, and a lasso credibility model. The lasso penalization model will still perform similarly to the GLM on this data set. However, we will see that lasso credibility, with the countrywide model as a complement of credibility, can outperform both models due to the extra information included in the complement.

Lasso credibility can also be the best-performing model on the **medium state subset**. In this example, we will detail and discuss the role of a judgmental increase of the penalty parameter when building a lasso credibility model.

The **small state** by our definition is one that cannot be fit reliably with a GLM, making it impossible to compare lasso credibility models to GLMs. Instead, the countrywide model is used as a comparison. We detail two possible behaviors of the small-state use case. First, we consider a small-state data set that has material differences from the base modeling data set. We show that not only is it possible to build a lasso credibility model on this small data set, but the **lasso credibility model outperforms the countrywide model**. Then, we consider a second small-state data set that has an identical underlying distribution to the base modeling data set. In this case, **the lasso credibility model collapses to the countrywide model**. The lasso credibility model is able to correctly identify that the selected complement is a good representation of this data subset.

We focus on the countrywide modeling and state refit use case because it covers the entire range of credibility—nearly full, partial, and little/no credibility. This range of credibility is also applicable to a broader set of scenarios. For instance, data could be divided by time periods or other definitions rather than by state. Similarly, the complement of credibility could be the current filed and implemented model instead of a new countrywide model. We invite the reader to use the guidance of ASOP 25 to identify additional applications of lasso credibility in their own practice.

7.2.1. Data Description

The data generated for the case study is a synthetic commercial auto data set containing 3,500,000 total records. We show the risk characteristics of each record in Table 7.1. Those characteristics are assigned using a distribution that mimics a probable real-life distribution on a univariate level, but with no correlation between characteristics. For each value of a risk characteristic, a **true risk relativity** is assigned. **True risk relativities** are consistent within state-level subsets but **may be different between states,** as visualized in Section 7.2.2. Additionally, the true base rate varies within each state-level subset.

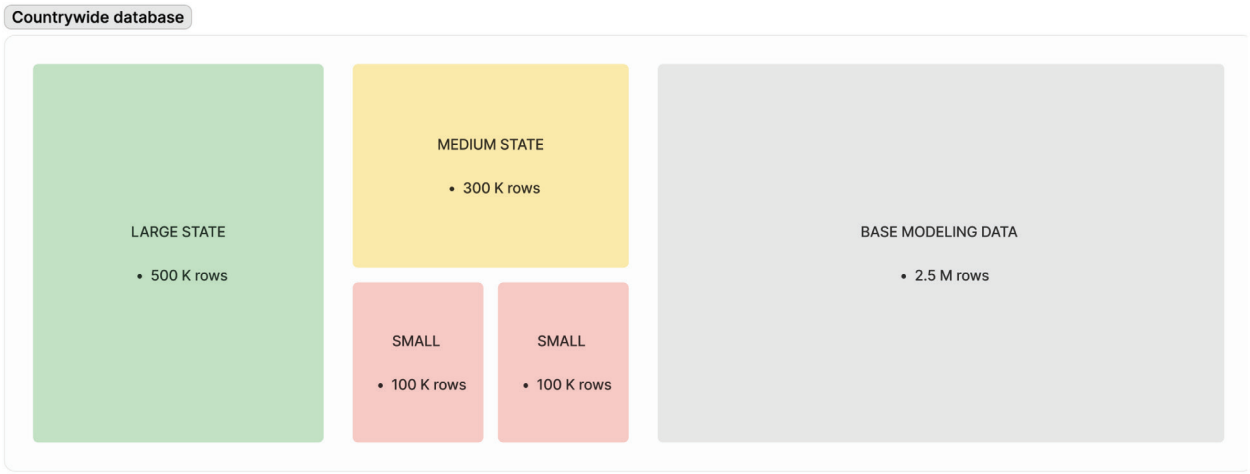
A **true pure premium** is generated for each risk by multiplying the true base rate by the true risk relativities assigned to each characteristic. We then simulate an **experienced pure premium** using a Tweedie distribution.[8]

Figure 7.1 illustrates the division of the data across five subsets, corresponding to states of various sizes:

- Base modeling data: 2,500,000 records
  - The largest part of the modeling data set is referred to as the base modeling data set. In practice, this segment would be made up of a combination of other large, medium, and small state subsets. For simplicity, we simulate this all at once.
  - This data set is included so that the countrywide model is highly stable and can be used as a complement of credibility for the smaller models. We will demonstrate

[8] The Tweedie random samples were simulated with a p-parameter of 1.6 and a \$% of 800. These parameters were chosen so that the resulting frequency is approximately 4%.

Figure 7.1. Visual representation of the various subsets in the simulated data. Risk characteristics are consistent within each subset but may (slightly) vary across states.



K = thousand; M = million.

that for this large data set, GLM and lasso penalized regression have very similar results. We do not model this data set on its own, but instead combined with all other data.

- Large state subset: 500,000 records
  - This subset represents the insurer's largest state in the U.S. market.
  - We model this data set using a GLM, lasso penalization, and lasso credibility. We will see that lasso credibility can be an improvement where data is large enough to build a stable GLM.
- Medium state subset: 300,000 records
  - This subset represents one of the insurer's growing states.
  - We build only a GLM and a lasso credibility model on this data set. We will see that lasso credibility provides more benefit on this data set than on the large state data.
- Small state 1: 100,000 records
  - This small-state data represents a subset containing minimal data.
  - This subset has **different risk relativities** than the base modeling data set.
  - We use this subset to show that, even with small data, lasso credibility is able to identify meaningful changes.
- Small state 2: 100,000 records
  - This second small state also represents a subset containing minimal data.
  - This data has **the same underlying risk relativities** as the base modeling data set.
  - We use this to show that a good complement creates a sparse lasso credibility model.

7.2.2. Predictor Variables

The generated data set contains seven risk characteristics, summarized in Table 7.1.

Each of our predictor variables is included to highlight a different scenario when comparing generalized linear modeling and lasso credibility. All charts reflect the exposure distribution and true risk relativity for each of the subsets of the data.

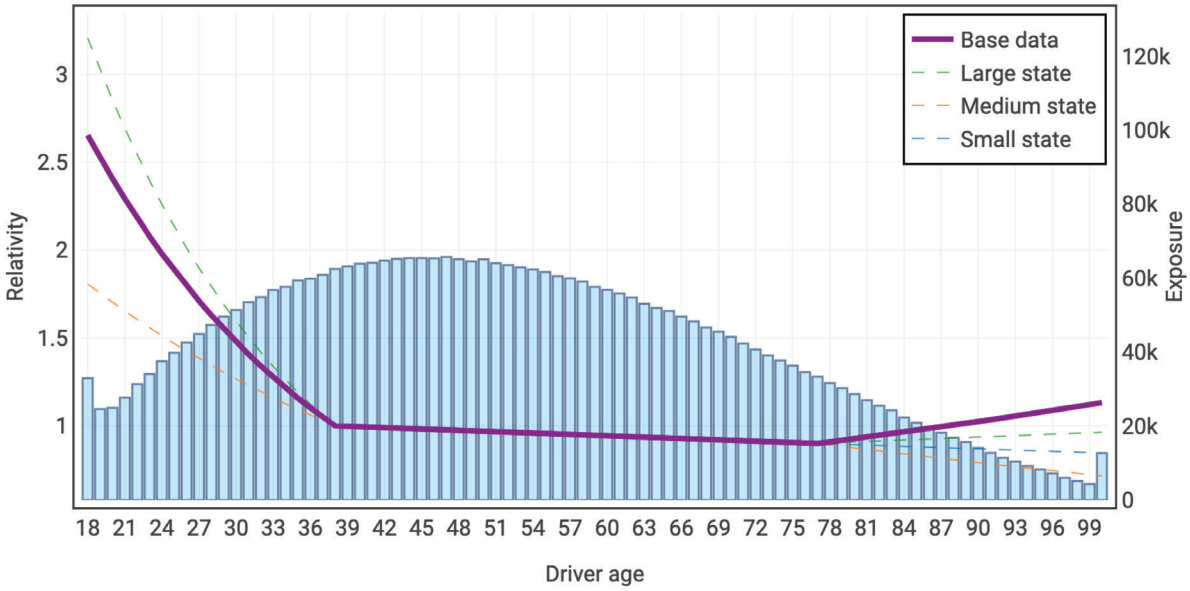
Table 7.1. List of Risk Characteristics in the Synthetic Database

Name	Type	Values
Driver age	Numerical	[18–100]
Vehicle age	Numerical	[0–19]
Industry code	Categorical	[Education, . . . , Fireworks]
Vehicle weight	Categorical	[Extra-Light, Light, Medium, Heavy]
Multipolicy discount	Categorical	[Yes, No]
x-Treme turn signal	Categorical	[Yes, No]

Driver Age

As we will see, lasso credibility does not prevent a model from identifying the steep increase for young drivers, and it can bring stability to the low-data tail beyond age 76 (Figure 7.2).

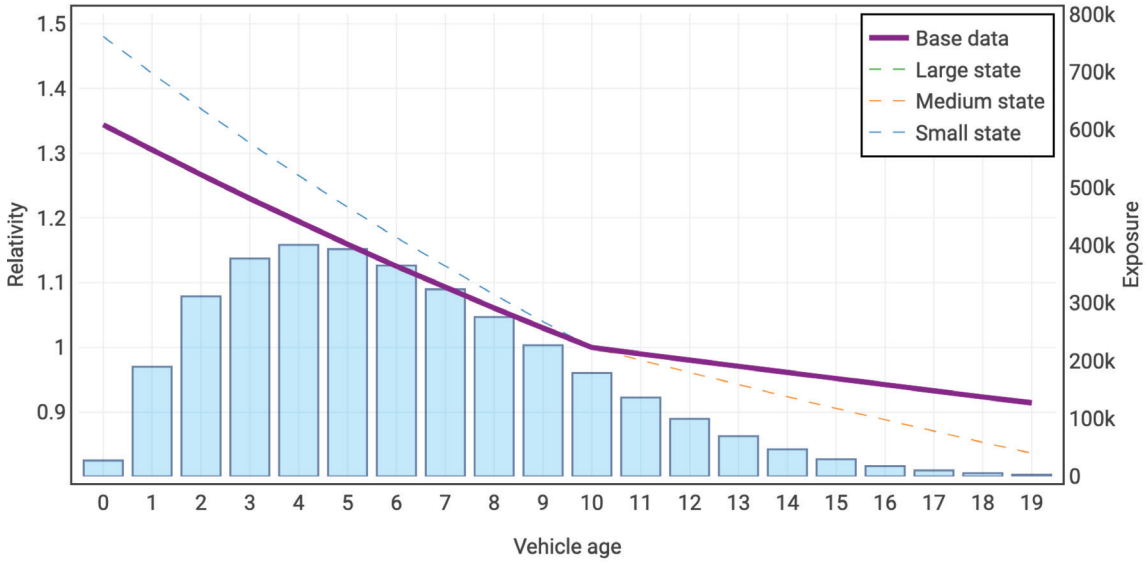
Figure 7.2. Driver age demonstrates that not all transformations within a continuous variable receive the same amount of credibility. This may seem trivial, but it is highly important to demonstrate that we cannot accurately rely on or even meaningfully calculate a credibility score (Z) when describing a lasso credibility model's behavior.



Vehicle Age

Figure 7.3 shows the risk relativities for the age of an insured's vehicle.

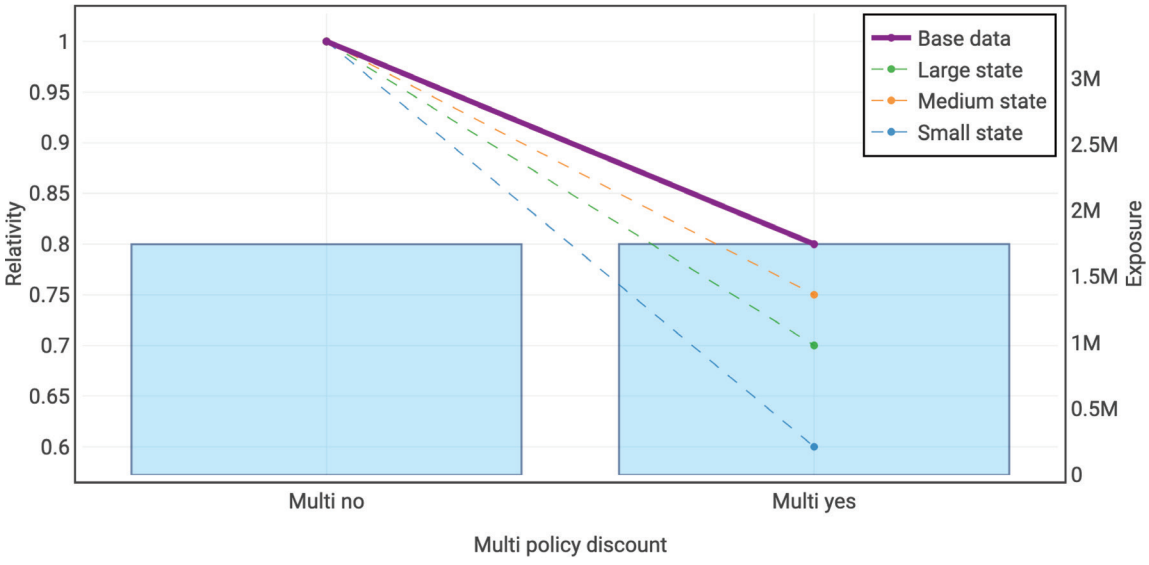
Figure 7.3. Across all data, this variable's true risk decreases from 0 to 10, then decreases at a less steep rate beyond 10. We will see that coefficients are highly stable from vehicles ages 0 to 10 and then become increasingly harder to model for older vehicles.



Multipolicy Discount

This indicator represents whether the insured has another policy with the insurer (Figure 7.4).

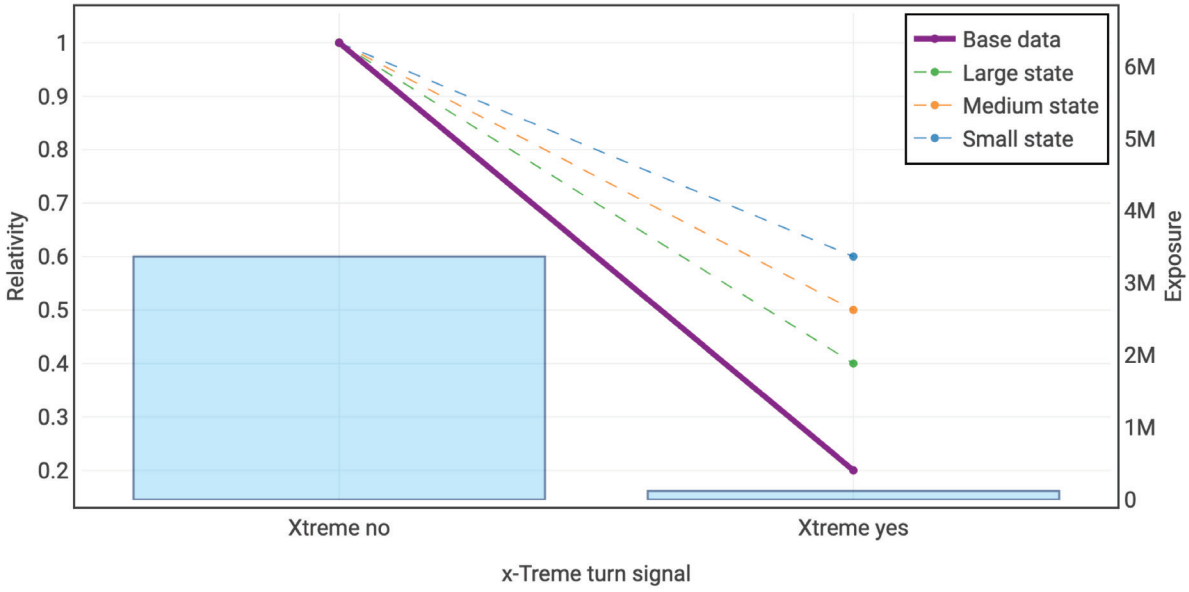
Figure 7.4. The "Yes" category always has a lower true risk relativity. We will see that multipolicy discount is relatively well predicted across all data sets because both levels have significant data and material signal.



x-Treme Turn Signal

This a fictitious new safety feature is in the early stages of adoption and greatly reduces accidents (Figure 7.5).

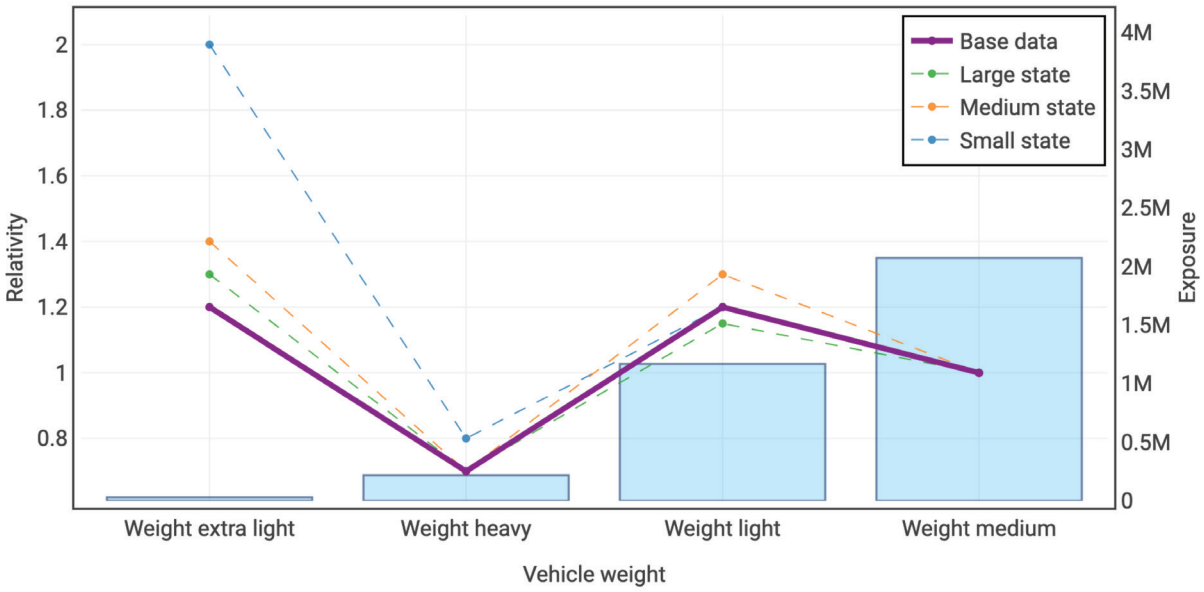
Figure 7.5. x-Treme turn signal has a very low relativity that varies between data sets. We will use this example to show that lasso credibility can react—without overreacting—to strong signal in small segments.



Vehicle Weight

This is a categorical variable with four categories: extra-light, light, medium, and heavy (Figure 7.6). While we are modeling vehicle weight as a categorical variable, we want to point out that vehicle weight could also be modeled as an ordinal variable. Unlike continuous variables, ordinal variables can reflect a link between adjacent categories without enforcing a numeric distance between these categories.

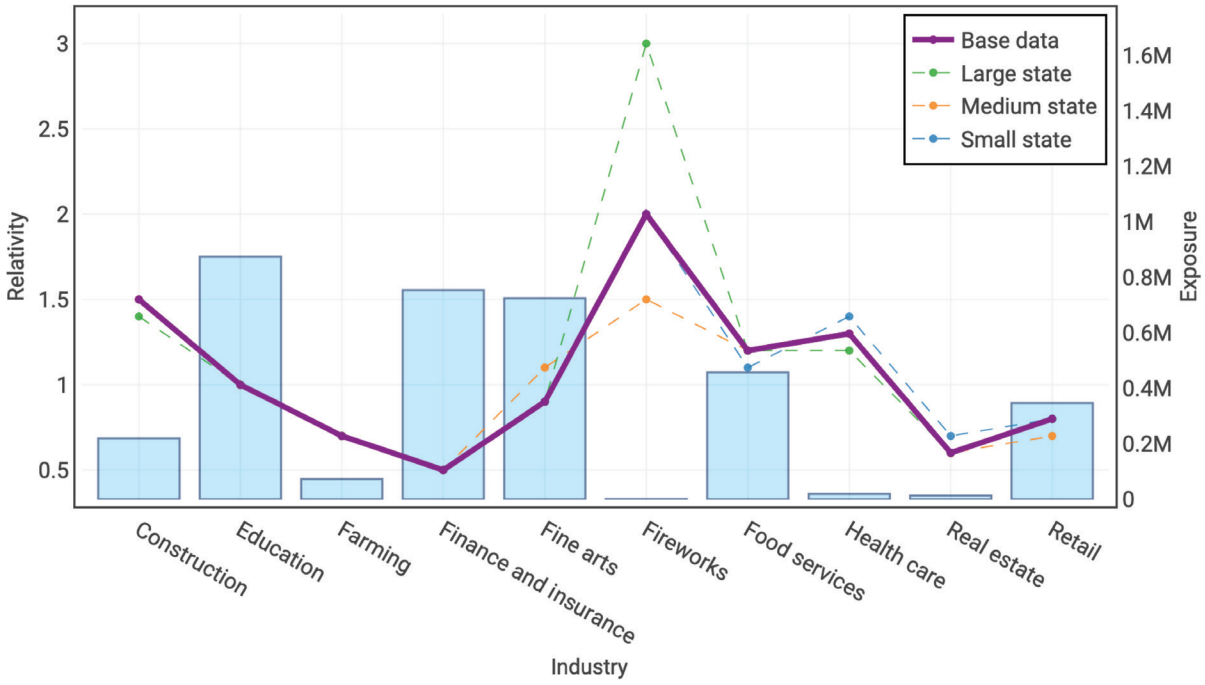
Figure 7.6. Risk increases as a vehicle becomes lighter, and may plateau with identical relativities for extra-light and light vehicles. Extra-light vehicle weight is very hard to predict due to its low exposure, but in most data sets it has a higher relativity than light vehicles. We will see that our complement of credibility allows us to model low credibility categories without grouping them together.



Industry Code

The industry code category is separated into 10 categories that are not easily consolidated (Figure 7.7).

Figure 7.7. Industry code provides a side-by-side example of categorical variables of various discounts, surcharges, and data sizes. The fireworks industry code is known to require a high surcharge but has extremely small data. We suggest that since none of these categories is easily grouped with another it would be best to predict a unique relativity for all categories.



7.2.3. Methodological Notes

The Use of Simulated Data

Working with synthetic/simulated data comes with its own advantages and disadvantages. We acknowledge that simulated data amounts to a strong simplification versus using an open, realistic data set, as, for example, in Wüthrich and Merz (2023) or Casotto and Holmes (2023). We still decided to work with simulated data because of the following benefits:

1. We know the true underlying relativities for risk characteristics, and therefore have knowledge of what is signal and what is noise. As a result, we can create charts comparing our model output to the true risk relativities instead of potentially noisy validation data.
2. We know that our risk characteristics are uncorrelated. Correlation will not affect the stability of either our GLM or lasso credibility models.
3. We know the variable transformations that can capture the true risk relativities. Such transformations would allow a model to fit perfectly to the true risk relativities

given sufficient data. This greatly simplifies the feature engineering process as we will use these known transformations for all models. We acknowledge that additional feature engineering would improve all models, but we think that the simplifying assumption is the best way to explain lasso credibility for instructional purposes.

4. The case study can correctly highlight situations where lasso credibility is performing well versus situations where it is performing poorly, and why the model exhibits this performance.
5. The underlying data can be resimulated for additional investigations.

Model Types

The generalized linear, lasso penalization, and lasso credibility models are built as follows (if not otherwise indicated):

- The GLM is built using the same feature engineering as the one used to generate the data.
- Lasso penalization uses the same feature engineering as the GLM. The penalty parameter λ is selected using cross-validation. The coefficients are standardized prior to the fit.
- Lasso credibility models are built identically to lasso penalization, but with the addition of a complement of credibility through the offset. The HDtweedie package used in the code does not directly support offsets, so the offsets are included by manually adjusting the response and weight columns through the methodology outlined in Shi (2010).

7.2.4. Prediction and Relativity Plots

We do not examine traditional test statistics like Gini or Tweedie deviance but instead focus on a direct comparison to the true relativities. This choice is made because the goal of the case study is to understand the behavior of lasso credibility, and that behavior is best represented as a visual representation of the estimated, simulated, and true relativities to provide insights on a variable-by-variable basis. By understanding the behavior of lasso credibility, a modeler can better understand why a model's test statistics will improve or degrade in various situations.

To efficiently compare true and modeled relativities and predictions, a relativity plot and a prediction plot are provided for all modeled variables.

The **relativity plot** contains the predicted GLM relativity, the lasso penalized regression relativity, and the true relativity. All relativities have been rebalanced to the base level for categorical variables, to age 38 for driver age, and to age 10 for vehicle age. Such relativity plots provide us with better validation metrics than comparing on holdout data because we can compare directly to the true relativities. The model with relativities closest to the true relativity is the better model at predicting on unseen data.

The **prediction plot** contains the true pure premium, the experienced (simulated) pure premium, the GLM prediction of pure premium, and the lasso prediction of pure premium. The prediction plot is not rebalanced for the full modeling data plots

but is rebalanced for the large, medium, and small state models. Rebalancing is done by multiplying the predictions by a constant such that the average prediction is equal to the average observed value. The difference between the experienced (simulated) pure premium and the true pure premium is determined by the random simulation of Tweedie distribution. Comparing these two quantities gives us a sense of how much noise is in the data. For example, if the true and experienced pure premiums are very close, the data is not noisy. If the true and experienced pure premiums are far apart, the mean of the simulated data is not similar to the true mean, and therefore our simulation has introduced noise.

These charts can be busy with overlapping items, so we recommend pulling the code and generating the charts in R so that you can click on items in the chart's legend to toggle their visibility on or off.

Double **lift charts** are provided to compare models. Records are ordered by the ratio of the models being compared and then bucketed into 10 deciles. Then, the ratios of the predictions to the true pure premium are plotted. The model that is closest to the horizontal line (a 1.0 ratio) is considered the best model.

7.3. Countrywide Model Results

Our countrywide models are built on the full 3,500,000-row data set. The GLM is quite stable, and our penalized regression model is only slightly penalizing coefficients. As expected, applying a credibility procedure to a large data set does not result in a large amount of weight being put on the complement of credibility.

7.3.1. Large Data Approaches Full Credibility

The coefficient chart (Table 7.2) contains GLM coefficients and lasso coefficients. All lasso coefficients are shrunk very slightly toward zero. When categories are sufficiently populated with stable experience, the shrinkage from penalized regression will have limited effect on the modeled coefficients. In large data, even smaller categories may approach full credibility.

The health care and fireworks industry codes are the most interesting results. Our GLM's p-values assign high significance to these categories, and they are treated as nearly fully credible in penalized regression. Relativity plots show that both models are overpredicting the relativity for these segments. Prediction plots show why: our simulated losses were particularly high in these categories. Neither GLMs nor penalized regression can identify when actual experience is unlucky (or lucky) and different from the true underlying risk.

7.3.2. Additional Exercises—Full Data

We recommend the following exercises to explore variable transformations and their effect on the shrinkage of continuous variables.

- Move the hinge point of vehicle age to see how the significance and shrinkage of the new variable transformations change from the current transformations.

Table 7.2. Comparison of Coefficient Results between GLM and Lasso Penalization

Variable	GLM Coefficient	Lasso Coefficient
(Intercept)	8.12177	8.12918
driver_age_18_38_hinge	–0.04886	–0.04886
driver_age_38_76_hinge	–0.00213	–0.00209
driver_age_76_99_hinge	0.00685	0.00667
ind_construction	0.37839	0.37833
ind_farming	–0.38933	–0.38591
ind_finance_and_insurance	–0.70583	–0.70410
ind_fine_arts	–0.09941	–0.09788
ind_fireworks	1.03998	1.02833
ind_food_services	0.18639	0.18654
ind_health_care	0.33816	0.33555
ind_real_estate	–0.50757	–0.50066
ind_retail	–0.19828	–0.19636
multi_yes	–0.25957	–0.25890
vehicle_age_0_10_hinge	–0.03199	–0.03192
vehicle_age_10_99_hinge	–0.01697	–0.01675
weight_extra_light	0.18419	0.18067
weight_heavy	–0.33246	–0.33117
weight_light	0.18095	0.18030
xTreme_yes	–1.33088	–1.32796

- Test different polynomial terms for driver age and vehicle age—do these terms behave similarly in GLM and lasso models?
- Replace the hinge terms with an ordinal encoding as referenced in Section 3.4.3. Include all of these variables in your lasso model. How do the predictions change? For example:
  - driver_age_over_18: 0 for age 18, 1 for all ages over 19
  - driver_age_over_19: 0 for ages 19 and below, 1 for all ages 20 and above
  - driver_age_over_20: 0 for ages 20 and below, 1 for all ages 21 and above
  - etc.

7.3.3. Full Data Conclusion—Lasso Penalization, but Not Lasso Credibility

Both models are fitting similarly to the data overall, and we see in Figure 7.8 and 7.9 that output coefficients are immaterially different. As expected, both models are performing similarly on the validate data in Figure 7.10 and Figure 7.11.

Figure 7.8. Both the GLM and Lasso Models Produce Similarly Accurate Predictions on This Large Data Set

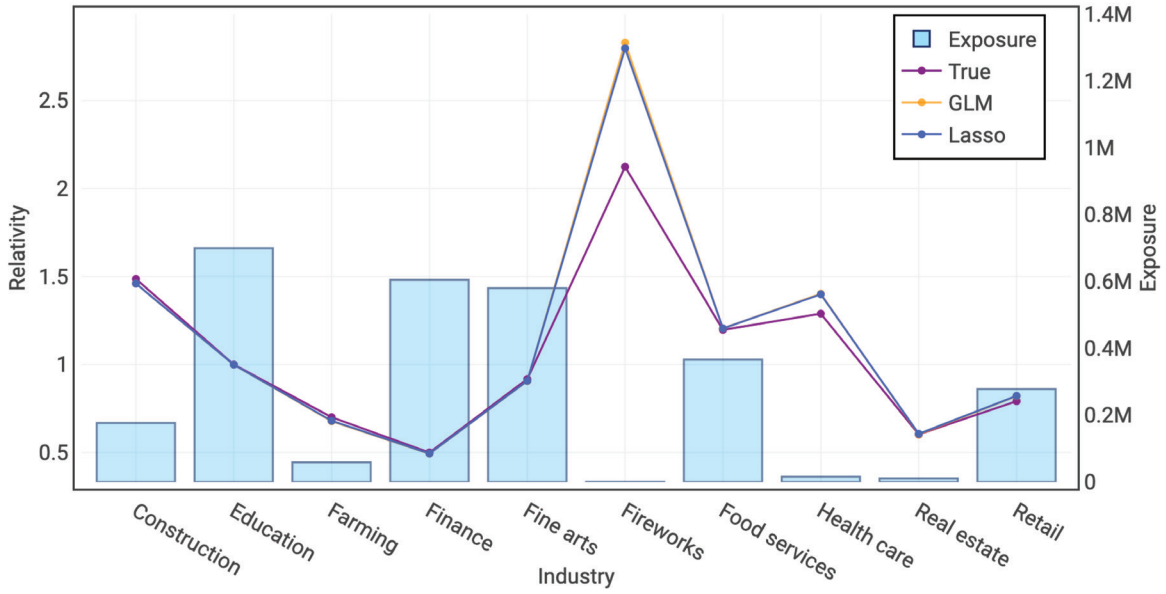


Figure 7.9. Neither Model is Able to Identify Situations Where the Experienced Pure Premium is Slightly Different Than the Experienced Pure Premium in Health Care and Real Estate

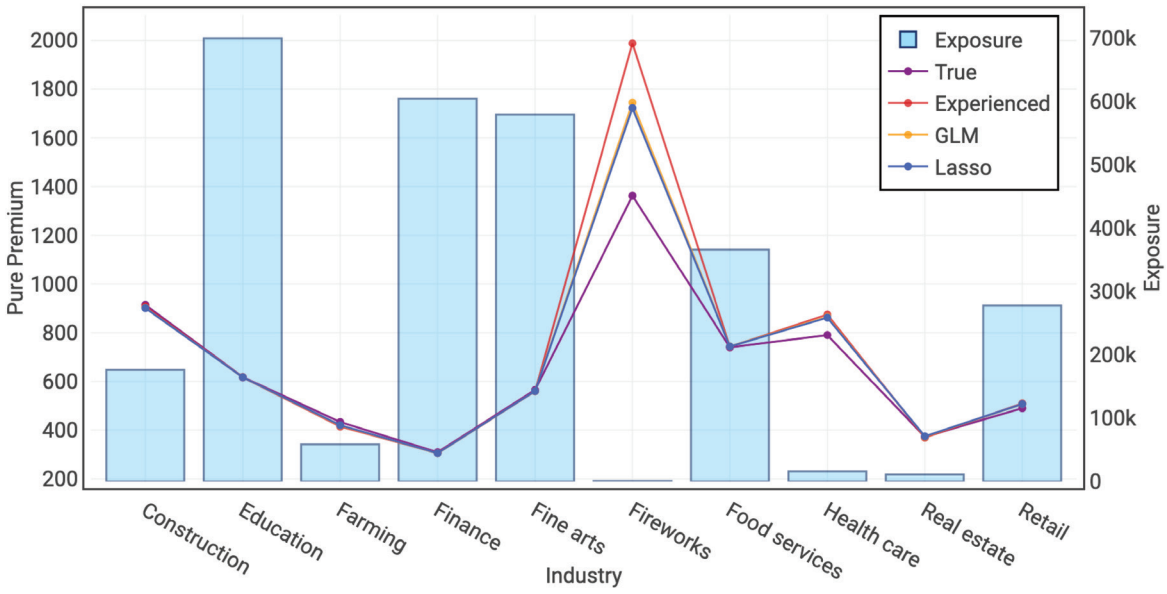
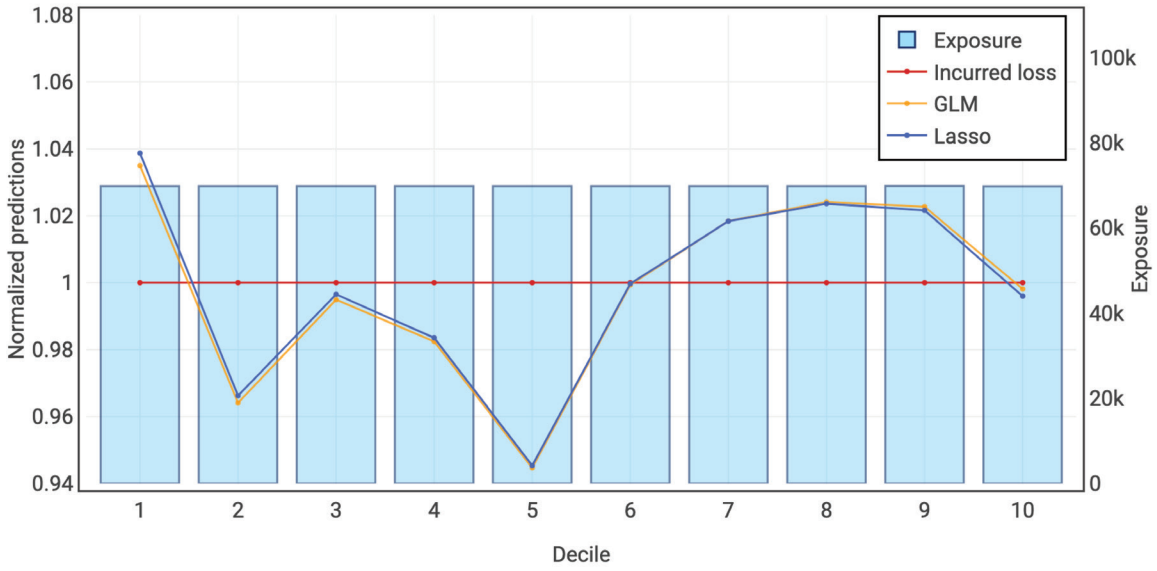


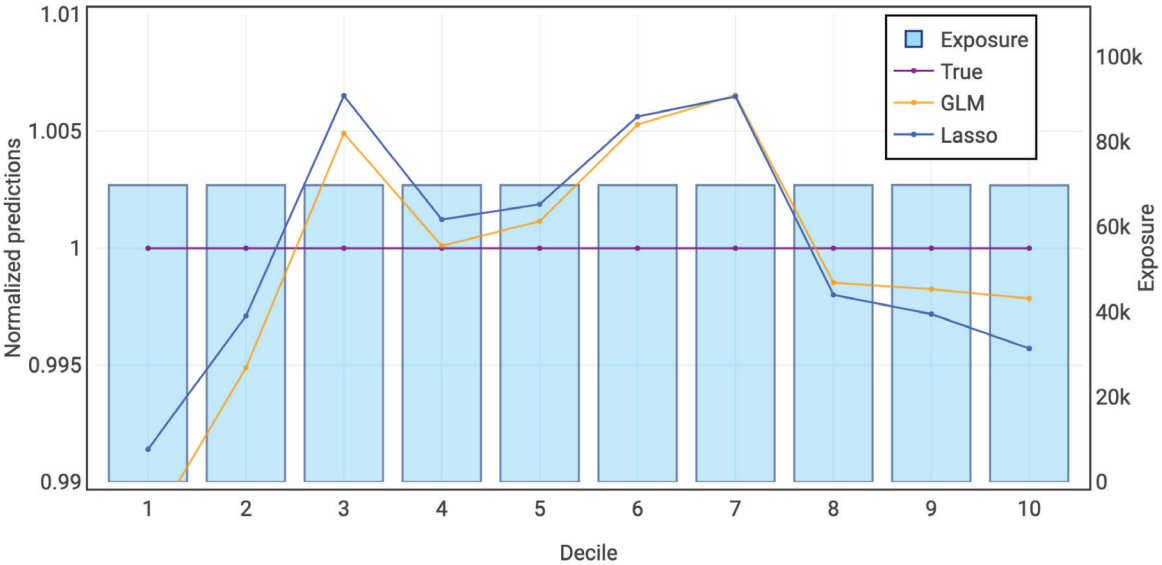
Figure 7.10. Both Models are Doing a Similar Job of Predicting the Simulated Losses in Our Validate Data Set



Because we know the true pure premium for each risk, we can create a lift chart using true values instead of simulated values. This lift chart is significantly less noisy and will be useful when evaluating model results on smaller data sets.

The penalized model output could be used to support rating factors identically to GLM output, but we would not refer to this use of lasso penalization as an application of lasso credibility. Instead, the model is behaving appropriately as a traditional

Figure 7.11. Both lasso and the GLM are doing an excellent job at identifying the true relativities overall in this large data set. This double lift chart is calculated on the validation set and the scale is extremely small.



use of penalized regression. To call it lasso credibility, we would have to justify that the selected default complement of credibility for all variables (a 1.0 relativity) is an actuarially sound complement of credibility. For young drivers and the fireworks industry code in particular, this would be a poor complement.

7.4. "Large State" Modeling Results

We build three models on the large state data:

1. A standard GLM model
2. A lasso penalization model
3. A lasso credibility model, using the previously fitted countrywide model as complement of credibility

We begin to see instabilities in the GLM, and those instabilities are automatically accounted for in our penalized regression model. The penalized regression model will outperform the GLM by shrinking coefficients when they are not sufficiently supported by the data.

7.4.1. Low Significance Correlates with High Shrinkage

Now that our data is smaller, some GLM coefficients have *p*-values above the 5% threshold. This is the case for the tail of the driver age variable in Figure 7.12 (driver_age_76_99_hinge) and the vehicle age variable in Figure 7.13 (vehicle_age_10_99_hinge), as well as the health care indicator for the industry code (ind_health_care). A modeler would have to remove these factors and try again with different variable transformations.

Figure 7.12. The Hinge for Older Drivers is Shrunk Toward Zero Slope in the Lasso Model, Not Reacting as Much as in Our GLM

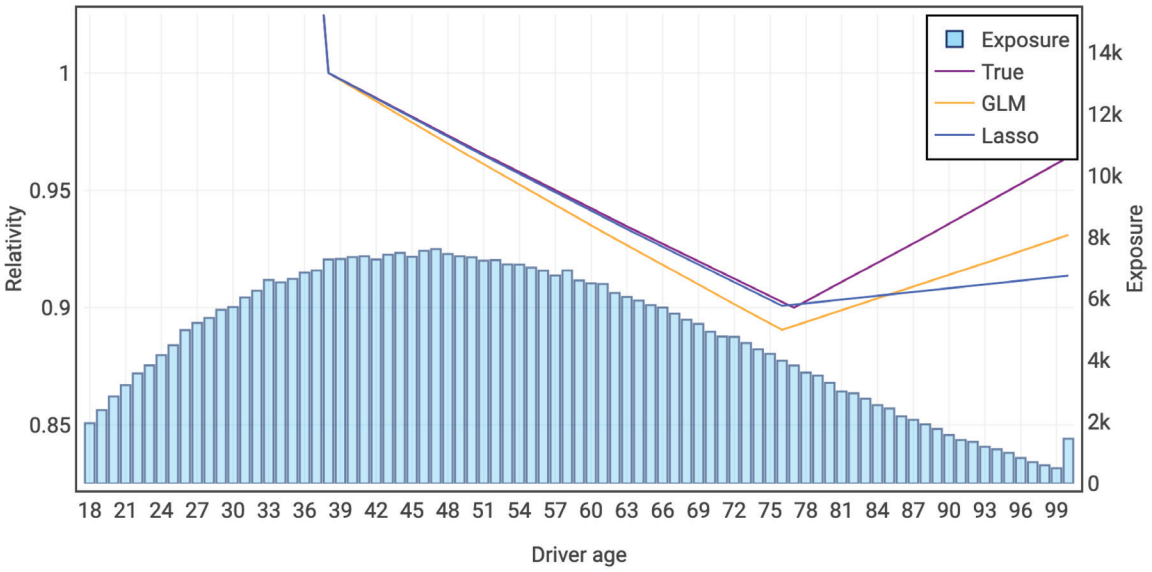
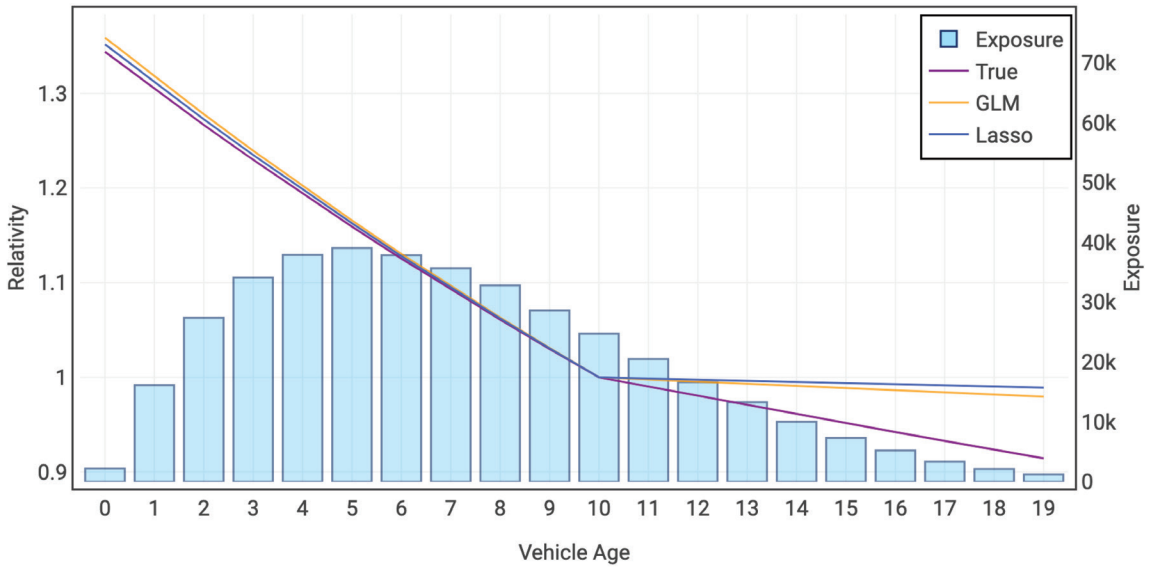
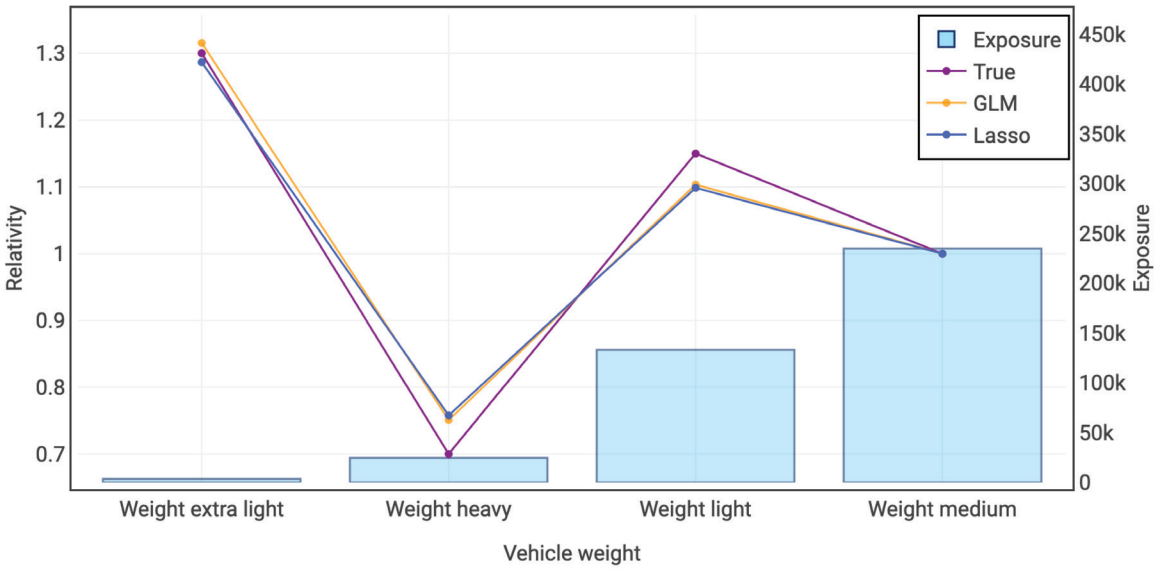


Figure 7.13. Both Models are Reacting Only Very Slightly to the Decrease in True Risk Relativity After Vehicle Age 10



Rather than rejecting all signal from a variable based on significance, penalized regression is able to reflect some of that experience by imposing a high shrinkage. Additionally, where *p*-values are closer to our 0.05 threshold, such as with extra-light vehicles in Figure 7.14, the model has automatically applied shrinkage to reflect that uncertainty. In GLMs, we would have to make this selection judgmentally after model fitting.

Figure 7.14. Extra-light vehicles are shrunk slightly toward 1.0. The GLM is overestimating the true relativity, whereas our lasso model is underestimating the true relativity.



7.4.2. Shrinkage Varies between Engineered Features

A corollary to low significance having high shrinkage is that shrinkage will vary for different transformations of continuous variables or for different levels of a categorical variable. The variable vehicle_age_10_99_hinge saw significantly more shrinkage than vehicle_age_0_9_hinge. This shrinkage is correlated with the exposure distribution of the continuous variable, as there are fewer exposures after vehicle age 10. As we noted earlier, lasso credibility is a likelihood-based credibility procedure and likelihood correlates heavily with exposure distribution. Therefore, the exposure distribution can be used by a modeler to understand where and why coefficients may see shrinkage.

7.4.3. Credibility and Feature Engineering

It follows that uncapped continuous variables and polynomial terms will not reflect credibility in a way similar to our segmented hinge feature engineering, which changes factors for only a portion of a given feature. A polynomial term such as vehicle_age_squared would have an effect across the entire distribution of vehicle ages and will use the "credibility" of the newer ages to extrapolate to the older ages. Intuitively, it is difficult to rationalize how the credibility of vehicle ages 0 to 9 can be used to support changes in ages 10 to 99. An ordinal treatment of variables as in derivative lasso or AGLM will apply shrinkage and credibility intuitively and appropriately. An ideal application of lasso credibility will use feature engineering that is easily understood through the lens of credibility.

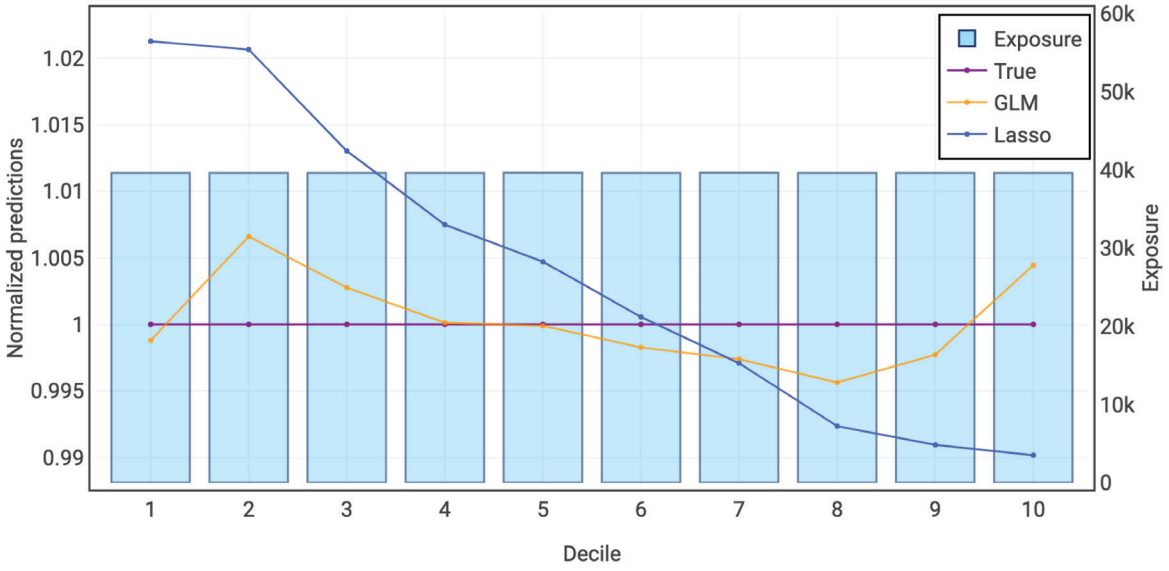
7.4.4. Penalized Regression Benefits

Let's compare the lasso penalization to two different versions of the GLM: one including all coefficients and another excluding those whose p-value is greater than 5%.

When we do not exclude insignificant variables, the GLM does slightly outperform the lasso model on a double lift chart (Figure 7.15). However, this GLM includes variables that have failed our significance test. Fortunately for the GLM, those unstable variables have experience similar to the true risk relativities. But without knowing the true relativities, how could we be sure that the insignificant variables are truly beneficial? Unfortunately, the GLM does not have the ability to assign partial credibility, and therefore we must refit the model without those variables. Even if the modeler takes a risk and gives the variables the benefit of the doubt, it is worth noting that the y-axis of Figure 7.15 is quite condensed and the lasso model is not being outperformed by a large margin.

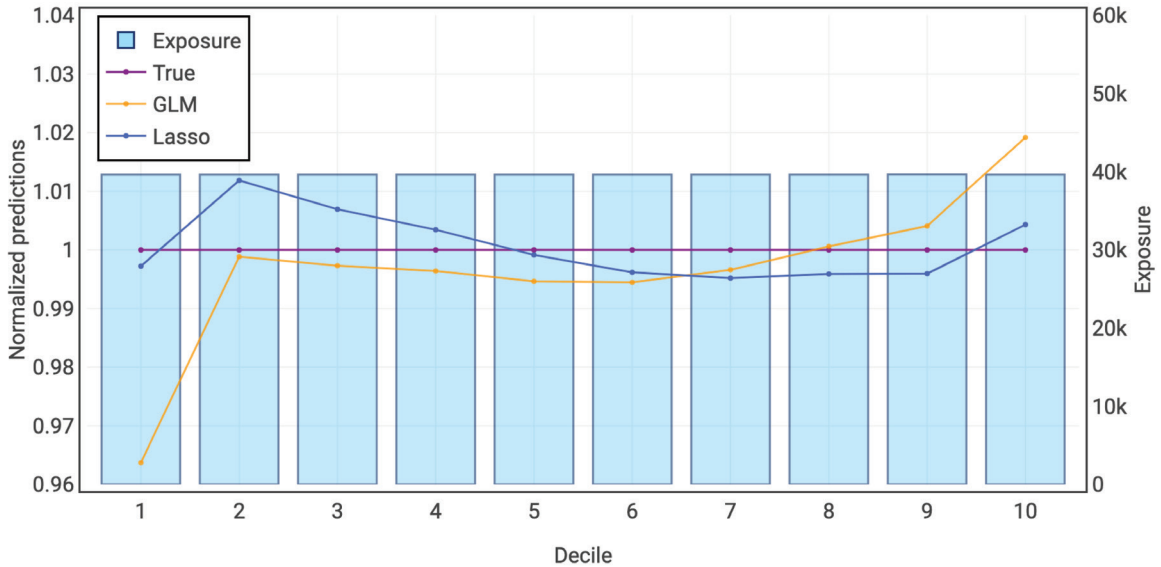
If we remove the insignificant variables from our GLM (Figure 7.16), the lasso penalized regression model performs significantly better in the tails of the double lift chart while performing within 1% of the true relativities for the middle deciles.

Figure 7.15. The GLM that Includes Insignificant Variables is Outperforming Our Penalized Regression Model by More Than 2% in the Lowest Decile



With this comparison we aim to illustrate that **some credibility is better than none**. Where we know that full credibility is not supportable, a partial credibility solution is better than the binary choice of removing a variable. Lasso penalization recognized this partial credibility and applied significant shrinkage to coefficients that were not significant in the GLM. That shrinkage resulted in a coefficient between our fully credible GLM coefficient and a 0.0 coefficient. A partial credibility treatment of coefficients is what allows the lasso model to outperform the GLM.

Figure 7.16. After Excluding Insignificant Coefficients, the GLM No Longer Outperforms the Lasso Penalized Regression Model



7.5. "Large State"—Lasso Credibility Versus GLM

We use the relativities from our full model as a complement for this subset.[9] By using our countrywide model as a complement of credibility, we are modeling credible differences from that model instead of creating an entirely new model from scratch. The inclusion of the extra information allows us to build a much more robust model than our GLM or lasso penalization model.

The large state data contains different true risk relativities than the base modeling data for the variables industry code (Figure 7.7), multipolicy discount (Figure 7.4), driver age (Figure 7.2), and vehicle age (Figure 7.3).

The shrinkage present in the lasso credibility model highlights the model's ability to react to various levels of credibility in the data. Some coefficients see high shrinkage where the difference between the complement and modeled relativity is set to zero, and others see low shrinkage and high reactivity. To illustrate this, relativity plots will now include each variable's complement of credibility. The penalized regression relativities are labeled as "lasso credibility" and are always between the complement relativity and the GLM relativity.

7.5.1. Coefficients of Zero Show Confidence in the Complement of Credibility

The lasso credibility coefficients for farming and food services are **completely removed** by the model's penalization. When a coefficient is at or very near zero in a GLM, the modeler can say that this characteristic is not predictive and should be removed from the model. On the other hand, lasso credibility concludes that there is "no credible difference" from our complement. Either the model has high confidence in the complement, low confidence in the data, or a combination of the two where the bias–variance trade-off is optimized through the application of bias.

Looking at Table 7.3, we can see that the difference between the fully credible GLM estimate and the complement of credibility is quite small. In high-exposure segments,

Table 7.3. A Comparison of Farming and Food Service Industry Code Relativities between Models

Model	Farming Relativity	Food Service Relativity
Complement of credibility	.680	1.205
Lasso credibility	.680	1.205
True relativity	.700	1.200
GLM	.668	1.171

[9] As always, an actuary should review the considerations in ASOP 25 when selecting a complement of credibility (see Section 5.4). The review should focus on this item in ASOP 25's Section 3.3, "Selection of Relevant Experience": "The actuary should consider the extent to which subject experience is included in relevant experience. If subject experience data is a material part of relevant experience, the use of that relevant experience may not be appropriate." We assume that our large state subset did not have an undue influence on the full model's predictions.

Table 7.4. A Comparison of the Construction Industry Code Relativities Across our Different Models

Model	Construction Relativity
Complement of credibility	1.460
Lasso credibility	1.436
True relativity	1.4
GLM	1.379

the likelihood benefit of small differences may produce credible deviations. However, in these lower-exposure segments, the likelihood improvement of moving toward the GLM estimate does not outweigh the penalty incurred by a nonzero coefficient.

7.5.2. Partially Credible Categories Avoid Overreactions

Other categorical coefficients do deviate from the complement of credibility. The lasso coefficient for construction is quite small: only −0.0163. The credibility-weighted relativity is moving toward the true relativity in our relativity plot but does not quite get there. On the other hand, moving from our complement relativity to the GLM relativity would overshoot the true relativity. The stability provided by lasso credibility can prevent unnecessarily large policyholder impacts caused by assigning too much credibility to noisy data while still moving closer to the true relativity as seen in Table 7.4.

7.5.3. Credible Categories React Quickly

The complement for multipolicy discount is not accurate in our large state example, and yet our lasso credibility estimate is still quite close to the GLM estimate as seen in Table 7.5. We can draw two conclusions from this:

1. Large categories can approach full credibility in the same model where small categories receive small or no credibility. (Figure 7.17 and Figure 7.18)
2. The complement of credibility is less material in categories with large exposure and more material in categories with small exposure. (Figure 7.17 and Figure 7.18)

Table 7.5. A Comparison of Factors for the Multipolicy Discount Variable

Model	Relativity
Complement of credibility	.772
Lasso credibility	.704
True relativity	.7
GLM	.694

Figure 7.17. Categories with High Levels of Exposure are Closer to the GLM Estimates Than Categories with Low Exposure

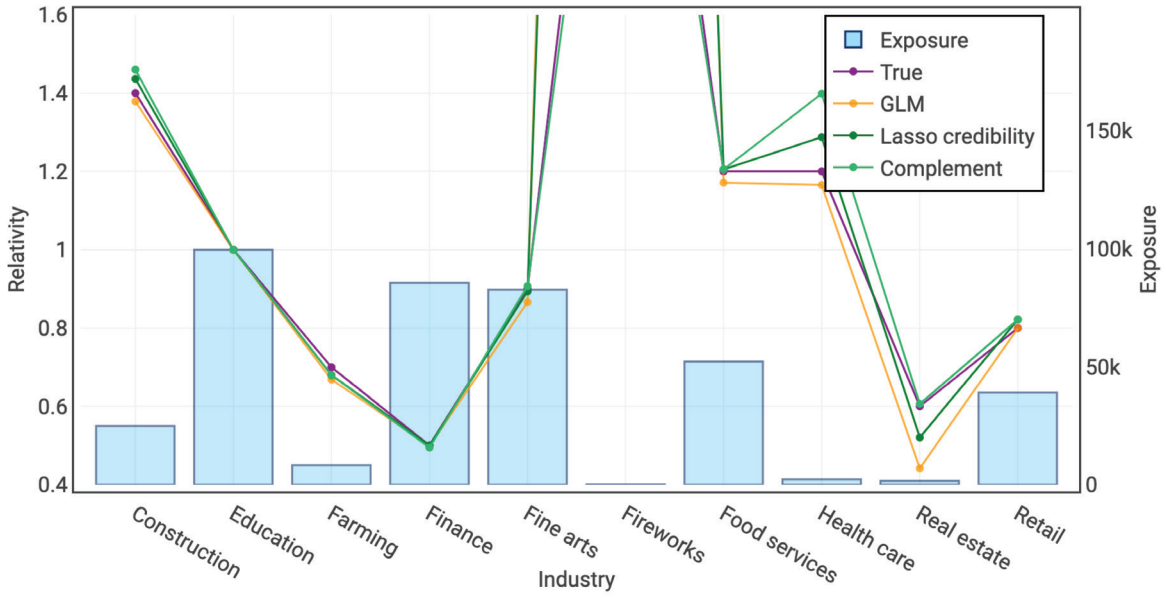
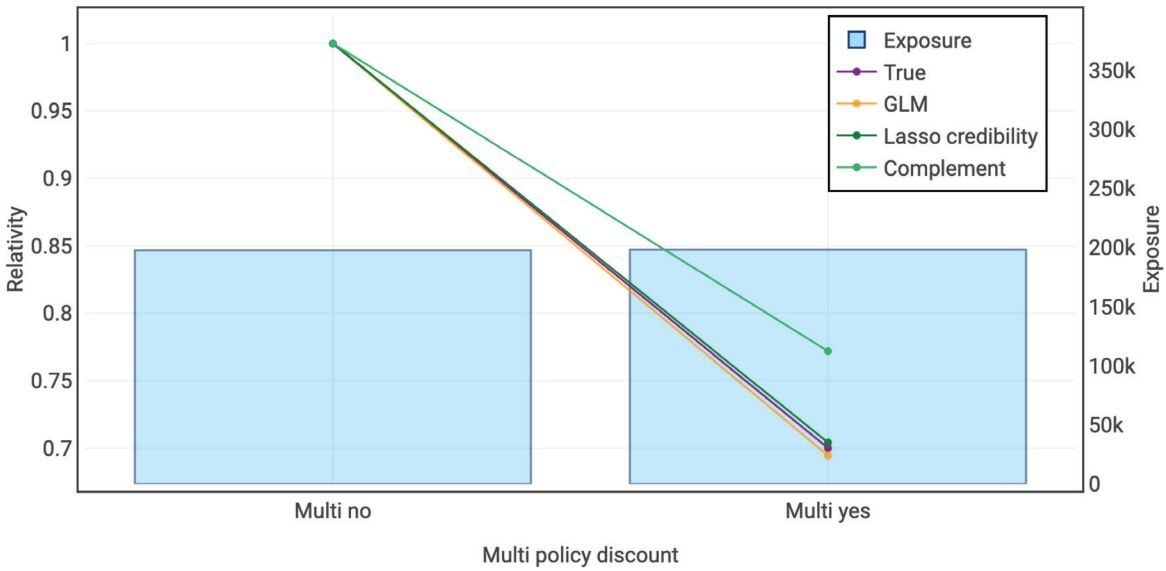


Figure 7.18. Categories with High Levels of Exposure are Closer to the GLM Estimates Than Categories with Low Exposure



The observation may seem trivial, but it is included to stress that not all components of the complement will be equally influential on model results.

7.5.4. Lasso Credibility Moves Toward Experienced Relativities

In Figure 7.19, shrinkage moves the driver_age_76_99_hinge coefficient toward indicated without overreacting, and the final driver age factor curve produced by our lasso credibility model is extremely close to the true relativities. This coefficient is insignificant in the GLM, and this variable would likely be completely removed from the model. By reflecting the credibility of our data, lasso credibility has achieved a great result.

On the other hand, Figure 7.20 shows that the 80% shrinkage applied to the vehicle_age_10_99_hinge coefficient assigns some credibility to the simulated experience, and this experience is quite far from the true relativity. This variable's insignificance in the GLM does not solve the problem, as removing the coefficient would result in an even more incorrect aggregate relativity.

Credibility procedures are effective in many cases, but it is still possible for particularly noisy data to draw indicated relativities away from the true relativities.

7.5.5. Performance Comparison: Lasso Credibility Versus Lasso Versus GLM

We can see in the double lift charts (Figures 7.21 and 7.22) that lasso credibility outperforms our GLM and penalized regression models in our large state example.

However, *this is not proof that lasso credibility will always outperform other model types.* When data is thin, lasso credibility's performance depends on a properly selected complement to pull the indicated relativities toward the true relativities. To illustrate

Figure 7.19. Lasso Credibility Prevents the Driver Age Relativity from Flattening Out Too Much for Older Driver Ages

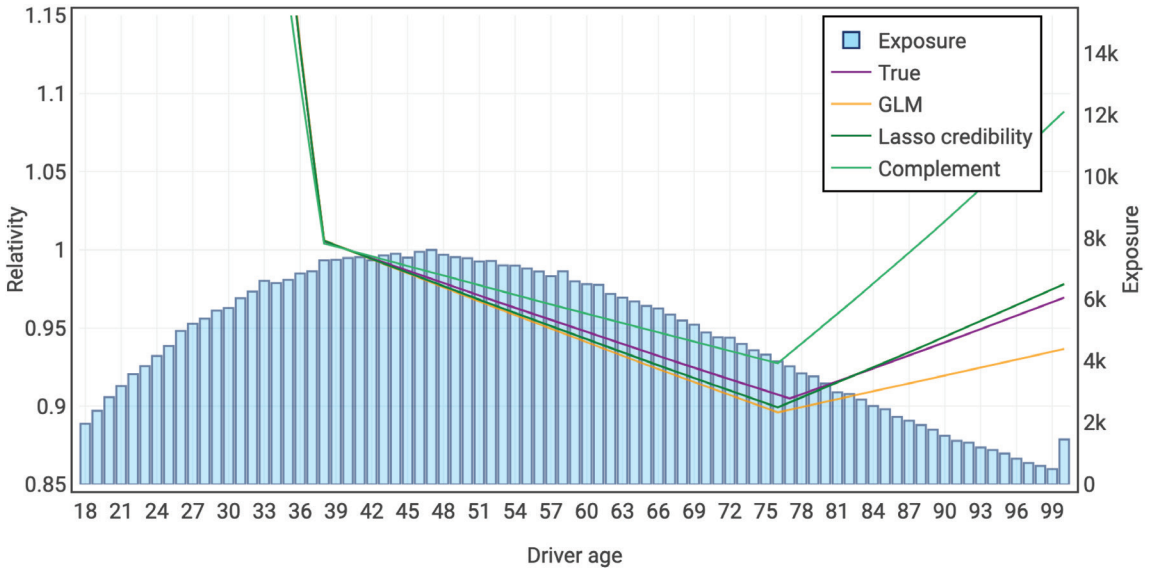
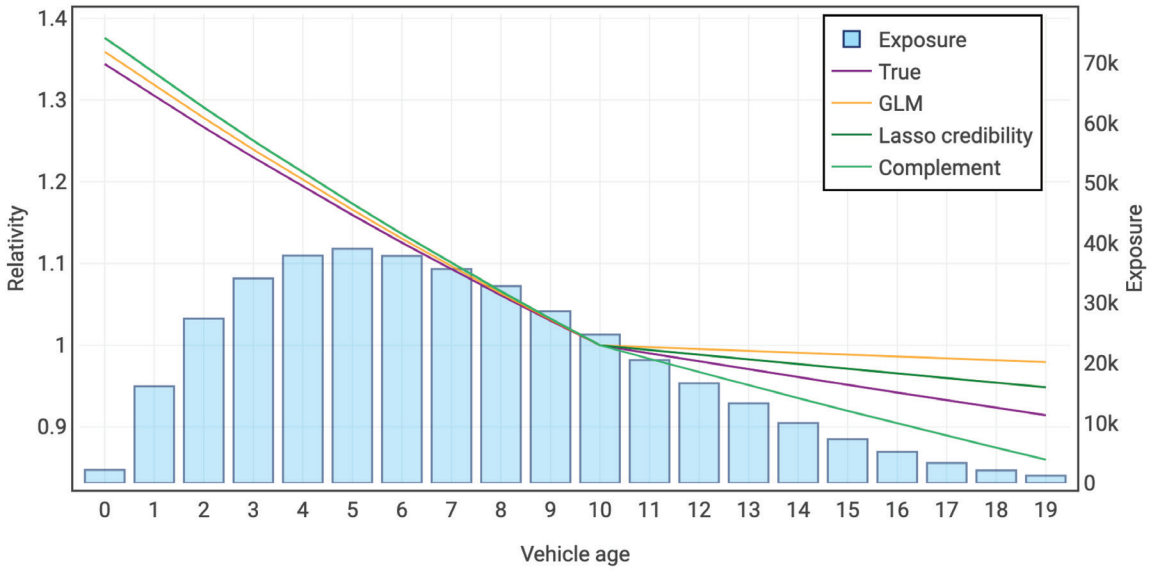


Figure 7.20. Lasso Credibility Prevents the Vehicle Age Relativity from Becoming Too Flat for Higher Vehicle Ages



this, examine the table below, which displays information on the indicated relativities of the health care and farming categories. The GLM produces an estimate closer to the true relativity for health care because the selected complements 1.399 and 1.0 are quite far from the true relativity of 1.2. The signal from the data was not strong enough to overcome the poor choice of complement. For the farming category relativity, the choice of complement is less material given the larger amount of exposure in this category.

Figure 7.21. The lasso credibility model clearly outperforms the GLM even when allowing lucky insignificant coefficients. After removing those coefficients, the lift is even more dramatic. This further supports our earlier opinion that some credibility is better than none.

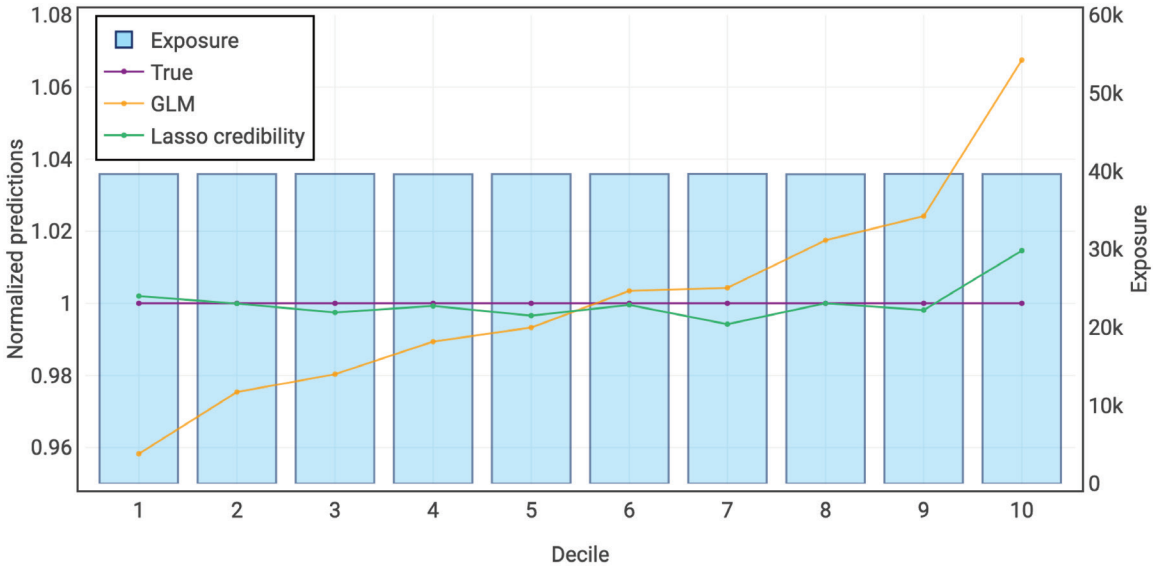
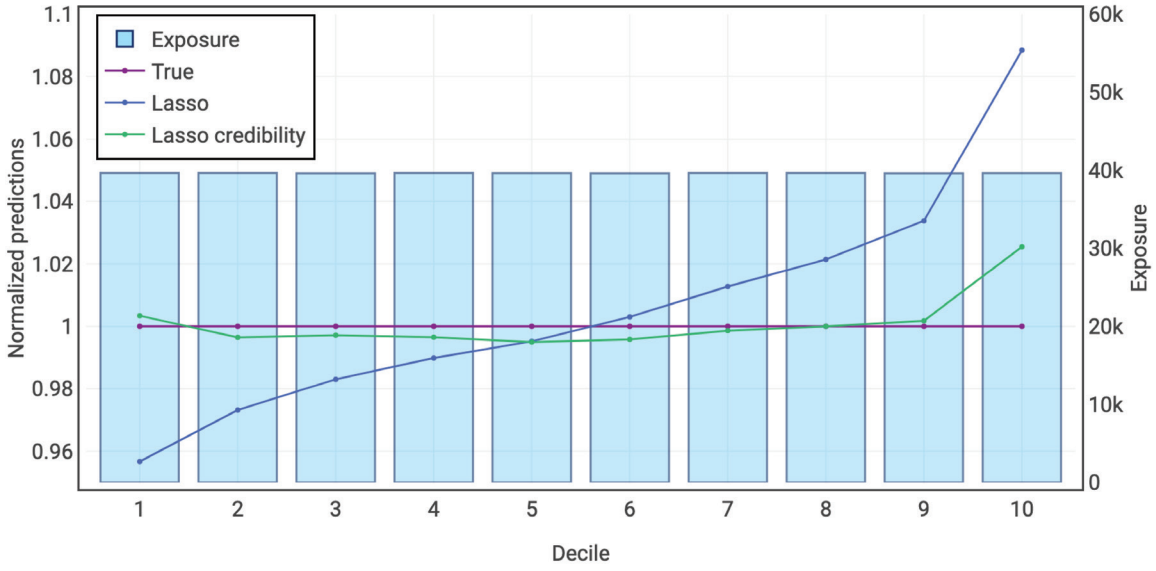


Figure 7.22. The Lasso Credibility Model Clearly Outperforms the Traditional Lasso Penalized Regression Model



Both a good complement of .68 and a poor complement of 1.0 produce similar modeled coefficients close to the true relativity as seen in Table 7.6.

In general, lasso credibility will be of more actuarial benefit than a GLM when the complement of credibility is pulling the indicated relativities in the correct direction toward their true relativities. Additionally, lasso credibility should outperform lasso penalization when the selected complement of credibility is more appropriate than the default assumption of 1.0. A poor complement can cause lasso credibility to perform worse than both lasso penalization and GLM.

7.5.6. Large State Conclusion

When we subset our data to this level, high-exposure segments still show minimal deviation in indicated relativities between the GLM and lasso credibility models. However, for smaller categories and some continuous variables, the GLM is already unable to provide significant coefficients. A modeler could explore additional feature engineering or the removal of insignificant variables in favor of an offset. Such additional adjustments and feature engineering would also benefit the lasso credibility model.

Table 7.6. A comparison of the Health Care and Farming Factors between Models

Category	True Relativity	GLM	Lasso Credibility Complement	Lasso Credibility with Original Complement	"Lasso Credibility" 1.0 Complement
Health care	1.2	1.165	1.399	1.288	1.101
Farming	.7	.668	.680	.680	.718

Lasso credibility is effective on data of this size, and we suggest that data of **all** sizes, even as large as our full modeling data set, can benefit from the application of lasso credibility. For every large data set, there is likely a not-fully-credible category. This not-fully-credible category can benefit from lasso credibility. From a theory perspective, this is motivated by the bias–variance trade-off (Chapter 4). For example, industry codes in the full modeling data set can be split into increasingly granular subsets until they begin to lose credibility and significance. Lasso credibility can help a modeler gain lift that would otherwise be unattainable for these segments.

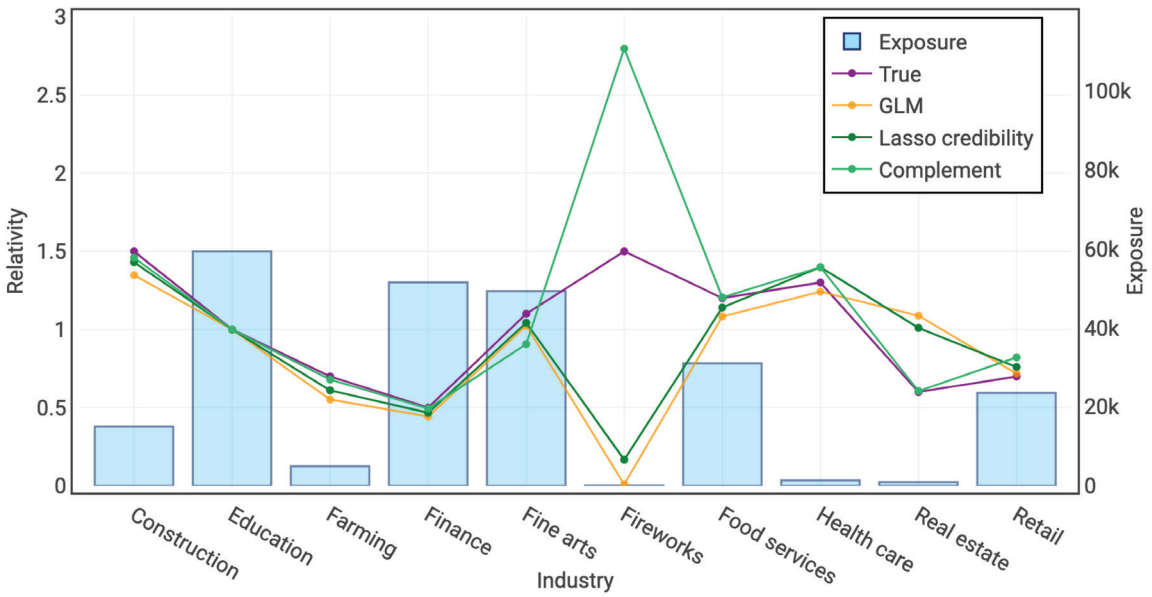
Here are some recommended exercises with the large state subset:

- Rerun the lasso credibility model with different complements of varying quality. Observe how the complement has more effect in low-exposure categories and segments than high-exposure segments.
- Resimulate the large state subset with a different seed. How consistent are the results from lasso credibility? How consistent are the results from a GLM?
- Resimulate the large state subset with different underlying relativities. How quickly does lasso credibility react to these differences? Is a GLM able to model these new relativities with significant coefficients?

7.6. "Medium State"—Lasso Credibility Versus GLM

A quick examination of the relativity plot for industry code (Figure 7.23) will show that both the generalized linear and lasso credibility models are producing indicated relativities quite far from the true relativities, and quite nonsensical ones for the fireworks industry code. Additionally, many of the coefficients are insignificant in our GLM. Is this data set too small or volatile to model? The answer is no: an adjustment of the penalty term in lasso credibility enables us to build an acceptable model on this data set.

Figure 7.23. Both Models are Producing Inaccurate Relativities for Real Estate, and the Indicated Relativity for Fireworks is Directionally Incorrect



7.6.1. Evaluating the Assigned Credibility

The relativity plot for industry code in Figure 7.23 has a couple of results that make us question the assigned credibility. First, in smaller data sets, a good complement would usually be given full credibility for some categories when their coefficients are penalized to zero. Only the health care coefficient is penalized to zero in this example. This by itself is not a bad thing, but an actuary should review the credibility of deviations for reasonableness. Second, some of the coefficients are directionally different than the selected complement. Does it make sense that our real estate relativities are slightly surcharged in this subset as opposed to receiving a sizable discount? Maybe—industry knowledge would be necessary to determine whether this is a reasonable movement. Third, and most importantly, we see a truly unintuitive result in the fireworks category. This category of only 69 exposures is deviating from a complement of credibility of 2.796 to an indicated relativity of .167. This is surely a misallocation of credibility.

As discussed earlier in the monograph, one can choose from a range of statistically reasonable lambdas. This result is a reason to select a smaller lambda and give more credibility to our complement.

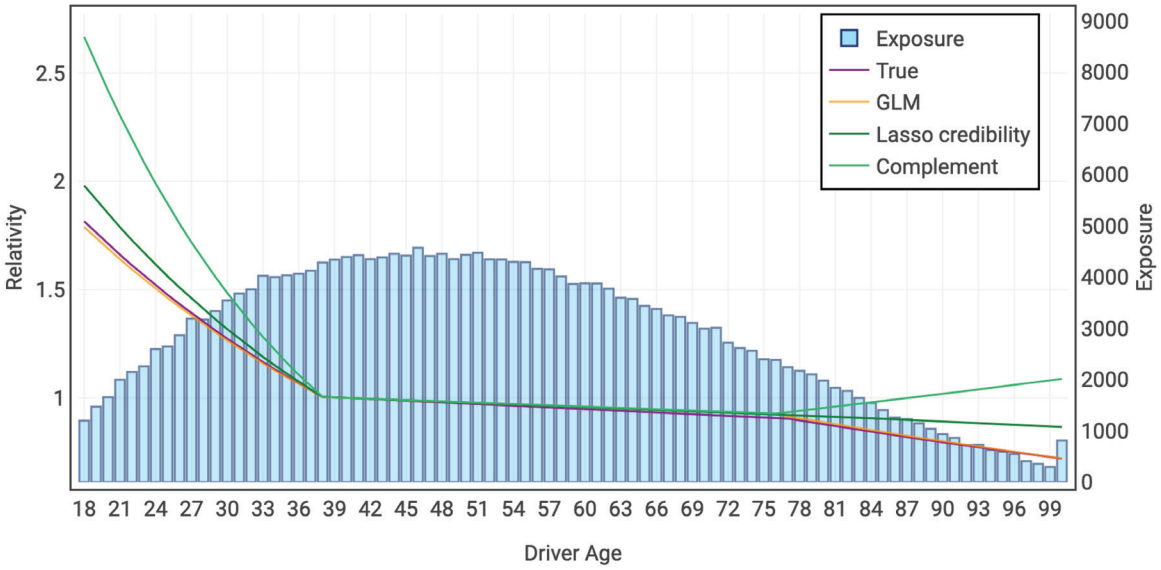
How can we be sure to recognize this misallocation of credibility in practice? A strength of lasso credibility is that the penalty is applied to all coefficients identically and therefore applies an equal credibility standard across all variables. If the credibility standard is incorrect for one variable, it is incorrect for all variables. In practice, this is easily addressed: a modeler should always start with the penalty that produces the most statistically sound model. If that produces unreasonable results for one or more variables, we recommend increasing that penalty term until the results are actuarially reasonable across all variables. An ordinal treatment of variables can make this evaluation quite clear.

Would *p*-values have fixed this? Not really. Our GLM has an indicated coefficient of 0.01 for fireworks with a *p*-value of 0.0645. This is very close to significance, and if it weren't so obvious that this result is wholly unintuitive, an actuary might use the result as support to judgmentally assign some discount to this industry code. Additionally, *p*-values would remove most of our continuous variable transformations and four of the other industry code coefficients. With enough work, a modeler might be able to create a reasonable model through robust variable transformations, but in its current state, the GLM would be unusable.

7.6.2. Some Credibility is Better Than None

We have talked about how some credibility is better than none for individual coefficients, and now we describe how some credibility is also better than none in the aggregate. The HDtweedie package allows us to view the range of tested lambdas, so we built models with increasingly high lambdas until the results appeared reasonable for all variables. This process assigns less credibility to our entire data set. Looking at the relativity plots and comparing to true relativities, we can see that this model is not perfect. Older drivers still have too much of a surcharge, real estate is erroneously deviating from the complement of credibility, and the change to our xTreme turn signal

Figure 7.24. Although still quite off, the lasso credibility model is a bit closer to the true relativity overall than the complement of credibility for the driver age variable.



variable is not fully recognized. Despite this, **the lasso credibility model is better than our alternatives**. (See Figures 7.24, 7.25, and 7.26).

The GLM fails to produce significant coefficients for half of the industry codes in our model, and similarly three of our five continuous variable transformations are insignificant. An actuary may be tempted to implement the countrywide model in this state instead of a state-specific model rather than perform significant work to turn this into a usable model.

Figure 7.25. Lasso Credibility Partially Reacts to Signal in x-Treme Turn Signal Variable

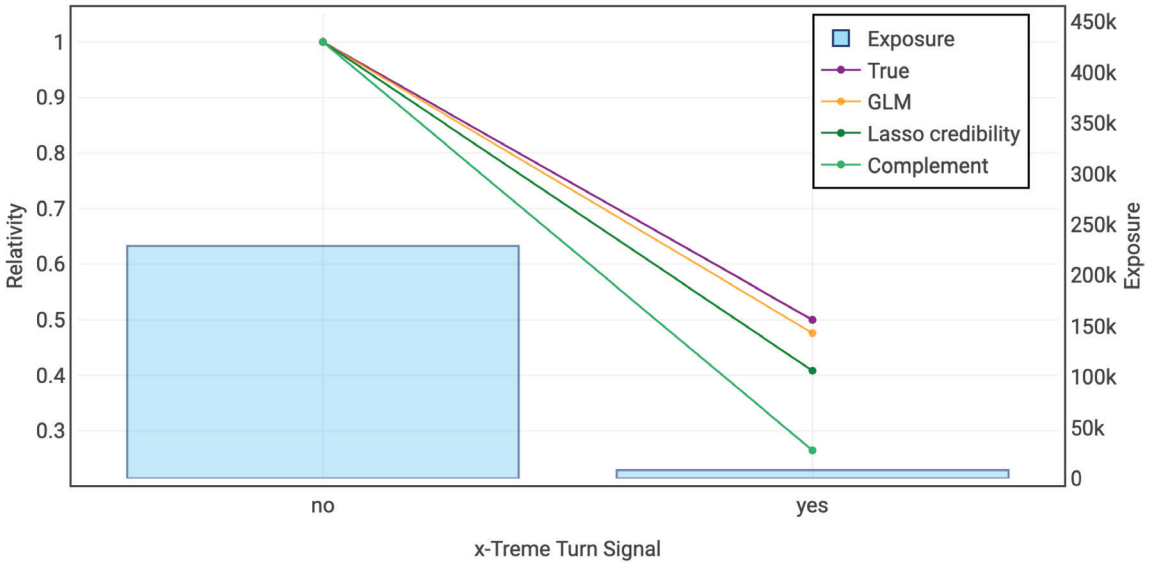
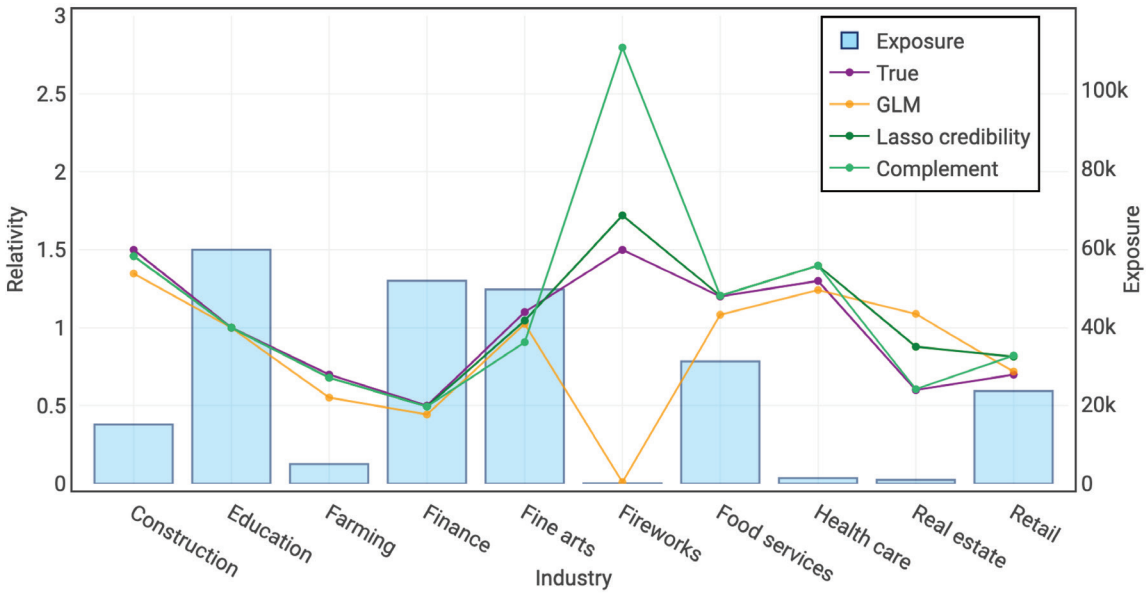
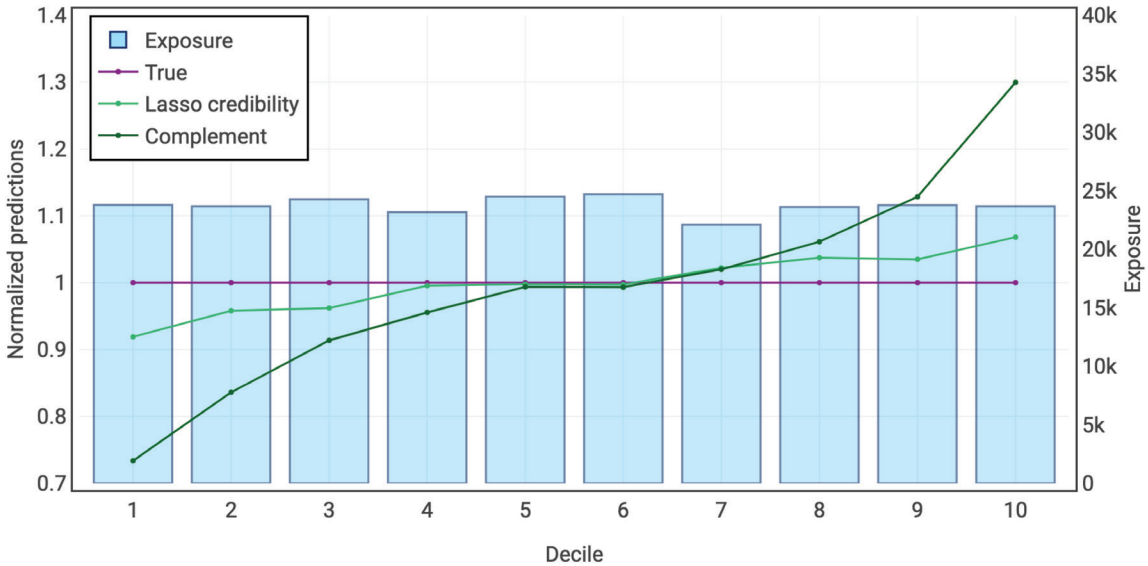


Figure 7.26. The Increased Penalty Term Has Placed Much More Weight on the Complement of Credibility



On the other hand, the lasso credibility model is improved simply by the tuning of the penalty parameter. Looking at the lift chart (Figure 7.27), we can see that the lasso credibility model is outperforming our countrywide model on this data set. By adjusting the penalty parameter, we are able to move closer to the state's true relativities without being overreactive. One of lasso credibility's key benefits is its ability to build models on data that would normally be considered out of scope for predictive modeling. Finding credible differences in a state's experience from a countrywide model is a key use case for lasso credibility.

Figure 7.27. The Lasso Credibility Model Outperforms the Countrywide Complement of Credibility



7.6.3. Medium State Conclusion

Our medium state data set is right on the edge of being able to produce a reasonable GLM, but lasso credibility models can still be built quickly. We can build better models on smaller data sets because **lasso credibility can incorporate prior assumptions and assign partial credibility to coefficients as opposed to assigning full credibility to all significant coefficients**.

7.7. "Small States"—Lasso Credibility Versus GLM

As a reminder, our first small-state subset has some different true relativities than the countrywide model, while our second small-state subset has true relativities that are identical to the countrywide model for all characteristics.

7.7.1. Lasso Credibility is Viable When GLM Fails

On this size of data, the GLM produces significant coefficients for only nine of the 19 variables included in the model, and a modeling project would likely be abandoned quickly. The lasso credibility model is not perfect—but it manages to build a mode that is better than our selected complement of credibility at identifying the true risk relativities overall (Figure 7.28). Rather than adopting the countrywide model directly, an actuary can use lasso credibility to quickly build state-specific models on small data where a GLM is not viable (Figure 7.29).

We can also judgmentally select a higher penalty term to move less toward indicated. Selecting a higher penalty term still produces better relativities than our complement of credibility and will result in smaller policyholder impacts (Figure 7.30). This model still performs better than the countrywide complement.

Figure 7.28. The Lasso Credibility Model Outperforms a GLM on This Small Subset of Data

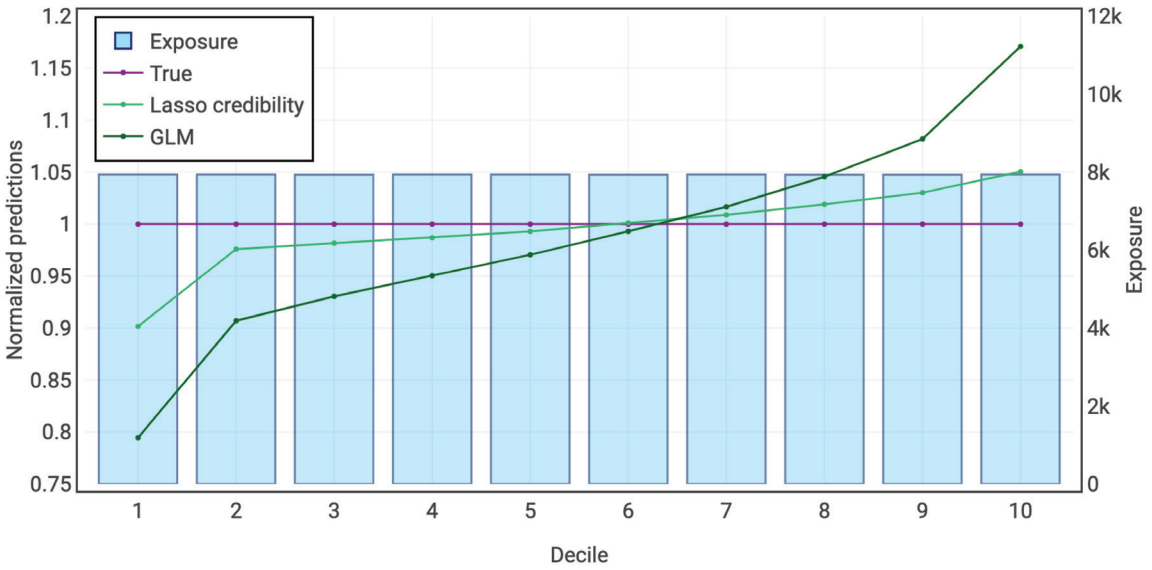


Figure 7.29. The Lasso Credibility Model Outperforms the Countrywide Complement of Credibility on This Small Subset of Data

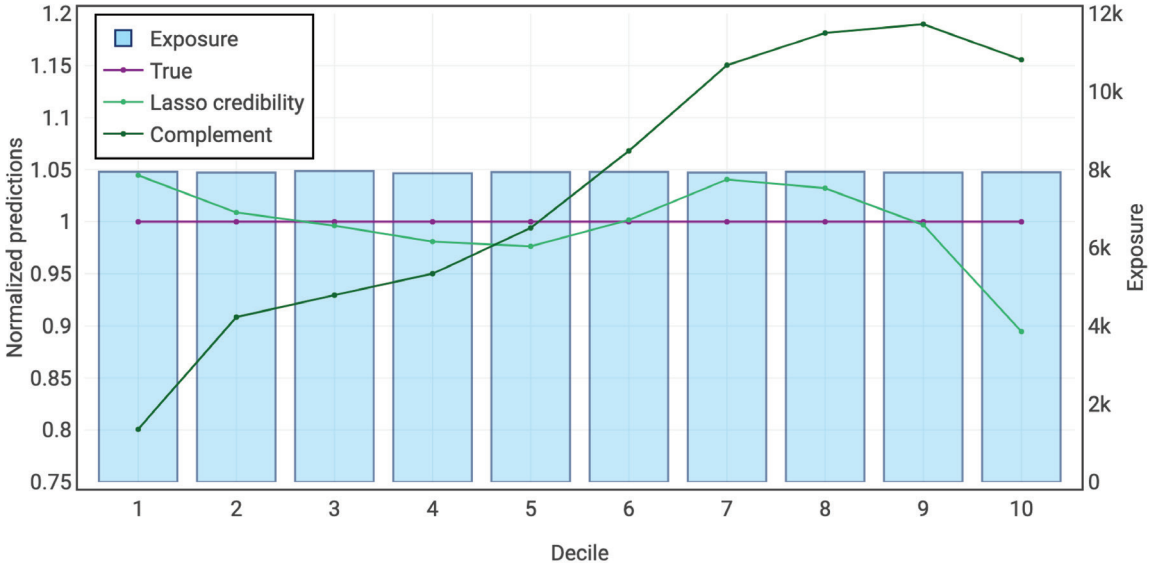
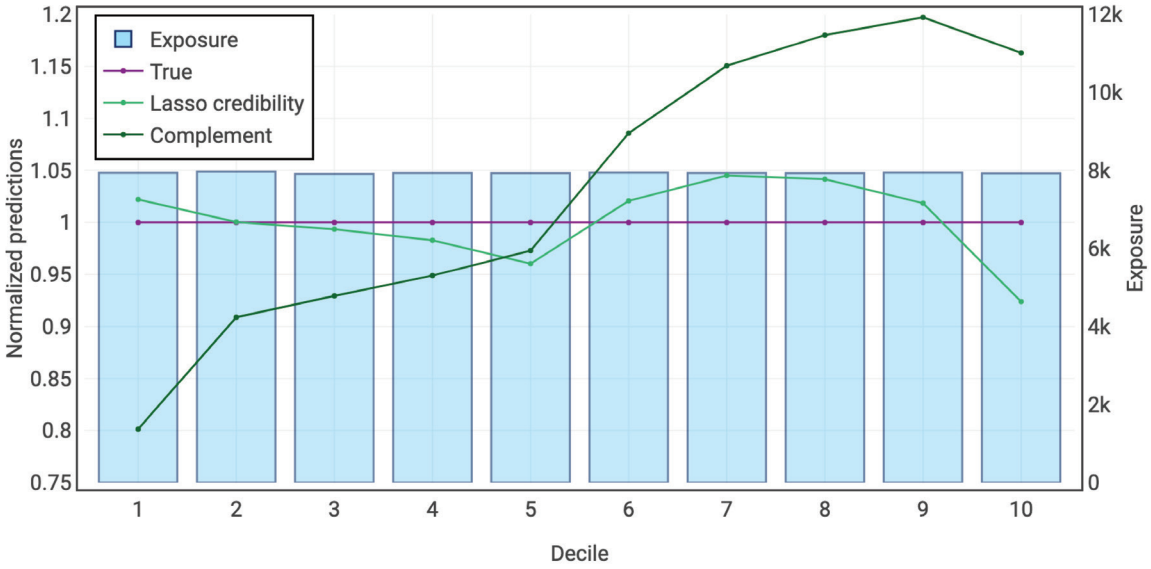


Figure 7.30. An Increased Penalty Creates a More Accurate Model by Preventing Overreactions in the Tails of This Lift Chart



7.7.2. A Good Complement Creates a Sparse Model

When we build a lasso credibility model on our second small data set (data with the same true risk relativities as our base data), the lambda indicated by cross-validation penalizes every single variable out of the model. Examining the relativity plots, we see that the complement of credibility is quite close to the true relativities for all variables. Without knowing the true relativities, an actuary would conclude that the data shows no credible differences from the complement.

This result is only possible by examining credibility instead of significance. **Rather than a process of evaluating *p*-values for significant differentiation from a null level, we are instead looking for credible deviations from a prior assumption**. When our complement is good, we can expect a model to output very few nonzero coefficients.

7.7.3. Small-State Conclusion

Lasso credibility models can be used to gain insights into data sets of increasingly small size. How can actuaries use this characteristic to apply lasso credibility in additional analysis?

For this, we turn to Actuarial Standard of Practice No. 56, *Modeling*. ASOP 56 provides significant guidance on the general usage of models, and we focus on one phrase that comes up repeatedly: the "intended purpose" of an analysis. In every aspect of modeling, an actuary should keep in mind the intended purpose and end usage of an analysis. Rather than for the intended purpose of "fully justifying new rates," an actuary can use lasso credibility for "identifying credible differences for further investigation." When using lasso credibility for analysis, its scope greatly increases beyond that of traditional GLM building.

Before an insurer has enough data to build a model for implementation, the actuary can use lasso credibility to identify the most credible deviations from that insurer's current rating plan. This multivariate analysis can save an insurer time when looking for profitability issues by quickly narrowing the scope of further analysis. An ordinal treatment of continuous variables is required in this approach, as the time required to identify the correct continuous variable transformations would likely make the analysis prohibitively long.

An actuary could also perform model monitoring by fitting a lasso credibility model on only the latest year of data. Whereas such a model would almost certainly not be appropriate for implementation, it would identify experience that is credibly out of pattern with the implemented relativities for further investigation.

7.8. Case Study Conclusion

We have shown how lasso credibility can outperform GLM and be used to build credible models where traditional GLM or penalized regression approaches would fail. This countrywide-to-state refit is expected to be one of the main use cases of lasso credibility, but we anticipate that actuaries will use the guidance of ASOP 56 to find many additional use cases for lasso credibility.

The key to lasso credibility is the switch from **significance** to **credibility.** This switch is only achievable through penalized regression and a proper modeling setup. In our example, we used known variable transformations as a shortcut for proper model setup. In practice, an ordinal treatment of variables is necessary to create a stepwise application of credibility. Without prior knowledge of feature behavior, the penalization will remove the less credible ordinal steps and leave only the most credible deviations. It is this ordinal treatment of variables that allows us to remove this simplifying assumption and apply lasso credibility on any data set.

We encourage the reader to rerun the provided code with alterations to explore various scenarios to solidify their understanding and test the limits of the application of lasso credibility. Examples include these:

- Start with a better complement of credibility.
- Start with a worse complement of credibility.
- Experiment across multiple starting seeds.
- Experiment with alternate variable transformations.
- Apply an ordinal treatment to continuous variables and/or vehicle weight.
- Perform "model update" scenarios by resimulating the modeling data.
- Start with more volatile or less volatile distribution assumptions.
- Incorporate "random noise" in the true pure premium distribution.
- Incorporate a variable in the true pure premium distribution that is not in the model.
- Incorporate a variable in the model that is not in the true pure premium distribution.

8. Conclusion—Overall

As we have seen, penalized regression can be applied as a credibility procedure and has strong mathematical links with widely accepted credibility procedures. Appendix A identifies a special case of equivalence between penalized regression and Bühlmann credibility, and it is important to note that equivalence is not necessary, as Appendix B defines penalized regression as a credibility procedure without relying on the link to other widely used procedures. Similarly, best practices applied to building lasso credibility models should be based on best practices of penalized regression in general—not taken from classical or Bühlmann credibility procedures directly. This change in perspective from lasso penalization to lasso credibility does not change the mathematical foundation of penalized regression, but rather it constitutes a new actuarial application of the existing technique.

Model reviewers should not directly take model validation requirements from either GLM or Bühlmann credibility techniques. Lasso credibility should be reviewed as a stand-alone credibility technique with its own model review standards. Those standards are taken from a combination of penalized regression and credibility procedure expertise. Straightforward application of existing model validation procedures to lasso credibility will be suboptimal.

We recommend an ordinal treatment of continuous variables in lasso credibility because it allows the model to identify credible differences from the complement during model fitting. The technique was popularized in Iwasawa and Wang (2022) and expanded upon in Casotto and Holmes (2023). We encourage the reader to investigate these techniques; they provide support for moving to a modeling approach where feature engineering is objective and fully credibility based.

Lasso credibility deserves a place in every actuary's analytical tool kit. By combining lasso penalization with a complement of credibility in the offset, one can use lasso credibility to analyze data sets of increasingly small sizes. Similarly, increasingly small subsets of larger data sets can be analyzed for insights and additional segmentation opportunities. Such complements of credibility can come from many sources and can be applied to many different types of analysis. Penalized regression is not just for big data!

We hope that this monograph has provided the necessary background for the reader to begin using penalized regression and lasso credibility. Happy modeling, and all the best in your actuarial and personal endeavors.

Appendix A. Bayesian Interpretation of Credibility

In this appendix, we explain why a GLM offers full credibility and how penalized regression mathematically aligns with Bühlmann credibility. After illustrating their differences, we delve into penalized regression's relationships with Bayesian statistics.

A.1. Why GLMs Give 100% Credibility to the Data

The statement "GLMs grant 100% credibility to the data" is a recurring theme in this monograph. To support it, we show that for all canonical link GLMs, predicted averages invariably match observed segment averages. The proof's essence lies in the computation of the GLM's parameter β, which maximizes the data's negative log-likelihood.

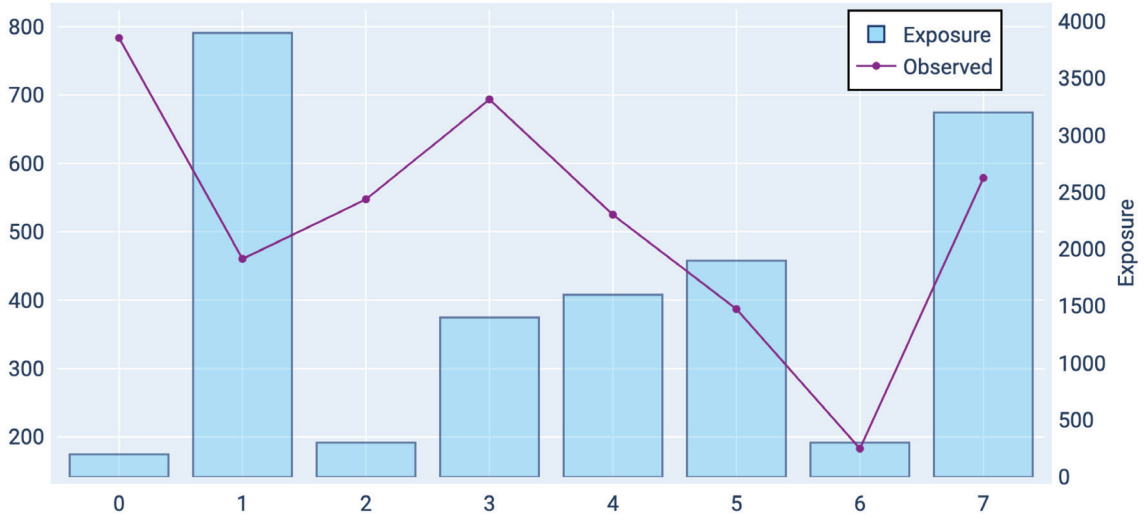
Setting the derivative (or gradient) to zero during likelihood maximization results in estimates where predicted averages coincide with observed segment averages, **independent of the underlying exposure**.

To introduce some mathematical notation, we start by considering an actuarial use case that consists of building estimates of loss costs for companies in a workers' compensation insurance rating plan. The modeler has access to a data set of historical loss experience, where each row represents the total loss observation for a specific company in a fixed 1-year period. Additionally, the class code describing the industry type is available for each company. There are several class codes and for most of them the number of observations is limited. An example of such a database is shown in Figure A.1.

The database contains n observations. The information on the companies is encoded in the matrix X, which provides the binary representation of the p class codes in the database. The coordinate x_{ij} of the matrix X will be 1 if company i belongs to class j, and 0 otherwise. X_i will denote the row vector of matrix X of size p. In general, index i will be used to represent a line in the matrix/database, and i takes values from 1 to n. The index j will be used to represent columns of the matrix, and j takes values from 1 to p.

Y represents the vector of the observed losses, meaning that Y_i will represent the observed loss for company i. The grand average is given by $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. y represents the differences of the observed losses from the grand average $\bar{Y}$, that is, $y_i = Y - \bar{Y}$ (so y is centered on zero).

Figure A.1. The purple line represents the behavior of the observed average loss deviation $\bar{y}_j$ by class code j on synthetic data. The blue bars represent the total number of observations n_j by class code j.



n_j denotes the number of observations belonging to class j. The set J represents the set of rows i belonging to class j, that is, i such that $x_{ij} = 1$. This implies that the cardinality of J is n_j. The constant $\bar{y}_j$ represents the observed average loss deviation by class code j, that is $\frac{1}{n_j}\sum_{i\in J} y_i$.

We recall from Section 1.1 that to fit a GLM we define the following items:

- The distribution is a normal distribution with constant variance σ^2.
- The link is the identity (canonical link).
- The target is the vector y representing the differences of the observed losses Y from the grand average $\bar{Y}$.[10]

The GLM coefficients β are estimated by maximizing the log-likelihood, or, equivalently, by minimizing the *negative* log-likelihood (NLL):

$$\begin{aligned}\hat{\beta} &= \underset{\beta}{\operatorname{argmax}}\, \text{Loglikelihood}\left(y, X, \beta\right) \\ &= \underset{\beta}{\operatorname{argmin}}\, NLL\left(y, X, \beta\right),\end{aligned}$$

which in the case of a Gaussian distribution with an identity link becomes

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - X_i\beta\right)^2.$$

[10] A problem is considered as numerically tractable if an optimization algorithm can provide its optimal solution in a reasonable time. This usually means that the problem is convex.

Since σ is constant, it doesn't affect the optimization problem, and we can ignore it. We can proceed to compute the gradient only of the summation, by differentiating with respect to each coefficient. This solution can then be found by setting the gradient to zero. The gradient can be simplified by application of the chain rule of the derivative:

$$\frac{\partial}{\partial \beta_j}\left[\sum_{i=1}^{n}\left(y_i - X_i\beta\right)^2\right] = \sum_{i=1}^{n}\left(X_i\beta - y_i\right)x_{ij}$$

We can further simplify the formula since the summand is not null only for $i \in J$ since $x_{ij} = 0$. Otherwise

$$\sum_{i=1}^{n}\left(X_i\beta - y_i\right)x_{ij} = \sum_{i\in J}^{n}\left(X_i\beta - y_i\right)$$

Since for all $i \in J$, $X_i\beta = \beta_j$; the $n_j\bar{y}_j$ term is by the definition of the average $\bar{y}_j = \sum_{i\in J} y_i / n_j$ the first addend $n_j\beta_j$ appears and we obtain

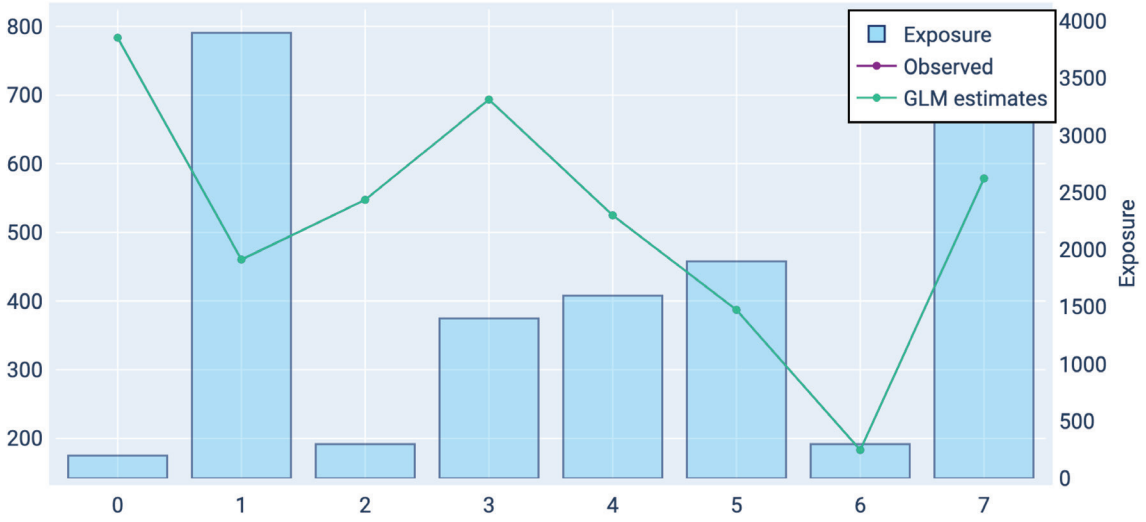
$$\sum_{i\in J}^{n}\left(X_i\beta - y_i\right) = n_j\beta_j - \sum_{i\in J} y_i = n_j\beta_j - n_j\bar{y}_j$$

Setting the gradient to zero implies that

$$0 = n_j\beta_j - n_j\bar{y}_j \quad \leftrightarrow \quad \beta_j = \bar{y}_j\,.$$

This proves that the coefficients β_j maximizing the log-likelihood are exactly matching the average $\bar{y}_j$ of each class.

Figure A.2. The green line compares the GLM estimates $\hat{\beta}_j$ with the data as represented in Figure A.1. The observed (purple line) coincides with the GLM estimates (green line).



This simple proof highlights two important features of generalized linear modeling:

- When modeling a single class in a binary encoding, a GLM will output the average of the observations for such a class. This is true as well for all distributions when the canonical link is used, and it can be proven equivalently by changing $X_i\beta$ with $\mu_i = g^{-1}(X_i\beta)$ as in Ohlsson and Johansson (2010).
- The formula applies as well on multivariate settings. Whenever a discrete variable is added in a binary format, the coefficients will be estimated so that the average of the predictions $X_i\beta$ on every level j will match the average of the observations, **regardless of the underlying exposure**.

In this sense, **any fitting procedure** that computes estimates by maximizing the likelihood of the data (or minimizing the deviance) **alone**, *effectively* ***assumes*** *that the underlying data sets are 100% credible, no matter their size.*

A.2. Credibility: A Bayesian Interpretation

Maximizing likelihood doesn't inherently blend credibility into both the model's factors and estimates. To address that, we must reconsider how a model interprets and estimates data.

Using the maximum likelihood formula to compute GLMs is characteristic of the "frequentist" statistical approach. We propose complementing this with a "Bayesian" perspective. "The Bayesian and classical versions have a lot in common, but they have a philosophical difference in that in classical statistics parameters are constants, but for Bayesians they have distributions" (Venter n.d., x).

Before modeling with GLM, two assumptions are essential. First, a probabilistic description of the observed response must be established, defining the data-generating distribution. Second, a link function should be selected to depict the relationship between the linear predictor (comprising parameters and covariates) and the target.

Once those are set, the parameters are estimated by maximizing the likelihood over the data. Such a maximization will try to replicate the observed data as closely as possible, giving 100% credibility to the data.

From a statistical perspective, this happens because the frequentist approach to modeling assumes that there exists a fixed set of true coefficients: the observed values of the target are assumed to have been generated assuming these fixed coefficients and the hypothesis assumed above. Our best guess is thus the set of coefficients maximizing the probability of observing the actual values of the target, motivating a maximum likelihood approach.

In a Bayesian perspective, on top of the standard hypothesis done over the observations, a distributional assumption is made on the coefficients of the model themselves (called **the prior distribution**). This assumption describes our a priori knowledge of and uncertainty over the values of the coefficients (hence the name), which are now random variables, and allows for the inclusion of some additional structure on the coefficients' estimates.

Bühlmann credibility can be considered from both a Bayesian and a frequentist point of view (Tse 2009).

The Bayesian equivalence has been proved by Jewell (1974): for all data-generating distributions used in GLMs, one can find an appropriate prior distribution such that the resulting Bayesian estimation coincides with the estimator obtained via Bühlmann credibility.

Jewell's result shows that for a given statistical hypothesis on the target, there exists a certain prior distribution for the coefficient β (called **conjugate**) that will return the Bühlmann estimator.

Jewell's result allows us to put Bühlmann credibility into the much more generic framework of Bayesian statistics. Under the Bayesian perspective, the connection with penalized regression will be evident.

To see how that connection works in practice, it is helpful to apply Jewell's result to the workers' compensation use case. The goal is to show that a Bayesian model with proper priors returns the exact same estimates as the Bühlmann credibility formula.

As a reminder, the Bühlmann credibility estimator is given by

$$\hat{\beta}_j = \frac{n_j}{n_j + k} \bar{y}_j,$$

where $k = \frac{\sigma^2}{\tau^2}$ is the ratio between the within-class variance and the between-class variance (see Table 2.1).

The initial assumption in the workers' compensation use case was that the loss deviations y are normally distributed around the estimations, that is, $y_i \sim N(X_i\beta, \sigma^2)$. The conjugate of the normal distribution with known variance is the normal distribution itself. For this reason we'll suppose that a priori β itself follows a normal distribution with constant variance τ^2 and mean zero: $\beta \sim N(0, \tau^2)$.

The choice of the normal **a priori** on $\beta \sim N(0, \tau^2)$ can be motivated as well by pragmatic considerations: deviations of class code losses from the grand average should be centered around zero. Furthermore, large deviations from the grand average should be

Table A.1. Pairings of statistical assumptions with their corresponding conjugate distribution for commonly used GLMs. Gamma (α) refers to a gamma distribution with known parameter α.

Statistical Assumption	Conjugate Distribution
y ~ Gaussian	β ~ Gaussian
y ~ Poisson	β ~ Gamma
y ~ Gamma(α)	β ~ Gamma
y ~ Binomial	β ~ Beta

considered a priori as less likely than minor deviations. In this sense, **the priors allow us to formalize commonsense intuitions in a robust mathematical framework**.

To define a Bayesian estimator for this model, it is necessary to derive the so-called *posterior* distribution, which updates through the likelihood (through the data) our *a priori* belief about the unknown parameter (the β should not deviate too much from zero, the mean). This update (by Bayes' theorem) takes the form of

$$p_{posterior}\left(\beta \middle| y, X\right) = \frac{1}{K} \times p\left(y \middle| \hat{y}\left(X, \beta\right)\right) \times p_{prior}\left(\beta\right),$$

where K is a constant that doesn't depend on β.

Under these specific assumptions of normality (see Section A.5 for more information), the solution can be computed via the maximum a posteriori formula:

$$\begin{aligned}\hat{\beta} &= \underset{\beta}{\operatorname{argmax}}\, p\left(y \middle| \hat{y}\left(X\right)\right) \times p_{prior}\left(\beta\right) \\ &= \underset{\beta}{\operatorname{argmin}}\, \mathrm{NLL}\left(y, X, \beta\right) - \log\left(p_{prior}\left(\beta\right)\right).\end{aligned}$$

The first summand is, for the workers' compensation use case, given by the same formula optimized by a GLM, that is,

$$NLL\left(y, X, \beta\right) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - X_i\beta\right)^2.$$

The second summand, since $\beta \sim N(0, \tau^2)$, is equal to

$$-\log\left(p_{prior}\left(\beta\right)\right) = \frac{1}{2\tau^2}\sum_{j=1}^{p}\beta_j^2 + C,$$

where C is a constant that does not depend on β, and can be removed from the optimization problem, which becomes

$$\begin{aligned}\hat{\beta} &= \underset{\beta}{\operatorname{argmin}}\, \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - X_i\beta\right)^2 + \frac{1}{2\tau^2}\sum_{j=1}^{p}\beta_j^2 \\ &= \underset{\beta}{\operatorname{argmin}}\, \frac{1}{2}\sum_{i=1}^{n}\left(y_i - X_i\beta\right)^2 + \frac{\sigma^2}{2\tau^2}\sum_{j=1}^{p}\beta_j^2. \qquad \text{(A.1)}\end{aligned}$$

Formula in Equation A.1 shows as well that the Bayesian estimates' optimization formula is equal to that of ridge regression with $\lambda = \frac{\sigma^2}{\tau^2}$. Furthermore, by the Jewell theorem, we already know that the optimal solution is equal to the Bühlmann estimates, with $k = \frac{\sigma^2}{\tau^2}$.

A.3. Penalized Regression: A Bayesian Interpretation

The equivalence of ridge estimates to this specific use case isn't accidental. As Miller (2015) demonstrated, every penalized regression can be framed as a Bayesian model, rooted in the maximum a posteriori (MAP) formula:

$$\begin{aligned}\hat{\beta} &= \underset{\beta}{\operatorname{argmax}}\, p\left(y \middle| \hat{y}(X)\right) \times p_{prior}(\beta) \\ &= \underset{\beta}{\operatorname{argmin}}\, \mathrm{NLL}(y, X, \beta) - \log\left(p_{prior}(\beta)\right). \end{aligned} \tag{A.2}$$

There is a one-to-one correspondence between prior distribution (prior assumption on the coefficients) p_{prior} and penalties:

$$\mathrm{Penalty}(\beta) = -\log\left(p_{prior}(\beta)\right). \tag{A.3}$$

We just showed how the ridge penalty corresponds to a normal prior on the coefficients with $\lambda = \frac{\sigma^2}{\tau^2}$. Furthermore, under the hypothesis of the workers' compensation use case, **ridge regression and Bühlmann credibility are equivalent** with $k = \lambda$.

In the case of the lasso, the associated prior is the Laplace distribution.

For context, the Laplace random variable with mean $\mu = 0$ and scale γ has the distribution

$$f_{Laplace(0,\gamma)}(x) = \frac{1}{2\gamma}\exp\left(-\frac{|x|}{\gamma}\right).$$

Figure A.3 compares the Gaussian and Laplace distribution.

Replacing Equation A.3 with the prior assumption that $\beta \sim Laplace(0, \gamma)$, then

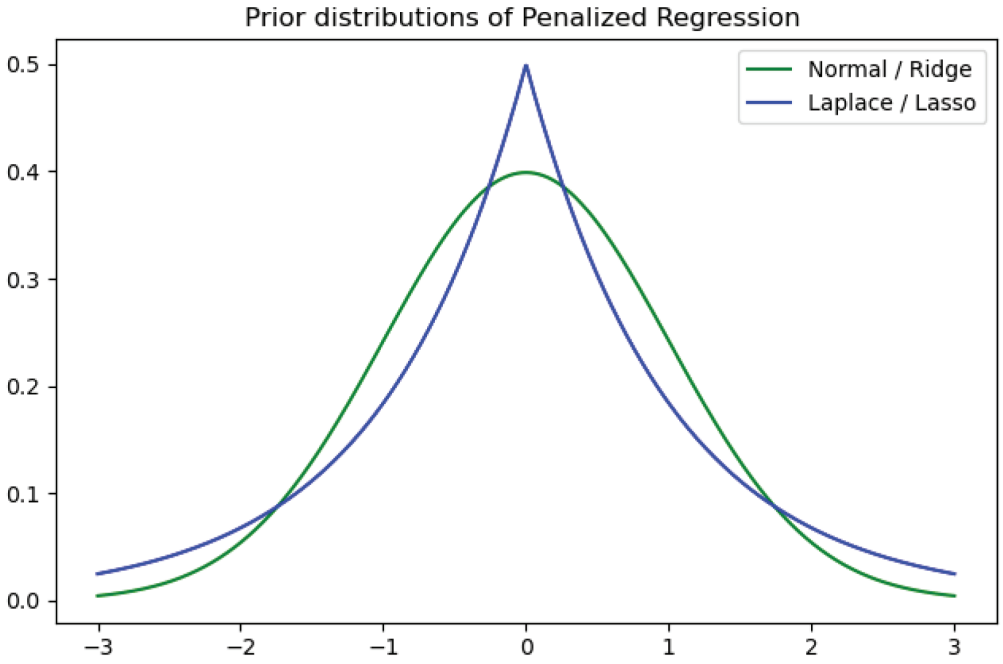
$$-log\left(p_{prior}(\beta)\right) = \frac{1}{\gamma}|\beta_j| + C,$$

where C is a constant independent on β, which is ignored when computing a MAP estimator. The formula is equal to the lasso penalty with $\lambda = 1/\gamma$.

The formulas highlight a strong connection between Bayesian and penalized regression modeling:

1. Bayesian modeling offers a general framework with which to model the uncertainty of the estimation β when the number of observations is limited. This uncertainty translates into a choice of a **prior** hypothesis on the coefficients β, i.e, a certain distribution that the coefficients β are assumed to follow.

Figure A.3. Comparison of Densities of Both the Normal/Ridge Prior and the Laplace/Lasso Prior



2. The estimator of the Bayesian model can be found by maximizing the posteriori log-likelihood $p_{posterior}(\beta|\, y, X)$. When taking the logarithm, its structure decomposes naturally into two terms: the log-likelihood (equal to the GLM formula) and the log-probability of the prior (which acts as a penalty).
3. Bühlmann credibility is an instance of a specific Bayesian model where the error distribution and the prior are conjugates (Table A.1). The choice of conjugate distributions allows us to compute the estimates in an explicit form. Explicit formulas were required due to the lack of computational tools, which are now available to everyone. Overcoming this bottleneck allows us today to both choose a range of more appropriate priors and to easily adapt the credibility framework to a multivariate setting via penalized regression.

A.4. Practical Comparison

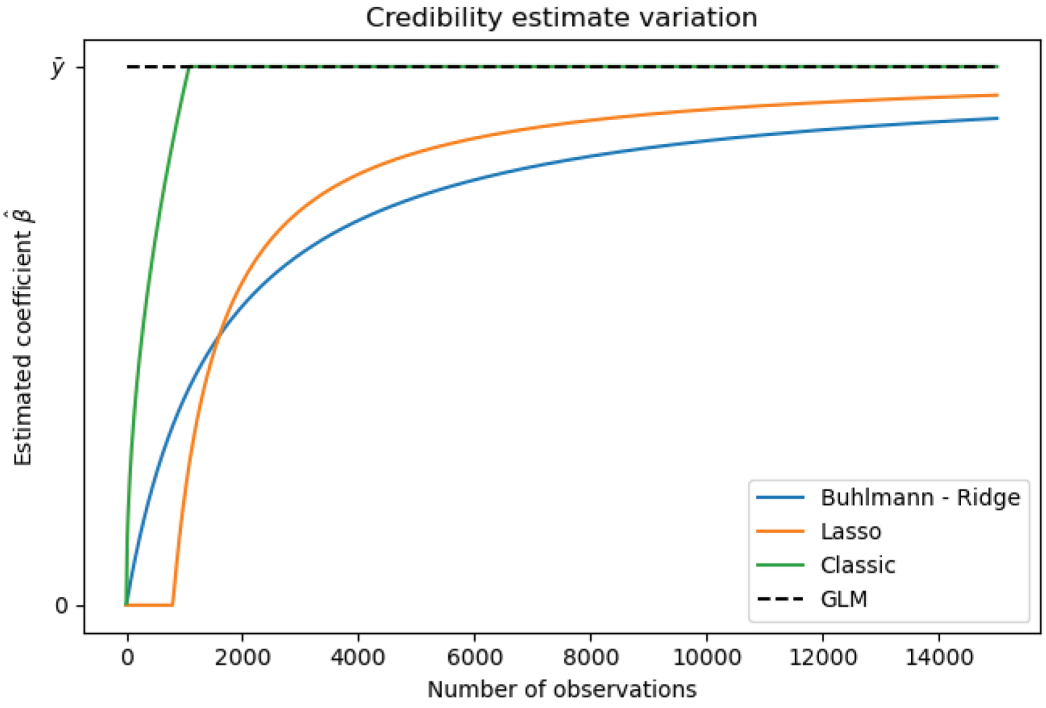
Ridge and lasso, viewed as credibility methodologies, can benefit from a comparison with other credibility methods. We consider two comparison scenarios:

1. Increasing underlying exposures and fixed observed average
2. Increasing underlying average and fixed exposures

A.4.1. Comparison with Increasing Exposures (Fixed Observed Average)

A meaningful comparison can be given by showing how the credibility estimates evolve as the underlying amount of exposures increases. Figure A.4 displays the evolution

Figure A.4. Plot of the estimates $\hat{\beta}$ with $\bar{y} = 5$ and the number of observations varying. Parameters for each model were as follows: classical credibility $N_{full} = 1{,}082$, Bühlmann $k = 1{,}600$, ridge $\lambda_{Ridge} = 1/k$, lasso $\lambda_{Lasso} = 4{,}000$. The formulas used to compute the estimates for the use case can be found in the relative section in the paper. The parameters were chosen arbitrarily to best display the differences of trend. Depending on the parameters, the curve would more or less be similar.



of the estimates of an effect with an increasing number of observations (e.g., a workers' compensation class code risk) when the average $\bar{y}$ for such effect is kept fixed.

The behavior of the lasso is different from the others, as it exhibits a minimum amount of observations (here $\lambda_{Lasso}/\bar{y}$) for which the recorded experience does not influence the estimates $\hat{\beta}$ and the predictions are equal to the complement of credibility. This can be seen as equivalent to considering a specific level of significance to include a variable in a GLM. The choice of including or excluding a specific level in a GLM may be based upon the result of a statistic (p-value) that will depend, among others, on the number of observations. If the statistic is below a certain threshold (e.g., 5% for p-values), then the factor will be included in the modeling, giving an underlying 100% credibility. The lasso regression leads to similar results, but it allows interpolation between 0 and full credibility instead of a binary split (a yes-or-no decision).

When the number of observations is above the lasso's threshold, experience gains weight onto the final estimate much faster than in the ridge/Bühlmann estimates.

The visualization in Figure A.4 shows, in a univariate example, how the signal is interpolated from the complement of credibility to the observed data by credibility

and penalized regression. It also shows how both frameworks can be derived from a Bayesian prior hypothesis on the coefficients' distribution, demonstrating how the penalized regression approach extends GLMs by integrating credibility in a multivariate context.

A.4.2. Comparison with an Increasing Observed Average (Fixed Exposure)

We now focus on comparing the estimates of GLM, ridge, and lasso when the number of observations are fixed but the observed average effect $\bar{y}$ changes. This toy example is helpful as it provides a sense of how these methodologies "learn" from the observed data.

First, we need to compute the estimates of each of the methodologies, i.e., the values of β_j as a function of $\bar{y}_j$.

GLM Solution

Because a **GLM** gives 100% credibility to the data,

$$\beta_{\text{GLM},j} = \bar{y}_j.$$

Ridge Solution

To compute the **ridge** estimate, we need to solve for $\hat{\beta}$:

$$\hat{\beta}_{\text{Ridge}} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{n}\left(y - X\beta\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{p}\beta_j^2.$$

This solution can be found by means of computing the gradient and setting it to zero, similarly to the GLM. The solution is given by the vector that sets the gradient to zero, that is,

$$\hat{\beta}_{\text{Ridge},j} = \frac{n_j}{n_j + \lambda}\bar{y}_j.$$

The addition of the penalty term effectively "shrinks" the observed estimates $\bar{y}_j$ by a number that is dependent on the number of observations of the segment in question.

Lasso Solution

The same procedure should be applied to the **lasso,** which solves the formula

$$\hat{\beta}_{\text{Lasso}} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{n}\left(y - X\beta\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|.$$

We prove in Section D.1 that

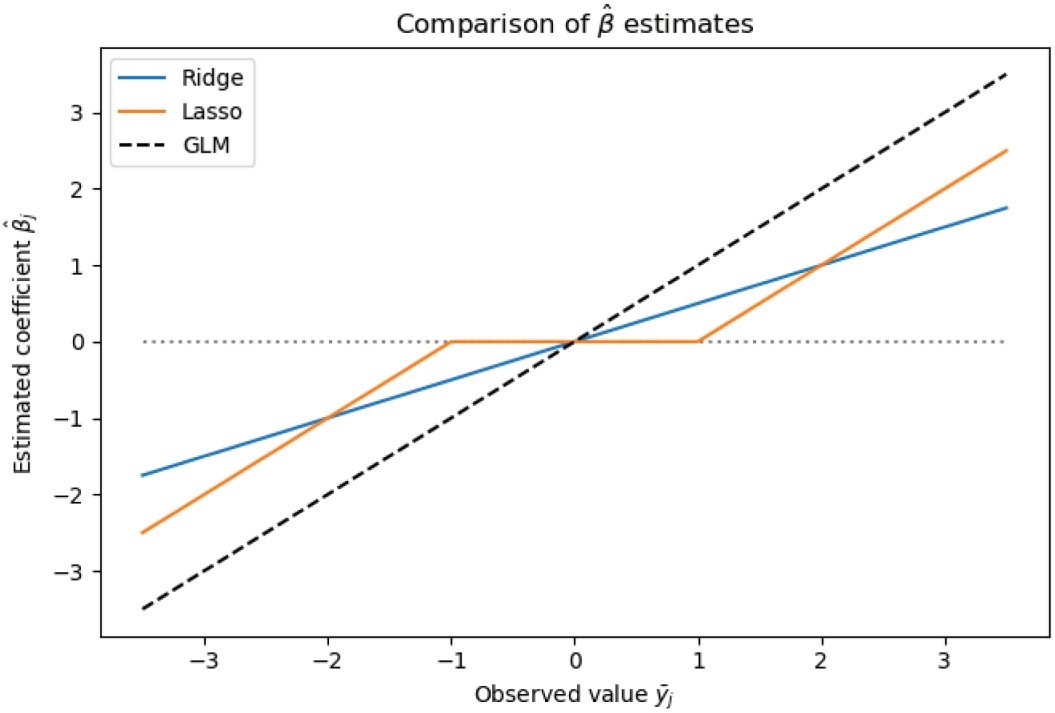
$$\hat{\beta}_{\text{Lasso},j} = \begin{cases} \bar{y}_j - \dfrac{\lambda}{n_j} & \text{if } \bar{y}_j > \dfrac{\lambda}{n_j} \\ \bar{y}_j + \dfrac{\lambda}{n_j} & \text{if } \bar{y}_j < -\dfrac{\lambda}{n_j} \\ 0 & \text{otherwise.} \end{cases}$$

Whenever the quantity $n_j\bar{y}_j = \bar{y}_i < \lambda$, the coefficient for the class j will be set equal to zero, hence assigning no credibility to the data if the quantity of signal is not relevant enough. Appendix D explores in depth the modeling consequences of the structure of the lasso solution.

A.4.3. Final Comparison

Figure A.5 displays how the different estimates β_j differ in the workers' compensation use case for fixed λ, n_j.

Figure A.5. Plot of the correspondence between observed value $\bar{y}_j$ with estimates $\hat{\beta}_j$ for GLM and penalized regression. For GLMs, the dashed line represents the identity function. For the ridge, the relationship is linear, by a factor of $\dfrac{n_j}{n_j + 2\lambda}$. For the lasso, the relationship is piecewise linear.



Compared to the GLM estimates ($\hat{\beta}_j = \bar{y}_j$), the lasso reduces the coefficient by a factor λ/n_j accordingly to the sign of $\bar{y}_j$. If the value of $\bar{y}_j$ is lower than such a threshold, then such value is set to zero.

Compared to the ridge estimates ($\hat{\beta}_j = \frac{n_j}{n_j + 2\lambda}\bar{y}_j$), the lasso shrinks high values of the observed $\bar{y}_j$ to a lesser degree (since lasso penalty grows linearly with the coefficients, and ridge quadratically), but does the opposite for small values of the observed (setting them to exactly zero).

A.5. Degrees of "Bayesian-ness"

We showed that Bühlmann credibility and ridge regression coincide when we do compute the Bayesian estimates via the maximum a posteriori, or MAP, formula. The MAP formula is not the only way to compute the estimates of Bayesian models—there are various other ways to do it.

As mentioned earlier, Bayesian statistics treats parameters as distributions themselves (the "posterior" distribution), and in general, penalized regression provides as prediction the mode of such estimators. Another typical estimator could be the posterior mean, and it can be proven to coincide with the mode in the case of the Bühlmann Gaussian example mentioned in the previous sections.

Ideally, we would be interested in computing the whole posterior distribution of the parameters β, but outside of a limited number of convenient combinations of likelihoods and relative conjugate priors, the posterior distribution might not be known analytically. Inference can quickly become intricate when delving into details. For instance, while we set prior distributions for the parameter β, those priors (like the normal distribution) depend on additional parameters, such as the standard deviation σ. In Bayesian statistics, these metaparameters might have their own prior assumptions. To maintain practicality, various approximations are available to modelers based on desired complexity. As Murphy (2012) illustrated, a hierarchy can be constructed where the more integrals executed, the "more Bayesian" the approach becomes (Table A.2).

The hierarchy serves as a guide through various actuarial methodologies that incorporate Bayesian statistics to varying extents. GLMs are based on the maximum likelihood method, while penalized regressions are MAP estimates of Bayesian models.

Table A.2. From Murphy (2012)

Method	Definition
Maximum likelihood	$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}} p(D\|\boldsymbol{\theta})$
MAP estimation	$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}} p(D\|\boldsymbol{\theta}) p(\boldsymbol{\theta}\|\boldsymbol{\eta})$
ML-II (empirical Bayes)	$\hat{\boldsymbol{\eta}} = \text{argmax}_{\boldsymbol{\eta}} \int p(D\|\boldsymbol{\theta}) p(\boldsymbol{\theta}\|\boldsymbol{\eta}) d\boldsymbol{\theta} = \text{argmax}_{\boldsymbol{\eta}} p(D\|\boldsymbol{\eta})$
MAP-II	$\hat{\boldsymbol{\eta}} = \text{argmax}_{\boldsymbol{\eta}} \int p(D\|\boldsymbol{\theta}) p(\boldsymbol{\theta}\|\boldsymbol{\eta}) p(\boldsymbol{\eta}) d\boldsymbol{\theta} = \text{argmax}_{\boldsymbol{\eta}} p(D\|\boldsymbol{\eta}) p(\boldsymbol{\eta})$
Full Bayes	$p(\boldsymbol{\theta}, \boldsymbol{\eta}\|D) \propto p(D\|\boldsymbol{\theta}) p(\boldsymbol{\theta}\|\boldsymbol{\eta}) p(\boldsymbol{\eta})$

Generalized linear mixed models, which are examples of empirical Bayes estimates, were popularized in the actuarial realm by Klinker (2011). In practice, GLMs share many similarities with ridge penalized regression. Both use Gaussian priors $\beta \sim \mathcal{N}(0, \sigma^2)$ and yield shrunken parameter estimates for β. Full Bayes models, on the other hand, necessitate specialized coding languages or Bayesian tools like Stan.

It's worth noting that the sparsity lasso introduces applies to maximum posterior estimates and doesn't necessarily extend to entire distributions.

Concluding this section on Bayesian interpretation, it's essential to recognize that there is no single best approach. Each method has its merits depending on the use case. While more Bayesian models might excel in certain scenarios, our monograph's purpose is to introduce sophistication to address practical challenges within our statistical methodology.

GLMs, widely accepted and comprehended within the actuarial community, have set a foundation. Penalized regressions, however, emerge as a natural progression from GLMs. They not only introduce a touch of Bayesian thinking, such as prior assumptions and the complement of credibility, but also bridge the gap between traditional statistical methods and machine learning techniques. This intersection with Bayesian statistics offers a pragmatic balance, especially considering that penalized regressions require tuning just a single parameter.

In the authors' view, this blend of simplicity and sophistication makes penalized regression an invaluable tool, serving as a practical introduction to more advanced Bayesian concepts within the standard approach.

Appendix B. Alignment of Lasso Credibility with ASOP 25

Lasso credibility is a new technique, but it aligns with the existing considerations in Actuarial Standard of Practice No. 25, *Credibility Procedures.* In this appendix, we spend time detailing that alignment. First, we introduce how the existing terminology of ASOP 25 can be applied when using the default complement in lasso credibility. Then, we confirm that lasso credibility is in ASOP 25's defined scope and show that lasso credibility has all of the characteristics of an actuarially sound credibility procedure. Finally, this section will confirm that these conclusions hold when the selected complement of credibility is applied through the offset and differs from the default assumption of lasso credibility.

B.1. Definitions: Default Complement of Lasso Credibility

Our first consideration is to verify that penalized regression is a credibility procedure by definition. Note that for the purposes of discussion, in this appendix the displayed text comes from the relevant sections of ASOP 25.

> *2.2 Credibility Procedure*
> *A process that involves the following:*
>
> *a. The evaluation of subject experience for potential use in setting assumptions without reference to other data; or*
> *b. The identification of relevant experience and the selection and implementation of a method for blending the relevant experience with the subject experience.*

When using the default complement of credibility (no offset), we define the above-mentioned terminology as follows:

- **Relevant experience:** the overall average relativity of a particular segment
- **Subject experience:** the experienced relativity of a particular segment
- **A method for blending the relevant and subject experience:** the penalized regression framework through the application of a penalty while maximizing the likelihood during the fitting process

Therefore, lasso credibility is a *credibility procedure* under ASOP 25's definition of the term.

B.2. Considerations and Scope of ASOP 25

Now let's look at Section 1.2.c of ASOP 25 to confirm that penalized regression is within the scope set out by ASOP 25.

> *This Standard applies to actuaries when performing actuarial services involving credibility procedures in the following situations:*
>
> *[. . .]*
>
> *c. When the actuary is blending or considering blending subject experience with other experience; [. . .]*

As lasso credibility is a credibility procedure and we are blending the subject experience with other relevant experience, it is within the scope of ASOP 25. Having determined that lasso credibility is within the standard's scope, we should now further scrutinize the application of the default complement of credibility to make sure our relevant experience satisfies the other considerations of ASOP 25.

2.4 Relevant Experience

> *Sets of data, that include data other than the subject experience, that, in the actuary's judgment, are predictive of the parameter under study (including but not limited to loss ratios, claims, mortality, payment patterns, persistency, or expenses). Relevant experience may include subject experience as a subset.*

When modeling using our simulated countrywide data, we assume we have no prior knowledge for the risk relativities of the variables we include in our model. Therefore, a 1.0 relativity (there is no relationship between the characteristic and true risk) is appropriate. In general, the 1.0 assumption will be valid in the absence of other information.

The text of ASOP 25's Section 3.3, "Selection of Relevant Experience," can be split into four main points.

> *The actuary should use care in selecting the relevant experience.*

This first section is the most important. Below, we make generalizations that are appropriate for most modeling exercises. However, data, projects, and models will always have at least one uniqueness that will require special treatment. Please treat the guidance below as guidance for a "normal" project, and take care to identify situations that might require exceptions to these generalities. We continue to assume for now that we have no prior knowledge about the variable being evaluated.

Moving on to the second point:

> *Such relevant experience should have characteristics similar to the subject experience. Characteristics to consider include items such as demographics, coverages, frequency, severity, or other determinable risk characteristics that the actuary expects to be similar to the subject experience. If the proposed relevant experience does not meet and cannot be adjusted to meet such criteria, it should not be used.*

Satisfying this requirement relies on the assumption that we have no prior knowledge for a particular characteristic and that our data is sufficiently homogeneous to be modeled together in the first place. Let's look at the next point:

> *The actuary should apply credibility procedures that appropriately consider the characteristics of both the subject experience and the relevant experience.*

The blending of subject and relevant experience in penalized regression is determined by the penalty parameter. The penalty parameter behaves similarly to Bühlmann credibility's k parameter as demonstrated in Appendix A. We suggest that if a model fitted through the maximization of likelihood is appropriate for this analysis, then a lasso credibility approach is an appropriate method of blending the subject and relevant experience. Now, to the fourth point:

> *The actuary should consider the extent to which subject experience is included in relevant experience. If subject experience data is a material part of relevant experience, the use of that relevant experience may not be appropriate. In some instances, no relevant experience is available to the actuary. In this situation, the actuary should exercise professional judgment, considering available subject experience, in setting an estimate of expected values.*

The goal of this consideration is to ensure that the complement of credibility is not self-deterministic. For example, if the subject experience is 80% of the relevant experience, and the subject and relevant experience are given 50% weight each, we are really giving 90% of the credibility to the subject experience.

The model structure when using the default complement of credibility will automatically fulfill these considerations as the selected complement is not influenced at all by our modeling data. The "null relativity" assumption for categorical and ordinal variables is truly independent of our subject experience. This consideration will be highly material only when applying a selected nondefault complement of credibility through the use of an offset.

We discuss one more item before moving on, ASOP 25's Section 3.2.c.

> *The actuary should use an appropriate credibility procedure when determining if the subject experience is fully credible or when blending the subject experience with relevant experience. The procedure selected or developed may be different for different practice areas and applications. Additional review may be necessary to satisfy applicable law.*
>
> *In selecting or developing a credibility procedure, the actuary should consider the following criteria:*
>
> *[. . .]*
>
> *c. whether the procedure is practical to implement while taking into consideration both the cost and benefit of employing a procedure.*

The use of penalized regression carries additional computation cost. The time required to fit an individual penalized regression model with a fixed penalty value is quite similar to that needed for an equivalent unpenalized GLM. The computation cost of penalized regression increases because of the need to test various penalty parameters in a cross-validation routine. Such time can be costly on very large data sets, but the ubiquity of cloud computing has made penalized regression more accessible on data sets of all sizes.

The preceding sections justify that the default assumptions of lasso credibility are appropriate for a "normal" actuarial loss model when there is no prior knowledge about a given variable. In the next section, we explore how ASOP 25 applies when using a different complement in lasso credibility.

B.3. Alternate Complements in Lasso Credibility

When using a nondefault complement of credibility, as in the case study, we need to realign our terminology with that of ASOP No. 25. That alignment is the same as earlier with the exception of our relevant experience, which we define as follows:

- **Relevant experience:** relativities produced by our lasso credibility model on the full countrywide data set

Therefore we revisit only ASOP 25's guidance relating to selection of relevant experience.

> *2.4 Relevant Experience*
>
> *Sets of data, that include data other than the subject experience, that, in the actuary's judgment, are predictive of the parameter under study (including but not limited to loss ratios, claims, mortality, payment patterns, persistency, or expenses). Relevant experience may include subject experience as a subset.*

In our case study we assumed that the countrywide data is similar enough to individual states for the countrywide model to be used as a complement. In the case study, the relevant experience does include the subject experience as a subset.

Section 3.3, "Selection of Relevant Experience," has additional considerations for relevant experience that are satisfied through the selection of current rates or a countrywide model as an offset for a state-specific model.

> *The actuary should use care in selecting the relevant experience. . . . Such relevant experience should have characteristics similar to the subject experience. Characteristics to consider include items such as demographics, coverages, frequency, severity, or other determinable risk characteristics that the actuary expects to be similar to the subject experience. If the proposed relevant experience does not meet and cannot be adjusted to meet such criteria, it should not be used.*

Now to the next relevant point of ASOP 3.3:

> *The actuary should consider the extent to which subject experience is included in relevant experience. If subject experience data is a material part of relevant experience, the use of that relevant experience may not be appropriate.*
>
> *In some instances, no relevant experience is available to the actuary. In this situation, the actuary should exercise professional judgment, considering available subject experience, in setting an estimate of expected values.*

Previously, this consideration did not require significant attention as the "null" complement was not related to our subject experience. Now, actuaries should take care to understand the potential influence of their modeling data on the selected complement of credibility. If industry rates or competitor rates are selected, it is unlikely or impossible that the modeling data is a material subset of the selected complement. In our example, the large state data is a subset of the full modeling data set and did have an influence on the creation of the complement of credibility. For our example, we assumed it did not have undue influence. In practice, an actuary should more thoroughly consider the data's influence on the selected complement.

B.4. Lasso Credibility and ASOP 25 Summary

Lasso credibility meets the definition of an actuarial credibility procedure, and the considerations of ASOP 25 should be directly applied when using this methodology. Considerations for the selection of a complement of credibility are not changed when using lasso credibility instead of a different credibility procedure. However, the application of professional judgment in lasso credibility requires both industry knowledge as well as knowledge of penalized regression. This application of judgment in lasso credibility will be most common in the adjustment of the lambda penalty parameter.

Appendix C. Miscellaneous

C.1. Rebasing Model Output

To make modeling output more interpretable and comparable, we often **rebase** the coefficients and factors by modifying the intercept, and we assume a log-link for this example. The term *coefficients* refers to the un-exponentiated betas produced by the model for a particular variable. The term *factors* refers to the exponentiated combination of coefficients with their corresponding X values. This post-modeling rebasing will have no effect on the scoring of the model as the rebasing factor may be incorporated into the intercept.

This is not to say that the selection of the base level during modeling is immaterial. The base level should still be selected to be the most statistically sound in penalized regression. For example, the optimal base level for a categorical variable should be the level with the highest level of exposure. After modeling, factors can be rebased to a selected base level for implementation and the base rate can be adjusted to achieve the desired rate level.

The tables below show an example of rebasing in a model using only one categorical variable with levels A, B, and C. The original base level is A, and we are rebasing to make level B the new base level.

Category	Coefficient	Rebased Coefficient	Factor	Rebased Factor	Rebased Intercept
A	0	$0 - (\beta_1)$	1.0	$\exp(-\beta_1)$	$\exp(\beta_0 + \beta_1)$
B	β_1	$\beta_1 - \beta_1 = 0$	$\exp(\beta_1)$	$\exp(\beta_1 - \beta_1) = 1.0$	$\exp(\beta_0 + \beta_1)$
C	β_2	$(\beta_2 - \beta_1)$	$\exp(\beta_2)$	$\exp(\beta_2 - \beta_1)$	$\exp(\beta_0 + \beta_1)$

Category	Coefficient	Rebased Coefficient	Factor	Rebased Factor	Base Rate	Rebased Base Rate
A	0.0	–.2	1.0	0.819	100	122.1
B	.2	0	1.221	1.0	100	122.1
C	.5	.3	1.549	1.340	100	122.1

The tables represent the mathematical transformations necessary to rebase the coefficient. In the model fit, category A was originally the base level, as highlighted by its 1.0 factor. In the rebased columns, such reference is changed to the factor B, which has a factor of 1.0 in the "Rebased Factor" column.

C.2. Penalized Regression and Near Aliasing

Near aliasing can cause large instabilities in GLMs, but models using penalized regression do not exhibit such volatility. We can explain this through the lens of credibility. Take for example two indicator variables that overlap entirely except for a handful of characteristics. Those non-overlapping risks are implicitly being given full credibility by the GLM, as one indicator will increase drastically and the other will decrease until both segments are perfectly identified. This extremely high or low prediction for the small segment can greatly skew test statistics. When using lasso penalization, it is highly likely that one of the indicators will be highly penalized, and that the other will be included and will correctly account for the overlapping risks.

C.3. Penalized Regression and the AIC

In the context of model selection in GLMs, the modeler may need to decide on one model between two alternatives. For example, such a comparison could be based on evaluating performances in a holdout, or testing, data set. Another way to compare models is not to use any testing set at all but instead use statistical theory to approximate the generalization power of a model using training data alone. This approach doesn't rely on new data sets to assess the generalization power of the model, avoiding the problems described above.

The Akaike information criterion, or AIC, is one of the most popular of such metrics, whose formula is

$$\mathrm{AIC}(\beta) = 2\mathrm{NLL}(\beta) + \#\text{ of degrees of freedom.}$$

Here, NLL represents the negative log-likelihood computed in the training data, and # of degrees of freedom represents the number of nonzero entries of the GLM coefficient β.

Given two different modeling alternatives, the modeler can compute the AIC for such models. Since the AIC formula "penalizes" the decrease of the NLL of the more complex model with the degrees of freedom (hence the complexity), the model with lower AIC should be preferred.

The curious reader may notice that the AIC formula looks similar to the penalized GLM Equation 3.1:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, \mathrm{NLL}(y, X, \beta) + \lambda\, \mathrm{Penalty}(\beta).$$

Both formulas consider the trade-off between the goodness of fit (negative log-likelihood) and the complexity of the model (in the AIC, the number of degrees of freedom, in penalized regression, the penalty).

The connections between AIC and penalized GLM, with lasso in particular, are much more deep. First, we can define a penalized GLM version of the AIC, by just defining

$$\text{Penalty}(\beta) = \#\text{ of degrees of freedom} = \#\text{ of nonzero entries of } \beta,$$

The solution of this problem is known as "best subset selection" problem. Many authors consider this the gold standard for building models: who would not compute the most performant GLM model given a "complexity" budget determined by the degrees of freedom? The model is not popular because it is proven to be numerically intractable. However, one can prove that the best "numerically tractable"[11] approximation to that problem is the lasso. The mathematically inclined reader can refer to Hastie, Tibshirani, and Tibshirani (2020) for a more in-depth discussion of this approximation result.

That paper also contains an unexpected result: lasso outperforms best subset selection in conditions characteristic of insurance data. This result may sound counter-intuitive, because from a theoretical perspective, best subset selection is superior to lasso. But this does not mean that it will perform better in practice for every (or even a typical) high-dimensional regression problem that we might want to solve. Best subset selection tends to have much higher variance than the lasso, because there is shrinkage inherent in the latter's coefficient estimates. As a result, which estimator performs better in practice really depends on a lot of factors, such as the signal-to-noise ratio.

[11] A problem is considered numerically tractable if an optimization algorithm can provide its optimal solution in a reasonable time. This usually means that the problem is convex.

Appendix D. Sparsity: A Convex Optimization Perspective

The monograph details the intimate connections between penalized regression, with lasso in particular, and credibility, Bayesian statistics, and machine learning. There is one last connection, for the more mathematically inclined reader—that with convex optimization.

Lasso regression is known for its ability to achieve sparse solutions, where some of its estimated coefficients will be exactly equal to zero. The mathematical reasons for why sparsity happens are, however, either not discussed at all or hidden in very technical explanations that require a certain level of familiarity with advanced convex optimization concepts. That does not need to be the case—and this section is an attempt to show that **sparsity is achieved because lasso fits the signal up to a threshold**.

Let's start to evaluate how, in a multivariate setting, the GLM estimates adapt to the data by examining the gradient of the GLM formula.

The gradient of a GLM (with canonical link) can be seen as the difference, for each level of a variable, of the total observations and the total estimates included in the GLM model (Ohlsson and Johansson 2010):

$$\nabla NLL\left(y, X, \beta\right)_j = \frac{\partial NLL\left(y, X, \beta\right)}{\partial \beta_j} = \sum_{i=1}^{n} \left(\mu_i - y_i\right) x_{ij},$$

where $\mu_i = \text{Link}^{-1}(X_i\beta)$ is the prediction for a given β. At the GLM solution $\hat{\beta}$, the gradient is null. In particular, we have

$$\sum_{i \in J} \left(\mu_i - y_i\right) = 0. \tag{D.1}$$

This is consistent with the results already described in Section A.1: GLMs give 100% credibility to the data, and average predictions will coincide with the average of the observations for each level j included in the model (regardless of the number of exposures).

If we want to leave some room for the complement of credibility and shrink the estimates), we could consider adding some "slack," so that the model can match the data only up to a certain threshold. For example $|\sum_{i \in J}(y_i - \mu_i)| \leq \varepsilon$ for a certain value ε. This is exactly how the lasso guarantees optimality.

The following sections will prove that the optimality condition for

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \operatorname{NLL}(y, X, \beta) + \lambda \sum_{j=1}^{p} |\beta_j|$$

is equivalent to (where $\mu_i = X_i\beta$)

$$\begin{cases} \left|\sum_{i \in J} (\mu_i - y_i)\right| \leq \lambda & \longleftrightarrow \quad \hat{\beta}_j = 0 \\ \sum_{i \in J} (\mu_i - y_i) = \lambda \operatorname{sign}(\hat{\beta}_i) & \longleftrightarrow \quad \hat{\beta}_j > 0 \\ \sum_{i \in J} (\mu_i - y_i) = -\lambda \operatorname{sign}(\hat{\beta}_i) & \longleftrightarrow \quad \hat{\beta}_j \leq 0. \end{cases} \tag{D.2}$$

Any final estimate of the lasso, regardless of fitting procedure, will satisfy the above optimality condition system (Equation D.2). We can clearly see the slack discussed above when matching observed and predicted averages: a coefficient will be deemed nonrelevant (and thus set to zero) if its contribution to the likelihood via the gradient falls below the threshold λ; when the effect is instead considered relevant the coefficient will moved to capture it but just until the error tolerance threshold λ is hit, rather than going all the way like it would on a GLM.

The next sections provide a smooth learning curve to understand the origin of Equation D.2. First, by considering a simple example, we show how sparsity naturally arises from the nondifferentiability of the absolute value $|\beta|$ contained in the lasso penalty. Then, we introduce the least amount of concepts from convex optimization necessary to provide the optimality guarantees for the lasso problem.

D.1. Simplified Proof of the Lasso Problem

Consider the simplest possible lasso regression expressed as

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2}(y - \beta)^2 + \lambda |\beta|, \tag{D.3}$$

where we have a one-dimensional parameter β aiming to approximate a single observation y.

Computing the solution by setting the gradient to zero is not possible as the absolute value is nondifferentiable at zero. Instead, one can write the function as a piecewise parabolic function:

$$\frac{1}{2}(y - \beta)^2 + \lambda |\beta| = \begin{cases} \frac{1}{2}(y - \beta)^2 + \lambda\beta & \text{if } \beta \geq 0 \\ \frac{1}{2}(y - \beta)^2 - \lambda\beta & \text{if } \beta < 0. \end{cases} \tag{D.4}$$

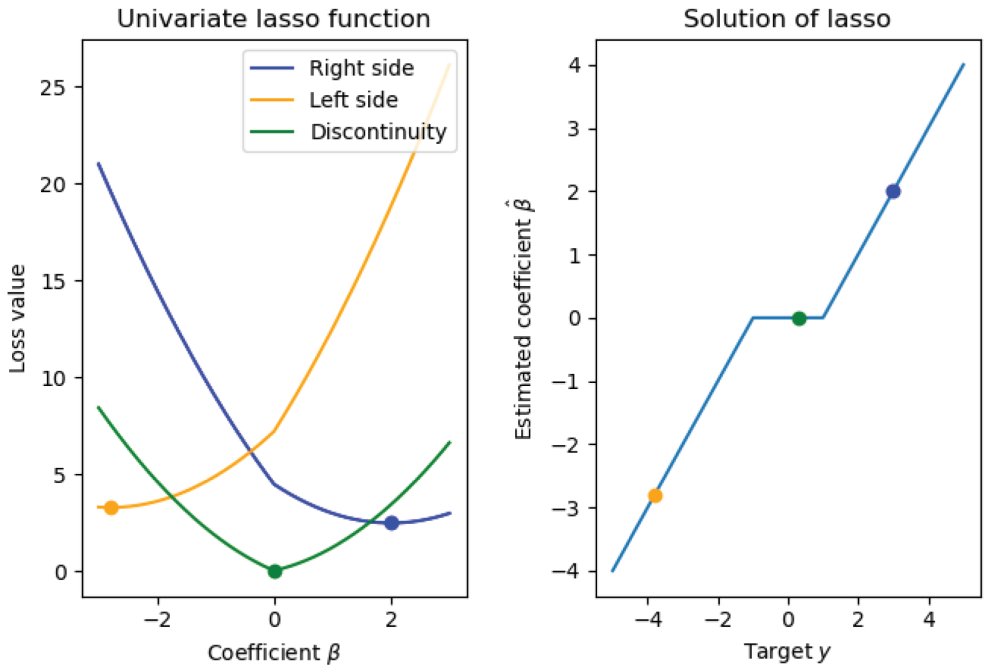
For every value of y and λ, this function is convex, meaning that there is one and only one global minimum. Figure D.1 highlights all the possible cases, depending on the value of y. It is clear that the optimum $\hat{\beta}$ of the piecewise function lies either at the global minimum of the parabola or at $\beta = 0$, where the two parabolas intersect.

We can then deduce that if the minimum of the function lies in the right interval $\beta > 0$, then the optimum will be $\hat{\beta} = y - \lambda$ (which is the global minimum of $\frac{1}{2}(y-\beta)^2 + \lambda\beta$). Equivalently, if it lies on the left part of the parabola ($\beta < 0$), then the optimum will be $\hat{\beta} = y + \lambda$. Combining the inequalities, we prove that the optimum of Equation D.3 is

$$\hat{\beta} = \begin{cases} y - \lambda & \text{if } y > \lambda \\ y + \lambda & \text{if } y < -\lambda \\ 0 & \text{otherwise} \end{cases} \tag{D.5}$$

Figure D.1 shows the plot of the optimum $\hat{\beta}$ as a function of the value of y.

Figure D.1. The left graph represents the plot of three different lasso formulas (Equation 12.4) with different fixed values of y and $\lambda = 1$. The right side represents $\beta \to 1/2(3 - \beta)^2 + |\beta|$, the left side is $\beta \to 1/2(-3.8 - \beta)^2 + |\beta|$, and discontinuity is $\beta \to 1/2(-3.8 - \beta)^2 + |\beta|$. The dots represent the optimum $\hat{\beta}$ for each function. The right graph represents the evolution of the optimum $\hat{\beta}$ as a function of the values y. The colored points represent the couples $(y, \hat{\beta})$ of the three functions of the left-side graph. The function is also called soft-thresholding in the literature.



The example highlights how the discontinuity in the lasso penalty $|\beta|$ achieves a sparse solution. The lasso estimate will be exactly equal to zero in correspondence of values of y smaller than λ in absolute value and will be equal to the value of y shrunken by a constant of size λ otherwise. It is thanks to its nondifferentiable nature that the lasso problem allows us to obtain sparse solutions.

We now review how to demonstrate the lasso solution in a general way using simple tools from convex optimization.

D.2. General Proof of the Lasso Problem

When required to find a minimum of a function analytically, the practitioner would naturally compute the gradient of that function and find the parameter β that solves the equality of the gradient to zero.

In the case of the lasso problem, we saw how this is not possible due to the nondifferentiability of the penalty at $\beta = 0$. As a matter of fact, it is still possible to compute a minimum of the lasso regression by setting the gradient to zero: we just need to generalize the definition of the gradient.

The gradient is defined as the slope of the tangent to the graph of a function. In the case of a discontinuity, there may be multiple slopes that are tangent to the graph of the function. The gradient loses its uniqueness property, and it is hence said that the function is "not differentiable."

A generalization of the gradient, the subgradient, is defined as the *set* of possible slopes that are tangent to a graph. Formally, given a convex function $f \in \mathbb{R}^p$, the subgradient is

$$\partial f(\beta_0) = \left\{ u \in \mathbb{R}^p \,|\, f(\beta) - f(\beta_0) \geq u(\beta - \beta_0) \right\}.$$

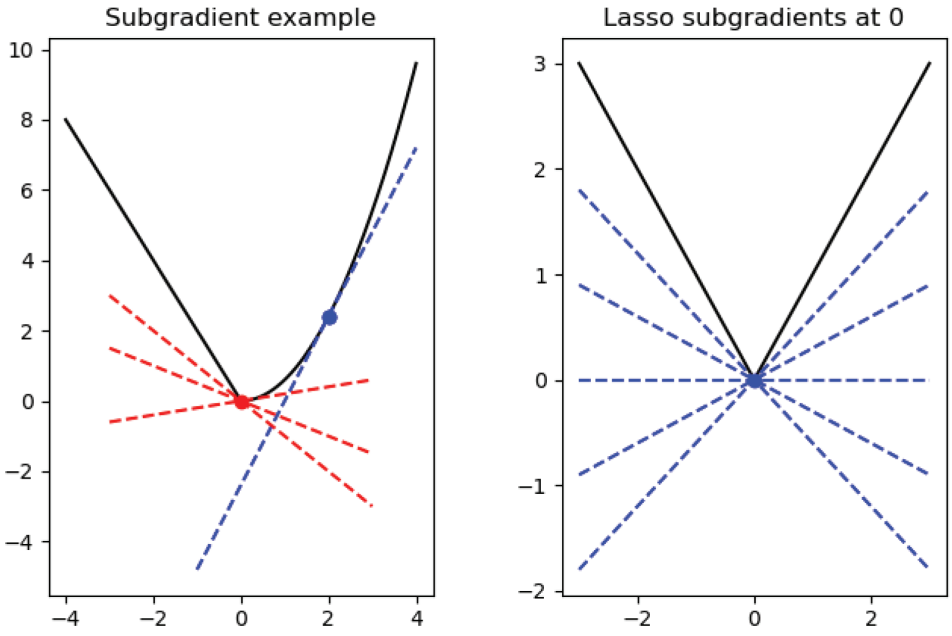
In the case of the absolute value function, since the subgradient is equal to the gradient when the function is differentiable, for all values strictly different than zero the gradient will be equal to the sign function, that is, 1 for all positive values and -1 for all negative values. In the discontinuity point at 0, it will take all possible values between -1 and 1.

$$\partial|\beta| = \begin{cases} -1 & \text{if } \beta < 0 \\ (-1, 1] & \text{if } \beta = 0 \\ 1 & \text{if } \beta > 0 \end{cases} \tag{D.6}$$

Generalizing the gradient to the subgradient allows us to compute the minimum for the lasso. It is established that if f is differentiable and convex, then

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} f(\beta) \quad \Leftrightarrow \quad 0 = \nabla f\left(\hat{\beta}\right). \tag{D.7}$$

Figure D.2. The graph on the left illustrates the subgradient for a piecewise function. In blue, we see the subgradient at value 2. As the function is differentiable, there exists only one subgradient, and it is equal to the gradient. Since the function at 0 is not differentiable, there is more than one tangent line to the graph. The subgradient is drawn in red. The graph on the right represents some possible tangent lines for the lasso function $\beta \to |\beta|$. Thus, we have a visual intuition of why $\partial|\beta|_{\beta=0} = [-1,1]$.



If f is not differentiable (but still convex), then the optimality condition becomes

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} f(\beta) \quad \Leftrightarrow \quad 0 \in \partial f\left(\hat{\beta}\right) \tag{D.8}$$

by the subgradient optimality condition (see Boyd and Vandenberghe 2004). Since the subgradient of a sum is the sum of the subgradients and the subgradient of a differentiable function is the gradient, in the case of the (simplified) lasso regression we can write the condition #eq-subdiffoptimal as

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2}(y-\beta)^2 + \lambda|\beta| \quad \Leftrightarrow \quad 0 \in \left(\hat{\beta} - y\right) + \lambda\partial\left|\hat{\beta}\right|,$$

where $(\hat{\beta} - y)$ is the derivative of $\frac{1}{2}(y-\beta)^2$.

Since the subgradient of $\beta \to |\beta|$ is given by Equation D.6, we can prove the optimality of Equation D.5: if the optimum is $\hat{\beta} = 0$, then there exists a number $|u| \leq 1$ such that $0 = -y + \lambda u$. This happens only when $|y| \leq \lambda$. For the other cases ($\hat{\beta} > 0$, $\hat{\beta} < 0$)

the lasso penalty is differentiable and by standard arguments one can verify the optimality of Equation D.5.

The subgradient definition provides the optimality conditions of the lasso regression in all its generality, both in a multivariate setting and using a generic negative log-likelihood. It provides as well the tools to understand the optimality conditions from Equation D.2. To see this, consider the general lasso problem

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \operatorname{NLL}\left(y, X, \beta\right) + \lambda \sum_{j=1}^{p} \left|\beta_j\right|.$$

To compute the optimal solution, first the **sub**gradient of the negative log-likelihood is required:

$$\frac{\partial}{\partial \beta_j}\left[\operatorname{NLL}\left(y, X, \beta\right) + \lambda \sum_{j=1}^{p} \left|\beta_j\right|\right] = \nabla NLL\left(y, X, \beta\right)_j + \lambda \partial \left|\beta_j\right|.$$

At optimum $\hat{\beta}$ zero must belong to the subgradient, which means that depending on the sign of $\hat{\beta}_j$ we have that

$$\begin{cases} \left|\nabla NLL\left(y, X, \beta\right)_j\right| \leq \lambda & \Leftrightarrow \quad \hat{\beta}_j = 0 \\ \nabla NLL\left(y, X, \beta\right)_j = \lambda \operatorname{sign}\left(\hat{\beta}_i\right) & \Leftrightarrow \quad \hat{\beta}_j > 0 \\ \nabla NLL\left(y, X, \beta\right)_j = -\lambda \operatorname{sign}\left(\hat{\beta}_i\right) & \Leftrightarrow \quad \hat{\beta}_j \leq 0, \end{cases}$$

which proves Equation D.5. The results of this section combined provide as well all required tools to prove the optimal solution formula.

References

Boor, J. A. 1996. "The Complement of Credibility." *Proceedings of the Casualty Actuarial Society* 83: 1–40. https://www.casact.org/sites/default/files/database/forum_95fforum_95ff323.pdf.

Boyd, S. P., and L. Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.

Bühlmann, H., and A. Gisler. 2005. *A Course in Credibility Theory and Its Applications*. Vol. 317. Berlin: Springer.

Casella, G., M. Ghosh, J. Gill, and M. Kyung. 2010. "Penalized Regression, Standard Errors, and Bayesian Lassos." *Bayesian Analysis* 5: 369–411.

Casotto, M., M. Banterle, and G. Beraud-Sudreau. 2022. "Credibility and Penalized Regression." Akur8 white paper - available at www.akur8.com/white-papers/credibility-and-penalized-regression

Casotto, M., and T. Holmes. 2023. "Derivative Lasso: Credibility-Based Signal Fitting for GLMs." Akur8 whitepaper – available at www.akur8.com/white-papers/derivative-lasso-credibility-based-signal-fitting-for-glms

Casualty Actuarial and Statistical (C) Task Force. 2020. *Regulatory Review of Predictive Models.* White paper. National Association of Insurance Commissioners.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani. 2004. "Least Angle Regression." *Annals of Statistics* 32 (2): 407–51.

Fujita, S., T. Tanaka, K. Kondo, and H. Iwasawa. 2020. "AGLM: A Hybrid Modeling Method of GLM and Data Science Techniques." Institut des Actuaires. Available at https://www.institutdesactuaires.com/global/gene/link.php?doc_id=16273&fg=1

Goldburd, M., A. Khare, D. Tevet, and D. Guller. 2016. *Generalized Linear Models for Insurance Rating.* CAS Monograph Series, No. 5. Arlington, Va.: Casualty Actuarial Society.

Greenland, S., S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N. Goodman, and D. G. Altman. 2016. "Statistical Tests, *P* Values, Confidence Intervals, and Power: A Guide to Misinterpretations." *European Journal of Epidemiology* 31: 337–50.

Hastie, T., R. Tibshirani, and J. H. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer Science+Business Media.

Hastie, T., R. Tibshirani, and R. Tibshirani. 2020. "Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons." Statistical Science 35 (4): 579–92.

Hastie, T., R. Tibshirani, and M. Wainwright. 2015. *Statistical Learning with Sparsity: The Lasso and Generalizations*. New York: CRC Press.

Hoerl, A. E. 1962. "Application of Ridge Analysis to Regression Problems." *Chemical Engineering Progress* 58: 54–59.

Iwasawa, H., and G. Wang. *Introduction to Accurate GLM.* 2022. 2022 CAS RPM Virtual Seminar. Willis Towers Watson. https://www.casact.org/sites/default/files/2022-07/M1_IntrotoGLM.pdf.

Jewell, W. S. 1974. "Credible Means Are Exact Bayesian for Exponential Families." *ASTIN Bulletin* 8 (1): 77–90.

Klinker, F. 2011. "Generalized Linear Mixed Models for Ratemaking: A Means of Introducing Credibility into a Generalized Linear Model Setting." *Casualty Actuarial Society E-Forum,* Winter. https://www.casact.org/sites/default/files/2021-02/pubs_forum_11wforumpt2_klinker.pdf.

Miller, H. 2015. "A Discussion on Credibility and Penalised Regression, with Implications for Actuarial Work." Presented to the Actuaries Institute 2015 ASTIN, AFIR/ERM and IACA Colloquia, August 23–27, Sydney.

Mosley, R., and R. Wenman. 2022. *Methods for Quantifying Discriminatory Effects on Protected Classes in Insurance.* CAS Research Paper: Series on Race and Insurance Pricing. Arlington, Va.: Casualty Actuarial Society.

Murphy, K. P. 2012. *Machine Learning: A Probabilistic Perspective*. Cambridge, Mass.: MIT Press.

Ohlsson, E., and B. Johansson. 2010. *Non-Life Insurance Pricing with Generalized Linear Models*. EAA Series. Berlin: Springer-Verlag.

Shi, S. G. 2010. "Direct Analysis of Pre-Adjusted Loss Cost, Frequency or Severity in Tweedie Models." *Casualty Actuarial Society E-Forum*, Winter. https://www.casact.org/sites/default/files/database/forum_10wforum_shi.pdf.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58 (1): 267–88.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight. 2005. "Sparsity and Smoothness via the Fused Lasso." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67 (1): 91–108.

Tse, Y.-K. 2009. *Nonlife Actuarial Models: Theory, Methods, and Evaluation*. Cambridge, UK: Cambridge University Press.

van Wieringen, W. N.. 2015. "Lecture Notes on Ridge Regression." arXiv:1509.09169. https://arxiv.org/abs/1509.09169.

Venter, G. n.d. "Bayesian Regularization for Class Rates." Available at http://www.garyventer.com/wp-content/uploads/2018/09/Venter-2018-Bayesian_Regularization_Ratemaking.pdf

Wasserstein, R. L., and N. A. Lazar. 2016. "The ASA Statement on *p*-Values: Context, Process, and Purpose." *American Statistician* 70 (2): 129–33.

Wüthrich, M. V., and M. Merz. 2023. *Statistical Foundations of Actuarial Learning and Its Applications*. Cham, Switzerland: Springer Nature.

Yan, J., J. Guszcza, M. Flynn, and C.-S. P. Wu. 2009. "Applications of the Offset in Property-Casualty Predictive Modeling." *Casualty Actuarial Society E-Forum*, Winter. https://www.casact.org/sites/default/files/database/forum_09wforum_yan_et_al.pdf.

ABOUT THE SERIES:

CAS monographs are authoritative, peer-reviewed, in-depth works focusing on important topics within property and casualty actuarial practice. For more information on the CAS Monograph Series, visit the CAS website at www.casact.org.

Expertise. Insight.
Solutions.

casact.org