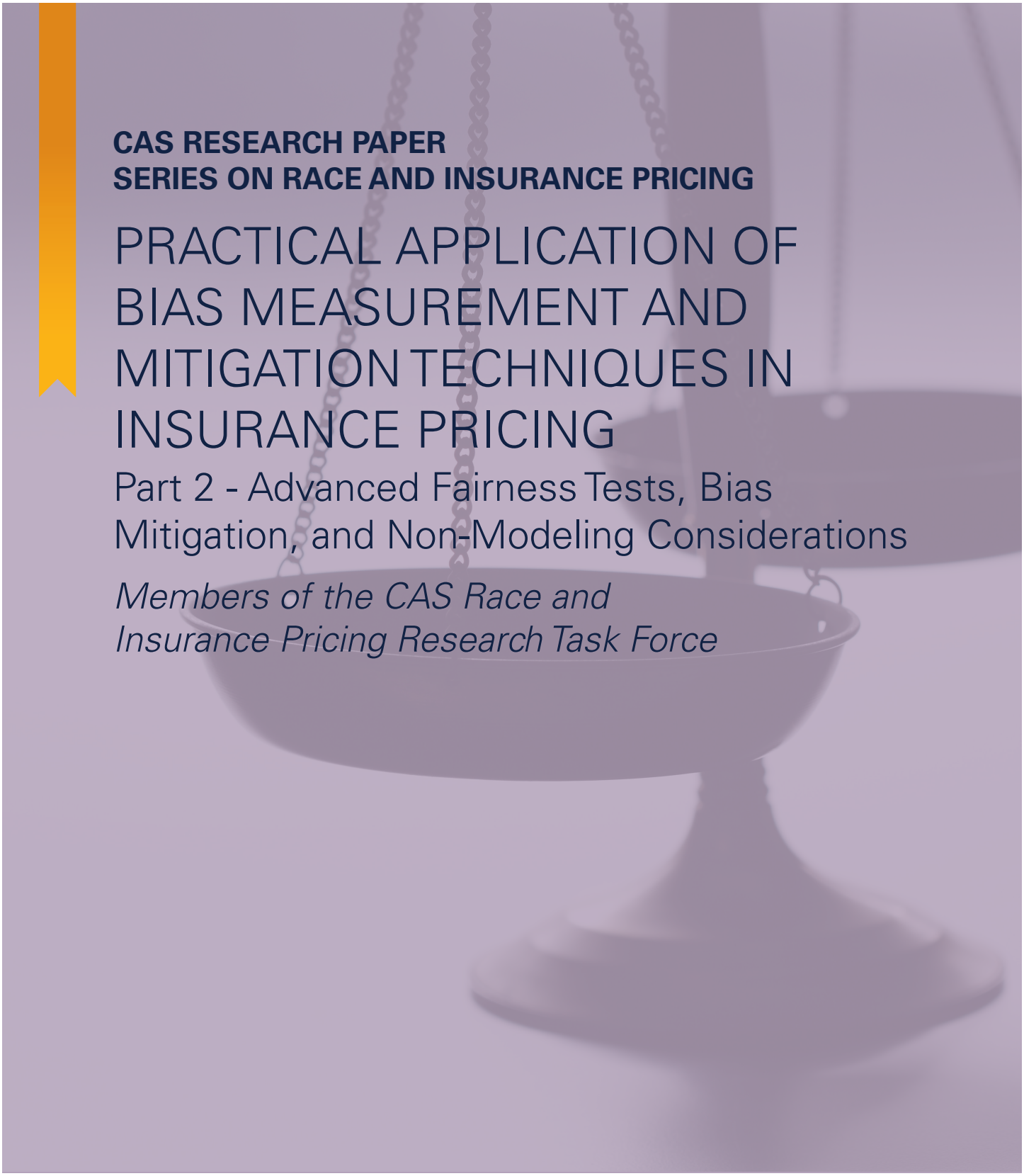


CAS RESEARCH PAPER
SERIES ON RACE AND INSURANCE PRICING

PRACTICAL APPLICATION OF BIAS MEASUREMENT AND MITIGATION TECHNIQUES IN INSURANCE PRICING

Part 2 - Advanced Fairness Tests, Bias Mitigation, and Non-Modeling Considerations

Members of the CAS Race and Insurance Pricing Research Task Force

CASUALTY ACTUARIAL SOCIETY

The CAS Research Paper Series on Race & Insurance Pricing was created to guide the insurance industry toward proactive, quantitative solutions that address potential racial bias in insurance pricing. These reports explore different aspects of unintentional potential bias in insurance pricing, address historical foundations and offer forward-looking solutions to quantify and handle possible bias. Through these reports, the CAS aims to encourage actuaries to discuss this topic with their stakeholders across all areas of insurance pricing and operations. For more information on the series, visit casact.org/raceandinsuranceresearch.

The Casualty Actuarial Society (CAS) is a leading international organization for credentialing, professional education and research. Founded in 1914, the CAS is the world's only actuarial organization focused exclusively on property-casualty risks and serves over 10,000 members worldwide. CAS members are sought after globally for their insights and ability to apply analytics to solve insurance and risk management problems. As the world's premier P&C actuarial research organization, the CAS reaches practicing actuaries across the globe with thought-leading concepts and solutions. The CAS has been conducting research since its inception. Today, the CAS provides thousands of open-source research papers, including its prestigious publication, Variance — all of which advance actuarial science and enhance the P&C insurance industry. Learn more at casact.org.

CAS RESEARCH PAPER
SERIES ON RACE AND INSURANCE PRICING

CAS RACE AND INSURANCE PRICING RESEARCH TASK FORCE

Mallika Bender, FCAS; Margaret (Peggy) Brinkmann, FCAS, CSPA; Eric Krafcheck, FCAS, CSPA; Craig Sloss, PhD, FCAS, FCIA; Gary Wang, FCAS, CSPA; Mike Woods, FCAS, CSPA

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

Contents

Practical Application of Bias Measurement and Mitigation Techniques in Insurance Pricing: Part 2 – Advanced Fairness Tests, Bias Mitigation, and Non-Modeling Considerations

By Members of the Casualty Actuarial Society Race and Insurance Pricing Research Task Force

Introduction to Part 2

Industry views on fairness in insurance pricing are evolving to include both the traditional understanding that insurance rates should not be "unfairly discriminatory" – that is, they should reflect differentials in risk among policyholders – and recognition that insurance rating may potentially result in "discriminatory effects" where certain legally protected groups are subject to disproportionately higher or lower insurance rates than others. In the United States, many jurisdictions are taking regulatory and/or legislative action to encourage or require insurers to evaluate their own data and models for both of these types of fairness.[1]

This paper is intended as a practitioners' guide for actuaries and insurance professionals responsible for building, maintaining, or updating insurance pricing models that satisfy multiple views of fairness.

The paper is presented in two parts, **of which this is Part 2**.

Part 1 of the paper can be found at casact.org/raceandinsuranceresearch and consists of three sections:

- Section 1 reviews different types of bias and how they can impact insurance data, model design, implementation, use, and monitoring.
- Section 2 provides practical guidance on applying the Bayesian Improved First Name Surname Geocoding (BIFSG) approach to impute race and ethnicity in cases in which insurers do not or cannot collect the information directly from policyholders.

[1] For more detail on recent regulatory and legislative actions, as of May 2024, refer to the following three papers in the CAS Research Paper Series on Race and Insurance Pricing: (1) "Regulatory Perspectives on Algorithmic Bias and Unfair Discrimination," (2) "A Practical Guide to Navigating Fairness in Insurance," and (3) "Comparison of Regulatory Framework for Non-Discriminatory AI Usage in Insurance."

- Section 3 illustrates the application of fairness tests based on three types of simple fairness criteria – independence, separation, and sufficiency[2] – to continuous insurance pricing models, through premium parity and loss ratio parity tests.

Part 2 of the paper consists of three sections:

Section 4 builds upon the three simple fairness tests illustrated in Part 1, Section 3, by considering approaches that also account for different distributions of protected groups within rating elements. Conditional demographic parity tests allow for a certain group of rating elements to be deemed "acceptable" for insurance rating before calculating parity metrics. The proxy ("control variable") test incorporates protected class into a rating model and evaluates the changes to the relative contribution of all other factors in a model. This provides an indication of which particular rating elements may be acting as a "proxy" for the protected dimension. This section closes by describing the nonparametric matching approach, which attempts to isolate the impact of one rating variable of interest on protected groups by creating control and treatment groups in which all rating attributes are exactly matched except for the protected attribute, so the impact of the rating variable of interest can be calculated.

Section 5 moves from diagnosing fairness concerns to addressing them. The fairness testing methods explored in Sections 3 and 4 of this paper may provide the actuary with insight into what type of mitigation approach, if any, may be most appropriate to the situation. This section covers multiple mitigation approaches that can be applied to model inputs, within the modeling process, or to model outputs, including their potential benefits and limitations. While there is value in exploring each of these mitigation approaches in the context of insurance pricing, this paper cautions that mitigation measures need to be evaluated carefully to ensure that correction of some discriminatory effects does not introduce new discriminatory effects as a byproduct.

Section 6 closes this practical guide with a discussion of several non-modeling considerations that can contribute to fairness concerns; these include marketing and underwriting decisions that impact modeling data, business decisions and regulatory adjustments to model outputs, and discounts or loads layered on top of the final rating model. Any of these practices could introduce or amplify existing discriminatory effects that may not be immediately apparent when applying fairness tests to the outputs of a rating model. This paper suggests performing these tests on the "final" price charged to policyholders, if that information is readily available and there are material differences between model indications and the final price.

While this paper may be read from start to finish, readers are invited to navigate directly to the part and/or section of the paper that is most relevant to their current responsibilities.

[2] These methods are introduced in the CAS Research Paper Series on Race and Insurance Pricing report, "Methods for Quantifying Discriminatory Effects on Protected Classes in Insurance".

Section 4. Accounting for Distributional Differences in Fairness Tests

Note: This paper is divided into two parts, **of which this is Part 2**. **Part 1** of the paper can be found at casact.org/raceandinsuranceresearch. This section builds upon the discussion presented in Part 1, Section 3, of three simple fairness criteria – independence, separation, and sufficiency – and tests that correspond to each one, by presenting more complex fairness testing methods that can account for distributional differences in insurance pricing data.

4.1. Testing for Disproportionate Impact and Conditional Demographic Parity

A limitation of the premium parity method of assessing fairness is that it does not consider whether disparities between protected groups can be explained by protected groups having differing distributions (i.e., correlations) across rating factors that are widely viewed as acceptable to use in insurance rating. If, for example, average home values vary between protected groups, all else being equal, one would reasonably expect average homeowners premium to also vary accordingly, since insurance companies use home values to establish how much coverage is provided by the policy. By not conditioning on home value when evaluating for bias in a rating model, one might conclude in this example that the homeowners rating plan is biased because it does not result in equal average premiums under an independence fairness criteria. Further, any efforts to mitigate this premium disparity would result in insureds with lower-value homes subsidizing insureds with higher-value homes.

One approach to addressing some of this concern with traditional demographic parity is to use a more general test, conditional demographic parity. This test is based on the idea that there is a specified set of rating factors that are considered unambiguously acceptable, and that after controlling for this set of "acceptable" factors, conditional demographic parity is achieved if the average premium is the same for all groups in the protected class. Conditional demographic parity corresponds to the notion of disproportionate impact, which is defined as occurring when "a rating tool results in higher or lower rates, on average, for a protected class, controlling for other distributional differences" (Chibanda 2022). An advantage of this approach is that it allows for more flexibility by considering the possibility that differences in premium could be explained by legitimate distributional differences of non-protected attributes. A disadvantage is that the need to specify a list of "legitimate" factors introduces a degree of subjectivity into the test, and modelers will generally need to rely on stakeholders outside the modeling team, such as a regulator or internal business partners, to provide this list as input to the technical modeling process.

To illustrate by example, consider an auto rating plan that uses the value of the vehicle and credit-based insurance score as rating factors. Suppose a regulator considers the value of a vehicle to be on the list of "unambiguously acceptable factors" and only allows rating plans that satisfy conditional demographic parity. Credit-based insurance score is not on this list

but is not explicitly forbidden as a rating factor. Conditional demographic parity would allow for a premium disparity among protected groups provided it can be explained by differences in vehicle values among the groups, as vehicle value is always allowed as a rating factor. However, credit-based insurance score would only be permitted as a rating factor if it can be demonstrated that it does not introduce a premium disparity after accounting for the effect of vehicle value.

To formalize this notion, assume that premium is calculated using an algorithm $f(R)$, where R is the full set of rating factors, which does not include the protected class A. Assume that there is a subset D of R that has been identified as a set of "acceptable" rating factors. The conditional demographic parity condition is

$$E\Big[f\big(R, A\big)\Big|D, A = a\Big] = E\Big[f\big(R, A\big)\Big|D, A = b\Big].$$

At the two extremes, when D is empty, this reduces to traditional demographic parity. When D is the set of all rating factors, the condition becomes

$$E\Big[f\big(R, A\big)\Big|R, A = a\Big] = E\Big[f\big(R, A\big)\Big|R, A = b\Big].$$

In this case, since $f(R)$ is fully determined by R, the condition is essentially the same as saying that the algorithm's output does not change based on the value of A, a notion referred to as "fairness through unawareness":

$$f\big(R, A = a\big) = f\big(R, A = b\big).$$

If D consists of a single rating factor, checking this condition corresponds to calculating average premium on a two-way basis by the rating factor and protected class. The condition is satisfied if, at each level of D, the average premium is equal for all protected groups.

When D consists of a larger number of rating factors, the two-way approach becomes impractical. Another limitation of the two-way approach is that even if it were repeated once for each rating factor in D, it would not be able to consider correlations among the variables. One approach to testing this condition would be to develop a function $g(D, A)$ such that

$$g\big(D, A\big) = E\Big[f\big(R\big)\Big|D, A\Big].$$

For example, the function $g(D)$ could be the output of a model that uses the premium $f(R)$ as the response, and D along with the protected class A as the set of predictors. Essentially, g is a surrogate model for f that uses a more limited set of predictors. The test then reduces to checking that

$$g\big(D, A = a\big) = g\big(D, A = b\big).$$

This condition is satisfied if A does not make a material impact on the output of g. The approach to measuring the impact of A on $g(D, A = a)$ will vary depending on the method used to determine g. For example, if g is the output of a generalized linear model (GLM), then the condition would be satisfied if the coefficient for A is negligible and/or has low statistical significance. If g is the output of a machine learning model, then this condition would be satisfied if A has a flat partial dependence plot.

The advantage of using a surrogate model is that it allows for multiple "acceptable" factors to be tested simultaneously. The disadvantage is that use of a surrogate model is less transparent than use of a two-way average premium table, and the ability to explain the test is an important consideration when testing models for bias. The process used to build the surrogate model could itself be influenced by the modeler's bias, but this could be mitigated by specifying strict criteria for how the model is to be built.

An alternate approach, when the number of factors in D is large, would be to check each individually, on a two-way basis, with A. This approach may provide a reasonable approximation of the surrogate model approach when the factors in D are not strongly correlated and produces results that are easier to explain.

4.2. Proxy ("Control Variable") Test

A limitation of tests such as premium parity, loss ratio parity, and conditional demographic parity is that while they can detect situations in which a bias may be present, they do not provide any insight as to which predictors in the model might be contributing to the bias. While correlation analyses could be used to provide insights, they do not consider multicollinearity that may exist among predictors. Accordingly, one approach to attempt to identify predictors that might be contributing to the bias is the proxy Test, informally known as a *control variable test*, which employs concepts that should be familiar to most practitioners of actuarial predictive models. In this test, the protected attribute is added into the baseline model as an explanatory variable, and the resulting model output is compared to the original model output (excluding the protected attribute) and examined for differences.

For example, if using a GLM with a baseline model that includes β coefficients for predictor variables X and response variable Y,

$$E[Y] = \sum_i \beta_i X_i,$$

then the model for the proxy test for $X_{Protected\ Class}$ would be as follows:

$$E[Y] = \sum_i \beta_i X_i + \beta_{Protected\ Class} X_{Protected\ Class}.$$

The intended rationalization of this test is that any association between the protected attribute and the response variable will be captured in the protected class control variable as opposed to other explanatory variables. This, in turn, would result in the effects captured by the remaining explanatory variables being agnostic with regard to the effect of the

Table 4.1. Comparison between Baseline Model and Control Variable Model – Example 1

	Baseline Model		Model Including Protected Attribute		Difference	
Explanatory Variable	Parameter Estimate	*p*-value	Parameter Estimate	*p*-value	Parameter Estimate	*p*-value
Variable 1	0.04	< 0.001	0.04	< 0.001	0.00	< 0.001
Variable 2	0.80	< 0.001	0.82	< 0.001	–0.02	< 0.001
Variable 3	1.40	< 0.001	0.02	0.200	1.38	–0.200
Variable 4	0.50	< 0.001	0.45	< 0.001	0.05	< 0.001
Protected Group A	–	–	0.90	< 0.001	–0.90	< 0.001
Protected Group B	–	–	0.00	–	0.00	–

protected attribute.[3] Thus, after adding the protected attribute as a control variable, any material differences measured in model parameters would indicate that some or all of the predictive effect of the protected attribute is being captured by other variables in the model. If there are parameter estimates for the model that are no longer statistically significant or approximately zero after the protected attribute is included in the model, this would suggest that one or more variables serve as a proxy for a protected attribute. This is illustrated in the model output in Table 4.1.

Table 4.1 is presented for theoretical purposes to help conceptualize the theoretical intent of the proxy test. In practice, however, it is not possible to control which variables pick up the effect when multicollinearity exists, without employing manual intervention in the modeling process. (Such an intervention could be an iterative modeling approach in which one of the correlated variables is added to a first iteration of the model, and the other correlated variable is added in a subsequent iteration of the model that is intended to model the residual effects not picked up by the first model.) A more likely outcome that may occur when two highly correlated variables are introduced into a model is illustrated by the example in Table 4.2, where the parameter estimates for the two variables are significantly large with offsetting effects.

While the example in Table 4.2 would similarly demonstrate that a variable included in the model has a proxy effect for the protected attribute, it highlights an inherent limitation of this test and one that should be familiar to modeling practitioners: the introduction of variables with high correlation (or, more generally, multicollinearity) in any model can lead to volatile results. As stated in Goldburd et al. (2019), "The GLM – forced not to double-count – will need to apportion the response effect between the two variables, and how precisely best to do so becomes a source of great uncertainty." While it is likely that the modeler will

[3] This type of approach is already routinely used in modeling for insurance purposes to account for the effects of other variables, such as the effect of trend and development via a time-dependent variable (e.g., accident year).

Table 4.2. Comparison between Baseline Model and Control Variable Model – Example 2

	Baseline Model		Model Including Protected Attribute		Difference	
Explanatory Variable	Parameter Estimate	*p*-value	Parameter Estimate	*p*-value	Parameter Estimate	*p*-value
Variable 1	0.04	< 0.001	0.04	< 0.001	0.00	< 0.001
Variable 2	0.80	< 0.001	0.82	< 0.001	–0.02	< 0.001
Variable 3	1.40	< 0.001	–14.00	< 0.001	15.40	< 0.001
Variable 4	0.50	< 0.001	0.45	< 0.001	0.05	< 0.001
Protected Group A	–	–	15.40	< 0.001	–15.40	< 0.001
Protected Group B	–	–	0.00	–	0.00	–

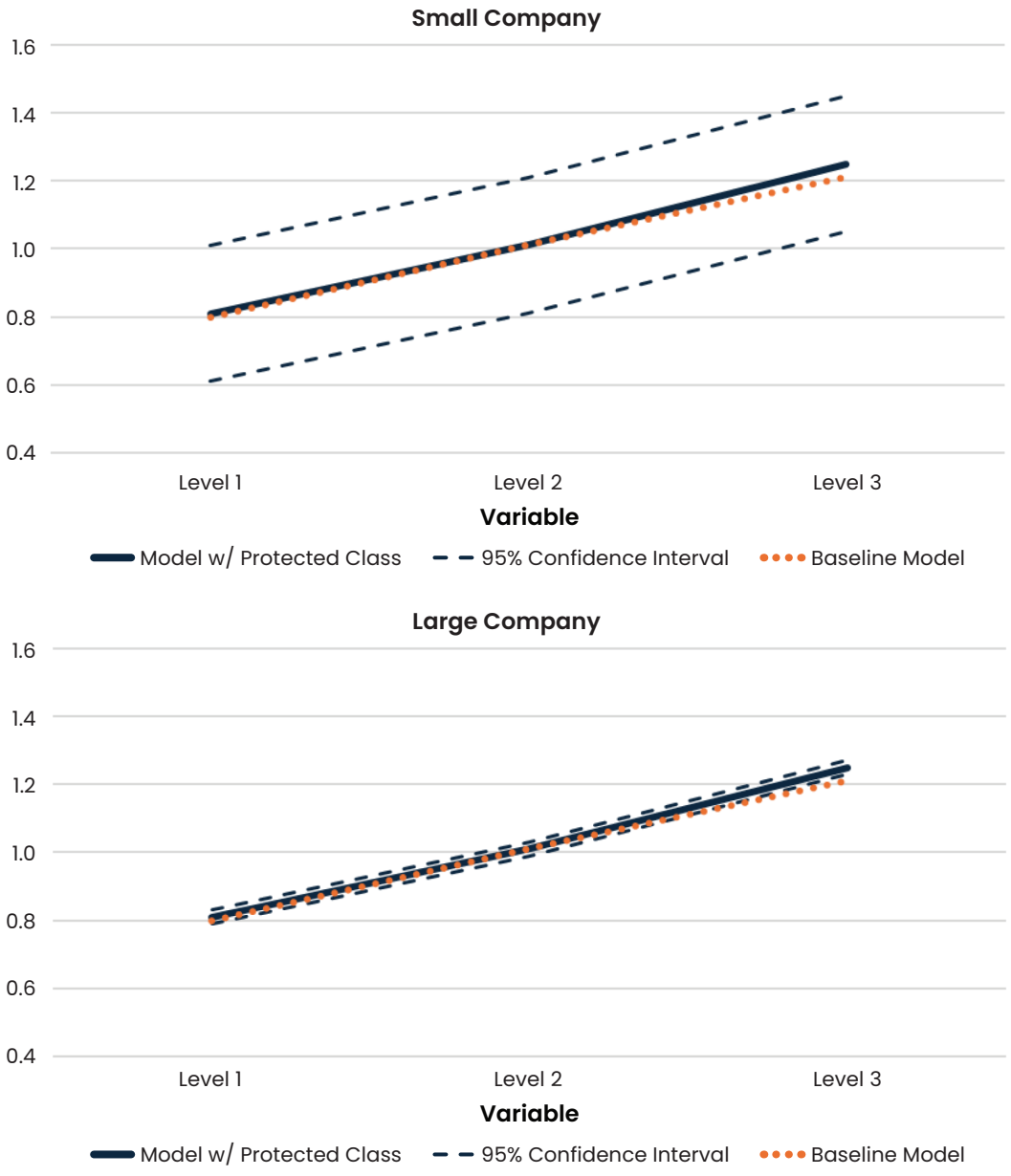
see parameter estimates change if two highly correlated variables are added to the model, it is not guaranteed, particularly when some levels of the protected attribute have low data volume.

When the protected classes are highly skewed, additional limitations on the proxy test can exist even when the protected attribute exhibits lower degrees of correlation with other variables. Essentially, when the protected attribute is omitted from the model, the model parameter estimate for the suspected proxy variable will reflect somewhat of a weighted average effect of the individual effects of the variable by each level of the protected attribute. So, when a protected class that has levels with low data volume is added as a control variable, the parameter estimate for the suspected proxy variable may not materially change because the majority class has the most influence on the overall effect of the suspected proxy variable. For this reason, while the proxy test can be useful in identifying proxy effects or identifying potential causes of premium disparity, users should not infer that premium parity is inherently achieved when a model passes the proxy test. In fact, this situation highlights a common misconception and misunderstanding of what it means for a model to "sort out the correlation" between two variables. The model will not magically undo correlations that exist between two variables such that the inclusion of one as a control would lead to model predictions that are uncorrelated with the control variable. Rather, what is meant by a model's ability to "sort out the correlation" is that the predictive effects of each variable will not be double-counted.

A challenge with the proxy test is defining what constitutes a "material difference" in model parameter estimates after the protected attribute is added as a control. While statistical significance or confidence intervals could be used to define this difference, the modeler should be cautious and not overly rely on these statistical measures without considering their limitations. This is because these statistical tests evaluate with how much confidence an estimate can be measured, which inherently will vary for companies with access to different volumes of data; this creates inequities in the standards applied to different companies.

For example, companies that have a significant volume of data can reliably measure parameter estimates with a high degree of confidence. Changes in these parameter estimates, therefore, may likewise be measured as statistically significant when there is no material difference in the model's prediction from a practical (e.g., premium dollars) perspective. Smaller companies that measure the same difference in parameter estimates may find that the difference is not statistically significant, since a smaller volume of data was used, which likely corresponds to higher standard errors. This is demonstrated visually in Figure 4.1, which shows how a small company and a large company may reach different conclusions if they rely solely on statistical

Figure 4.1. Comparison between Baseline Model and Control Variable Model for Small Company and Large Company



tests, even if the underlying model parameter estimates are the same with and without the control variable in the model.

Further, even statistical significance tests require the selection of an arbitrary threshold (e.g., 5% *p*-value), which may be subjective and could encourage p-hacking.[4] The modeler should therefore strike a balance between statistical measures (statistical significance) and materiality (practical significance) when evaluating results from this test (or any test for bias). For this reason, it is often useful and important to consider the results of other tests in conjunction with the results of the proxy test as opposed to relying exclusively on the conclusions of the proxy test alone.

4.2.1. Nonlinear Extensions

The discussion above has assumed that the underlying model produces explicit parameter estimates (and, based on the significance testing used, standard errors of the parameter estimates), such as in a GLM. However, the concept of the proxy test – to control for the effects of the protected class – can be extended to nonlinear models, such as random forests or gradient-boosted machines, as well. One could do this in a variety of ways. For example, one could review how measures such as variable importance change when the protected attribute is included in the model, as variable importance should change if proxies exist. Other measures, such as changes in partial dependence plots, could also be appropriate. Additionally, methods that may traditionally be used for mitigation via preprocessing could also be analogously leveraged and applied. For instance, one could measure the average effect of the protected attribute, adjust the response variable for this effect, rerun the specified model using the adjusted response variable, and measure how much the model predictions changed. Ultimately, the modeler should select the methodology or measure that is the most appropriate given the context of the model.

4.2.2. Advantages and Disadvantages

The proxy test is easy to understand and quick to implement; once a modeler has protected class data, they are limited only by the time that it takes to run the new models (though, as noted, implementing the test for nonlinear models may require additional steps). Additionally, in modeling techniques that are sensitive to high degrees of multicollinearity, the method can be used to detect explanatory variables that may serve as proxies for protected attributes.

However, as discussed above, there are also disadvantages and limitations to the method. It can be difficult to control where the signal goes when high degrees of multicollinearity exist, and highly skewed protected class distributions could potentially lead modelers to misinterpret the results. Finally, tests relying on statistical significance may result in inequities across companies of different sizes.

[4] P-hacking refers to when data or statistical analyses are selected until desired outcomes are supported by a statistical test (Head et al. 2015).

4.3. Nonparametric Matching

If the modeler needs to evaluate whether a specific variable or combination of variables causes model bias, and multicollinearity exists among the model predictors, a nonparametric matching approach can be used to give the modeler more control. While the nonparametric matching process itself is not a testing procedure, it is a tool that can be used to normalize for distributional differences that may exist across two cohorts within a dataset. This normalization of distributional differences is performed for all variables except for the specific variable(s) to be tested. Using the resulting matched (i.e., conditioned) dataset, bias testing can be conducted. In essence, the theoretical goal of the nonparametric matching process is similar to the goal of conditional demographic parity, wherein the modeler normalizes for a set of "acceptable" risk factors.

Note that the use of matching to infer causal effects has historical roots in experimental observational studies, where the goal is to evaluate the estimated treatment effect, such as the effect of a particular drug on a sample of the population. Consequently, most academic papers on the subject use "treatment" and "control" terminology. This paper uses the same terminology for consistency purposes but will use the term "treatment" to mean a protected group of interest and "control" to mean the remaining levels of the protected class.

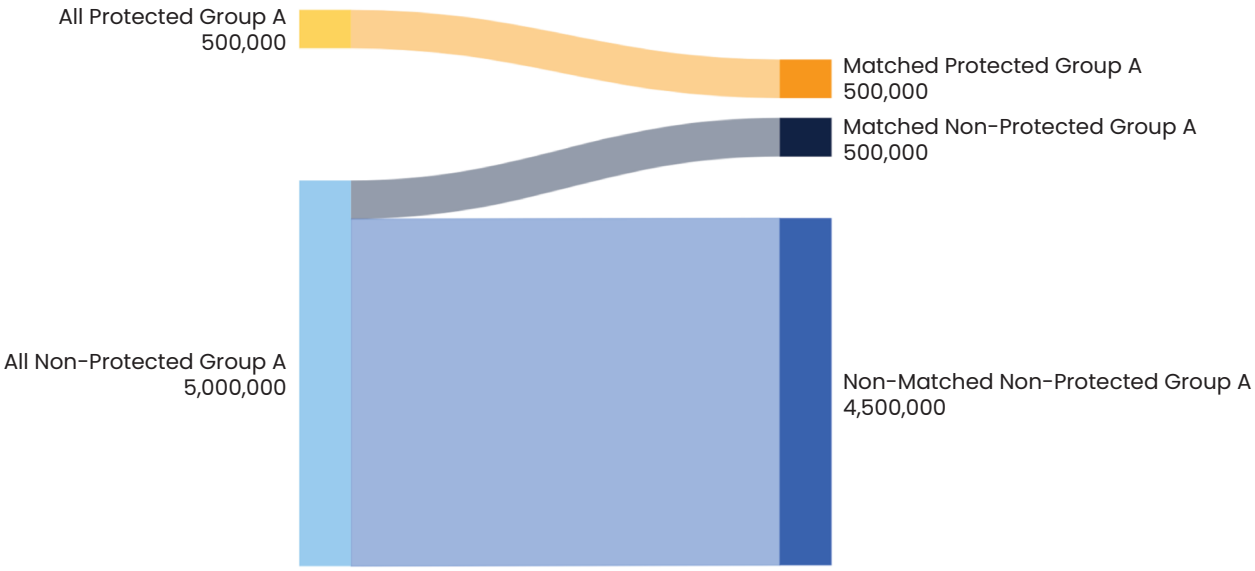
4.3.1. Rationale

Effectively, for every record in the treatment group, the goal of the matching procedure is to identify a record in the control group that is similar with regard to all risk characteristics except for the variable(s) to be tested, such that after the matching procedure is completed, the resulting dataset of matched records (a subset of the entire dataset) is balanced. That is, the matching procedure creates a dataset wherein the distribution of each predictor (except for the variable[s] of interest) is the same or similar for the treatment and control groups. For insurance purposes, this is useful because distributional differences between protected groups for variables that are well accepted by society as being fair for differentiating risk, such as the amount of coverage, can be normalized so that the effects of other variables can be isolated.

For example, assume there are two cohorts of a protected class, groups A and B. Then, for an individual from Protected Group A who lives in a certain geographic area, drives a certain type of vehicle, has purchased certain limits of coverage, etc., the goal of the matching procedure is to identify an individual from Protected Group B who lives in the same geographic area, drives the same type of vehicle, purchases the same limits of coverage, etc. This process would be repeated for every record in Protected Group A. The resulting matched dataset allows for more controlled comparisons between two protected groups such that the effects of the introduction of one or more variables can be isolated. An illustrative example of the subset of data that results from the matching process is displayed in Figure 4.2, where the numbers presented represent record counts in the data.

Once the matching procedure is completed, a variety of tests can be conducted to evaluate differences between the treatment and control groups, depending on the metric

Figure 4.2. Illustration of Matched Subset of Data



of interest. For example, models with and without the variable(s) of interest are fit on the whole dataset (not the matched data), and the predictions are calculated for all treatment and control observations in the matched dataset. From this, the fairness metrics discussed in Part 1, Section 3, such as premium parity and loss ratio parity, can be calculated for the matched dataset.

4.3.2. Defining the Treatment and Control Groups

The first step of the matching process is to define the two cohorts of protected classes that are to be evaluated: a treatment group and a control group. In most matching procedures, the treatment group is programmed to be the group to which the other records (the control group) are matched. The treatment group is therefore usually defined as a particular protected group of interest, and the control group is defined as all other groups within that protected class. Note that a limitation of the matching procedures available is that only two cohorts can be evaluated at the same time.

Care should be taken when selecting these two cohorts, and consideration should be given to the questions the test is trying to answer. For example, if the modeler is testing whether protected minority groups are adversely impacted by a model, they might segment the data into minority and nonminority cohorts. Alternatively, if the effect on a particular protected group is of interest, the modeler might segment the data into the protected group of interest and all other groups within the protected class, or the protected group of interest and the majority group. It is important to note that if the modeler is not careful, the selection of these cohorts themselves could bias the conclusions. For example, if multiple protected minority groups are adversely impacted in a similar manner by a particular variable, including some of those protected groups in the control group will mitigate any measured biases. Thus, in an ideal scenario, the modeler would evaluate multiple treatment and control group definitions.

Once the cohorts are defined, a binary treatment flag variable can be created for every individual in the dataset:

$$Treatment_{Flag} = 1, for\ individuals\ in\ the\ treatment\ group$$

$$Treatment_{Flag} = 0, for\ individuals\ in\ the\ control\ group$$

4.3.3. Matching Procedures

Many algorithms can be used for matching. A few sample procedures that may be relevant for insurance bias testing purposes are discussed below; these descriptions all draw on Griefer (2023). Many of these matching procedures can be carried out using standard statistical modeling software, such as R's *MatchIt* package. In practice, multiple matching procedures should be tested, and the resulting balance should be compared. Because the goal of the matching procedure is to achieve a balanced dataset (preferably without needing to remove any of the treatment records along the way), the method that produces the best balance should be used.

Note that regardless of the matching method used, it is critical that the predictor variable(s) of interest (for bias testing) be excluded from the matching process. That is, if one is interested in evaluating whether the addition of Variable 1 results in bias, Variable 1 must be excluded from the matching procedure; otherwise, the results of any tests on the matched data will always show that no bias exists in a balanced dataset.

Exact Matching

In Exact Matching, each record in the treatment group is matched with a record from the control group that has the same risk characteristics. Whereas other matching procedures match records that are approximately the same, exact matching requires the characteristics to be equal. In practice, exact matching may be impractical for many insurance applications due to the number and types (e.g., continuous) of predictor variables used in many models.

Nearest Neighbor Matching ("Greedy Matching")

In Nearest Neighbor Matching, the closest available control record is matched for each treatment record. Unlike Exact Matching, this method does not require the two records to be exactly equal. This approach requires a selected distance measure in order to evaluate the "closeness" of records in the dataset. For this purpose, modelers will typically use a propensity score – the predicted probability that an individual is a member of the treatment group, conditional on the predictors that should be balanced (excluding the variable[s] to be tested). This score can be calculated using a logistic regression model where the response variable is the binary treatment flag defined above and the explanatory variables are the set of predictors that one would like to balance.

Nearest Neighbor Matching is also known as "Greedy Matching," because it is "greedy" in the sense that the matching for each treatment record occurs without reference to the

matching of other records. This means that there is no effort to optimize any set of criteria relating to the overall quality of matches in the dataset.

Optimal Matching

Optimal Matching is similar to Nearest Neighbor Matching in that it matches records based on a distance measure, such as a propensity score. It is different, however, in that Optimal Matching considers alternative matches based on a designated optimization criterion, such as minimizing the sum of the absolute pair distances.

4.3.4. Other Matching Options

Statistical software packages offer a variety of matching options that the user can employ. A sampling of some of the more relevant options is discussed below; again, these descriptions draw on Griefer (2023). As with any statistical model, it is imperative that the user understand the options that are available and select those that are most appropriate.

Caliper

A caliper defines the maximum distance between two records for them to be eligible for matching. If two records have a distance that is greater than the caliper, the matching algorithm will not consider matching the two records. Often this distance is expressed as a multiple of the standard deviation of the propensity score for all records, though explicit distance measures can also be selected. If a treatment record has no corresponding control records within the caliper distance, it is discarded.

Calipers allow the user to add restrictions to ensure that matches are reasonably similar, which adds more control to the matching process. Small calipers more closely resemble exact matching, so it is preferable to start matching using a small caliper and adjust as needed. If treatment records are discarded during the matching procedure, a larger caliper should be tested to see if fewer treatment records can be discarded without materially sacrificing balance. If the resulting balance is poor, a smaller caliper should be tested.

Replacement

Matching with replacement allows for single control records to be matched to multiple treatment records. For the purpose of testing models for bias, this approach is generally not advisable, as certain policyholders would have disproportionate impact on the results.

4.3.5. Assessing the Balance of the Matched Data Set

After the matching algorithm is completed, the balance of the resulting matched dataset – or, in other words, the quality of the matches – must be assessed. This can be done using validation techniques that should already be familiar to most actuaries. The following description outlines commonly used techniques but is not intended to be an exhaustive list of all the validation techniques that exist. Except where otherwise noted, this section draws on Ho et al. (2011).

Comparison of Means and Variance

To gain an initial understanding of whether balance exists at a high level, the modeler can review a comparison of the mean and variance between the treatment and control groups for each numeric covariate. When comparing means, it is often helpful to review means that have been standardized using the standard deviation of the treatment group or the standard deviation of the treatment and control groups. This allows for multiple covariates to be reviewed simultaneously, which provides a better understanding of the overall balance. The reviewer can compare how the differences in means of the treatment and control groups compare when they are calculated using the entire dataset versus when they are calculated using only the matched dataset. A properly balanced dataset should produce differences in standardized means of the treatment and control groups that are close to zero. An example of this comparison is provided in Figure 4.3, which shows that the differences in the standardized means between the treatment and control groups are closer to zero after the matching procedure, which indicates that the standardized means of the treatment and control groups are similar in the matched dataset.

Comparison of Distributions

In addition to reviewing high-level metrics such as differences in means and variances, it is important to also assess balance at a more granular level by comparing the distributions of the treatment and control groups for each covariate. To do this, modelers can use many statistical measures and visualizations, such as Kolmogorov-Smirnov statistics, box-and-whisker plots, density plots, bar charts, and QQ plots. When reviewing distributional differences, it is helpful to compare the selected measure(s) on the entire dataset as well as the matched dataset to assess how well the matching process reduces differences between the treatment and control groups. Examples of different visualizations are shown in the figures and table below.

Figure 4.3. Comparison of Standardized Mean Differences between Treatment and Control Groups

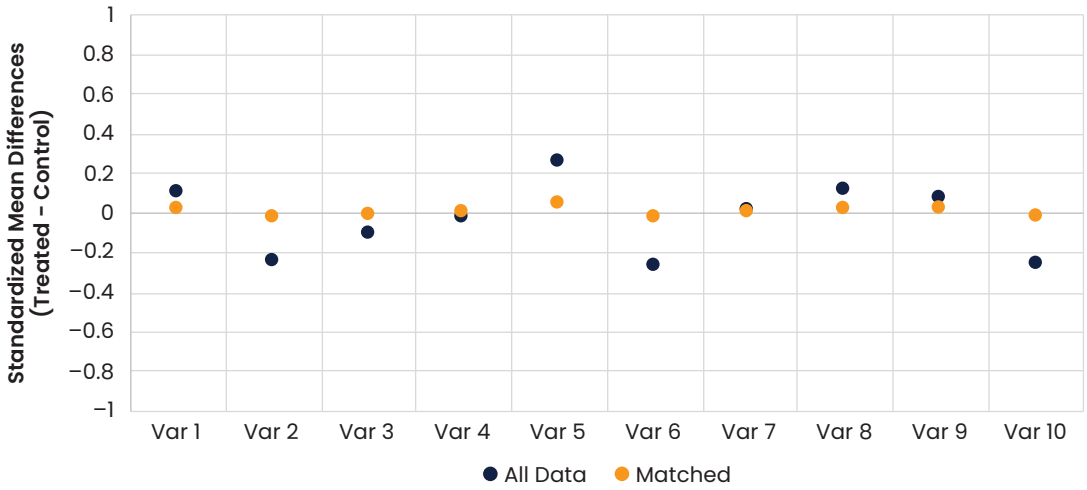


Figure 4.4 shows how the empirical distributions of a sample covariate (with levels 1, 2, and 3) compare for Protected Group A and Protected Group B based on the entire dataset, pre-matching ("All") and the post-matching ("Matched") dataset. The figure demonstrates that the distributions for this sample variable are different when all the data is considered but are more similar after the matching procedure is performed.

Figure 4.5 shows a comparison of QQ plots for a sample covariate pre- and post-matching, which demonstrates that the matching procedure better aligns the treatment and control groups for the sample covariate.

Note that while the similarity of propensity scores between the treatment and control groups can be reviewed, this should not be relied on fully when assessing balance if propensity scores are used in the matching process. Doing so may mislead the modeler with regard to the overall balance of the matched dataset since the matching procedure inherently attempts to minimize the difference in propensity scores.

It is important to note that hypothesis tests are not appropriate for assessing balance. As stated in Ho et al. (2007), "Balance is a characteristic of the observed sample, not some hypothetical population." Additionally, as hypothesis testing does not inherently establish

Figure 4.4. Comparison of Empirical Distributions of Variable Pre- and Post-Matching

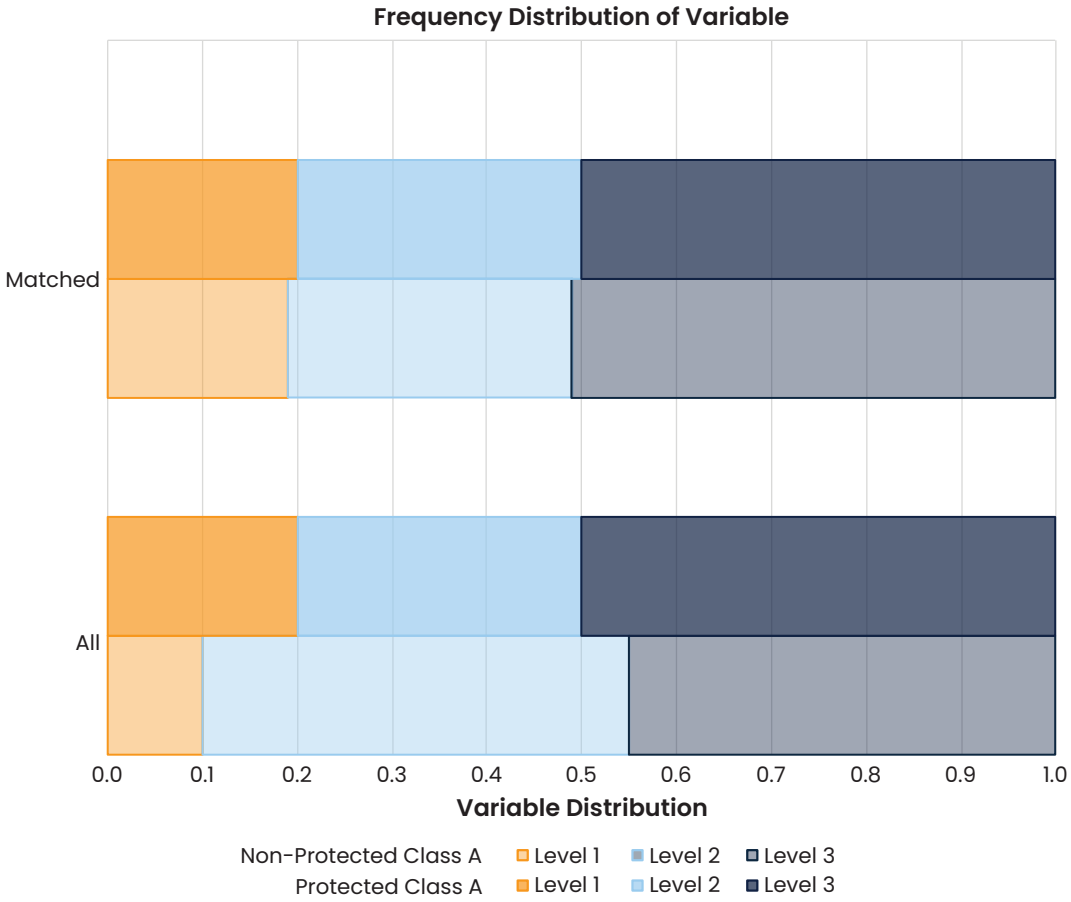
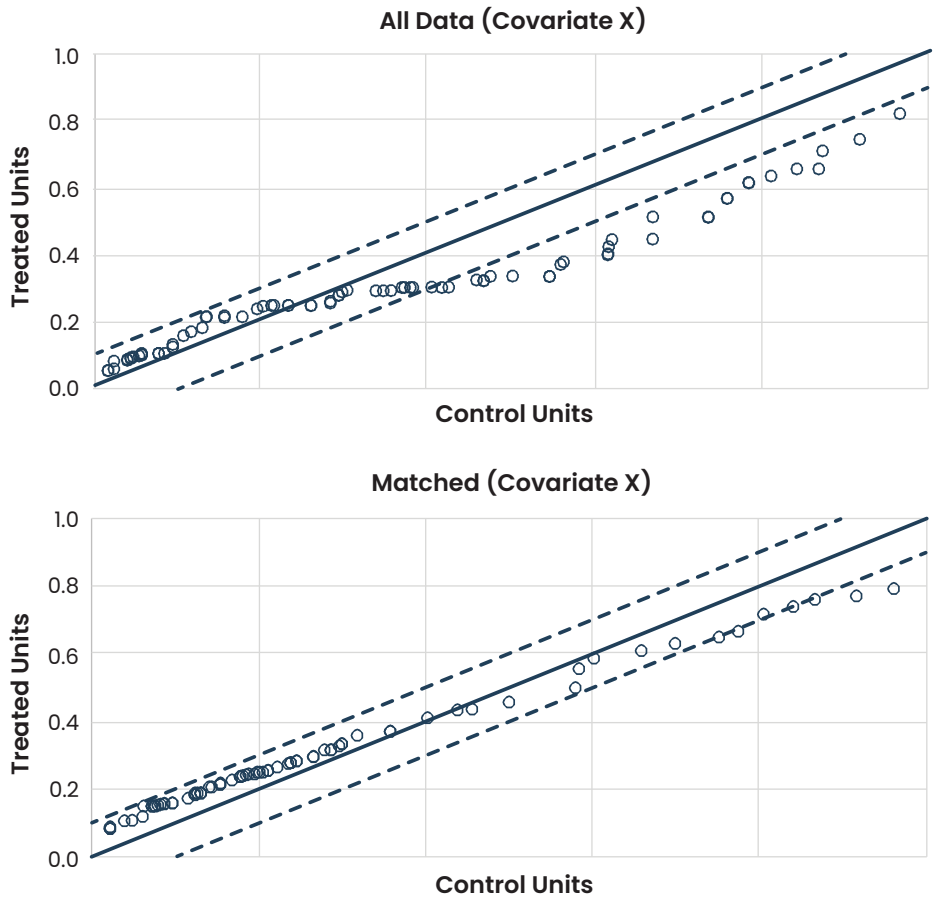


Figure 4.5. Comparison of QQ Plots Pre- and Post-Matching



a threshold for the level of imbalance that is acceptable, hypothesis testing could give the modeler a false sense of confidence in the level of balance in the matched dataset, which could make downstream bias test results misleading. This is further illustrated via an example in Ho et al. (2007) in which test statistics indicate that balance is "improved" as data is randomly discarded.

Ultimately, the modeler should strive to achieve as much balance as possible – a result that should be arrived at by examining multiple measures of balance and using sound judgment. When the measures discussed above indicate that balance has not been achieved, the practitioner should consider and test whether alternative specifications of the matching procedure would improve balance. Discarding records in the treatment (i.e., minority) group should be avoided, as doing so could lead to misleading bias-testing results. If a more balanced dataset cannot be achieved, tests on the matched dataset may still be insightful; however, the modeler should note the limitation on results, in that full conditioning has not been achieved.

4.3.6. Conducting Tests on the Matched Dataset

Once a sufficiently balanced matched dataset has been achieved, tests can be conducted to assess the isolated impact that a certain variable or set of variables has on the treatment

Table 4.3. Fairness Testing Using Matched Dataset

	Matched Data Predicted Pure Premium[2]	Average Pure Premium[1]		
		Protected Class A Policyholders	Non-Protected Class A Policyholders	Difference
(A)	Model without Variable of Interest	100	101	(1)
(B)	Model with Variable of Interest	140	100	40
(C)	**Actual Pure Premium**	**125**	**89**	**36**
			Difference of Predicted Differences [(A) – (B)]	**(41)**
			Relative Difference [(A) – (B)] / (B)	**–103%**

[1] Average pure premium represents the average for Protected Class A Policyholders (treatment group) and Non-Protected Class A Policyholders (control group) within the matched dataset.
[2] Predicted pure premium is derived from GLMs built using all training data, with the variable of interest included and excluded, respectively, as an explanatory variable.

and control groups. Recall from above that the matching process itself is only a method of dividing the data to be tested into two groups (treatment and control) that mirror each other with regard to all variables except for the variable(s) to be tested. The intent of this process is to condition the data on a set of "acceptable" variables. The matching process allows for comparison of how model predictions (premium parity) or predicted loss ratios (loss ratio parity) change by treatment and control group when the variable(s) of interest are added to the model.

To conduct these tests, it is first necessary to obtain model predictions for each record in the matched dataset for

1. the version of the model that excludes the variable(s) of interest and
2. the version of the model that includes the variable(s) of interest (the baseline model).

It is critical to use model predictions based on models that have been fit on the entire training dataset, not the matched dataset.

Once the model predicted values have been obtained for records in the matched dataset, the model predictions and historical loss experience can be summarized by treatment and control group for each model. This is demonstrated in the example in Table 4.3.

As demonstrated in Table 4.3, it is often useful to explicitly calculate a difference in the model predictions between the treatment and control groups. For the model excluding the variable(s) of interest, this difference should be zero or near zero,[5] as the matching procedure should

[5] Tolerance for differences will be dependent on the model and the model output's sensitivity to its inputs.

have been calibrated using the same variables in the model. Nonzero differences are indicative of an imbalanced matched dataset, in which case the modeler should revisit the matching procedure to test whether alternative matching assumptions, such as a smaller caliper, result in a more balanced dataset.

Material differences in the average model predictions for the model that includes the variable(s) of interest indicate that the introduction of those variables into the model disproportionately impacts one of the two classes; this suggests that after controlling for all other risk characteristics, there are distributional differences in the variable(s) of interest between the two classes.

In reviewing loss ratio parity, actual historical loss costs can be summarized by the treatment and control groups, and ratios of the historical loss costs to the model-predicted loss cost can be calculated under the two different model scenarios. Ratios that are close to 1.00 indicate that the model predictions align well with overall historical experience for each group. In the example above, the inclusion of the variable(s) of interest in the model improves model predictions in both the protected and treatment groups, which indicates that loss ratio parity is achieved.

For a more nuanced understanding of how the treatment and control groups compare with regard to the variable(s) of interest, distributions of the variable(s) of interest can also be compared between the two classes. This may provide a better understanding of the results of the premium parity and loss ratio parity tests discussed in Part 1, Section 3.

4.3.7. Advantages and Disadvantages

A main advantage of nonparametric matching is that it allows the modeler to isolate the effects of a particular variable while controlling for the effects of the other variables, all while making no assumptions about the model form. That is, it allows the modeler to more appropriately apply conditioning when the model includes many variables. For insurance purposes, this is particularly advantageous when there are risk characteristics that are widely viewed as acceptable to use for differentiating risks. To the extent that there is correlation between the protected attribute and other variables besides the variable of interest, the methodology effectively normalizes for this correlation, such that the relationship between the variable of interest and the protected attribute can be isolated.

A key limitation of nonparametric matching is that matching algorithms require a binary classification of the protected class. This means that if the available protected class information is inferred, for example using the BIFSG approach covered in Part 1, Section 2, the inferred probabilities cannot be directly incorporated into the analysis. Each individual must be assigned to a specific protected group.

Additionally, as discussed above, because binary classification is required, only two cohorts can be evaluated at a time. The above-discussed complications that arise from binary classification are exacerbated when intersectionality of multiple protected attributes is of interest.

Finally, matching may be difficult in situations in which a limited volume of data is available and the model or algorithm has high dimensionality (many possible combinations of risk classifications). In these cases, it might not be possible to obtain a balanced dataset without eliminating many of the explanatory variables, as the dataset may not reasonably have enough individuals with similar risk characteristics to complete the matching process.

Section 5. Approaches to Mitigating Bias in Models

Mitigating bias in models requires a comprehensive approach that leverages a variety of techniques at various stages of model development and implementation. These techniques can be broken down into three types of intervention methods: preprocessing, in-processing, and postprocessing methods.

5.1. Preprocessing Methods

Preprocessing techniques focus on mitigating potential bias before model training begins. Three examples of such methods are discussed below.

5.1.1. Removing Linear Dependence

Removing Linear Dependence is a method that seeks to remove one-way correlations between predictors and protected attributes (Berk 2008). Mathematically, this can be represented as

$$x_{residual} = x - \beta Y,$$

where X represents the predictor variables, Y represents the protected attribute, and β represents the coefficients obtained from regressing X on Y.

While this method aims to eliminate direct associations between predictors and protected classes, it overlooks complex interaction effects. For instance, if age and car type correlate with race, interactions between these variables may persist even after regression. Successfully implementing this method necessitates identifying and addressing all high-order interactions, which can be challenging.

5.1.2. Equalizing Outcomes

Modelers can equalize average outcomes across protected groups by adjusting weights, such as exposures (Berk et al. 2021). For example, if one protected group exhibits higher average losses, the weights are reduced for observations within that group until their average losses align with those of other groups. Mathematically, this can be represented as

$$\sum_{i=1}^{n} w_i x_i = \sum_{i=1}^{n} w_i y_i,$$

where w represents the weights of each observation, x represents the outcome of Protected Group 1 and y represents the outcome of Protected Group 2.

However, this approach often introduces a delicate balance between fairness and accuracy. Adjusting average outcomes may result in a trade-off, potentially sacrificing predictive power for the sake of parity in outcomes.

5.1.3. Perturbing Variables

Perturbation involves altering observations or predictor variables to achieve equitable outcomes across protected groups (Pedreschi, Ruggieri, and Turini 2008).

Mathematically, this can be represented as

$$X = X' + \epsilon,$$

where X' represents the perturbed predictor variables, X represents the original predictor variables, and ϵ represents the perturbation vector.

For instance, in this method, some observations from one protected group may be reassigned to another, or variables such as car models may be modified. These changes can be made randomly or strategically to equalize outcomes. While this approach addresses bias, it introduces complexities similar to base rate rebalancing, including potential unintended consequences and challenges in implementation.

5.2. In-Processing Methods

In-processing techniques intervene during the model fitting process to mitigate bias. Two examples of such methods are discussed below.

5.2.1. Including Protected Class as a Control Variable

As discussed in Section 4, integrating protected class as a control variable aims to remove associated signals from the model. By explicitly considering protected class during model training, this approach seeks to minimize direct associations between predictors and protected classes.

Mathematically, this can be represented as

$$b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n + b_{n+1}c + e = y,$$

where b represents the model coefficients, x represents the predictor variables, c is a control variable for the protected attributes, and e is an error term. After the model has been trained with the protected class control variable included, the coefficient for the protected class variable is ignored and the model parameters for all variables excluding the control variable can be implemented.

However, this approach may not fully eliminate all signals, particularly complex interactions between predictors and protected classes. Also, this method focuses on eliminating proxy discrimination but may allow other types of discrimination such as systemic discrimination.

5.2.2. Penalized Fitting Processes

Introducing a fairness regularizer penalizes associations between protected class membership and predictors (Kamishima, Akaho, and Sakuma 2011). This regularization technique aims to discourage the model from relying on protected information to make predictions.

Mathematically, this can be represented as

$$b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n + r + e = y,$$

where b represents the model coefficients, x represents the predictor variables, r is a regularization term for protected attributes, and e is an error term.

Determining an appropriate fairness calculation to optimize can be challenging. One example of a definition of a regularization term would be a term that seeks to minimize the disparity in predictions between the protected groups. However, the use of regularization terms can potentially lead to unequal treatment or unintended consequences.

5.3. Postprocessing Methods

Postprocessing methods focus on adjusting model outputs to achieve fairness. Two potential methods are discussed below.

5.3.1. Fairness Transformations

The fairness transformation approach involves applying transformations to the best-estimate prices to align with fairness axioms (Dwork et al. 2012). For example, rating factors can be modified to achieve loss ratio parity or parity in some other fairness metric. Mathematically, this can be represented as

$$\frac{\sum_{i=1}^{n} Losses_x}{\sum_{i=1}^{n} Premiums_x} = \frac{\sum_{i=1}^{n} Losses_y}{\sum_{i=1}^{n} Premiums_y},$$

where x represents the observations from one protected group and y represents the observations from a different protected group. In this example, premiums would be iteratively adjusted by changing rating factors until the loss ratios were equal among the two protected groups.

While this method achieves fairness at the output level without altering the underlying model, it may introduce price distortions and compromise interpretability. Additionally, justifying fairness adjustments can pose challenges, potentially leading to debates over fairness criteria and trade-offs between different dimensions of fairness.

5.3.2. Discrimination-Free Pricing via Adversarial Debiasing

Adversarial debiasing is a technique designed to mitigate both direct and indirect discrimination in insurance pricing models. This method uses adversarial learning to adjust

the model, ensuring that predictions do not rely on protected characteristics. The approach operates by introducing a secondary adversarial model that actively attempts to predict protected characteristics based on the primary model's output (Lindholm et al. 2022).

The steps of this process can be described as follows:

1. Identify discriminatory and nondiscriminatory variables.
2. Compute the best-estimate price using all available variables.
3. Compute the unawareness price by ignoring discriminatory variables.
4. Adjust for indirect discrimination by marginalizing over the distribution of discriminatory covariates.
5. Implement the discrimination-free price in practice.

The adversarial debiasing approach ensures that the pricing model is trained to be discrimination-free by reducing its dependence on protected characteristics, thereby addressing both direct and indirect discrimination. However, one challenge of this method is balancing fairness and accuracy, as removing all signals related to protected characteristics may lead to a reduction in the overall predictive power of the model.

5.4. Considerations for Bias Mitigation

In conclusion, there are multiple approaches to addressing bias in predictive models for insurance pricing at various steps in the process, including preprocessing, in-processing, and postprocessing techniques. It is important to consider both the potential sources of bias and the type of fairness to be targeted in selecting mitigation approaches. Each method has advantages and challenges, and thorough evaluation and careful consideration of fairness as well as accuracy, transparency, and interpretability are critical. In many cases, it may be appropriate to perform bias detection tests, such as those discussed in Section 4, both before and after applying any mitigation techniques, to gain a clear understanding of the impacts of the mitigation measures.

Section 6. Non-Modeling Considerations for Fairness Testing

While this paper has focused primarily on testing insurance rating models, these models are only a subset of a larger process for which fairness analyses may be necessary. Ultimately, stakeholders will be concerned about whether a company's use or implementation of a pricing model could result in discriminatory effects. The rating model outputs discussed earlier in this paper are just one input into the price that a policyholder will pay.

Before the modeling begins, it may be appropriate to evaluate the modeling data to understand the impact of any potential underlying biases. Targeted marketing practices, agency presence in different locations, and the company's overall risk appetite and desired customer segments can produce a policyholder dataset that is not representative of the full spectrum of potential

policyholders in the market. This can result in little or no data for protected groups of interest or data that represents only a subset of certain groups. There may also be systemic biases (see Part 1, Section 1) impacting what information is used to categorize risks for modeling purposes, how information is collected, or how complete or clean the data is.[6] These biases may result in rating factors that are not consistently reliable across protected groups. For example, a greater proportion of policyholders with no credit history may be members of protected minority groups, resulting in a systematic difference in the quality of the credit-based insurance score rating factor (2021 CAS Race and Insurance Research Task Force 2022). Such biases may or may not be apparent in the model outputs and/or model fairness tests or may limit the effectiveness of selected bias mitigation approaches.

Within the model building and evaluation process, it is important to be aware of human biases that can impact the data that is included or excluded from the model and the interpretation of model results. Statistical biases can even impact the fairness testing process itself. For example, if imputation methods are used to infer protected information, and imputation error rates are not consistent across those protected groups, as illustrated in Part 1, Section 2, that could introduce a bias into the fairness testing outputs.

While rating factors selected for implementation do often align with the outputs of a rating model, selected rates can deviate from the model recommendations for a variety of reasons. The underlying volatility or volume of the data may result in unexpected model outputs, such as reversals in the general pattern of modeled rates from one level of a rating factor to the next. For example, the model may generally indicate an increasing rating factor as age increases, but for one age range the indication does not fit with that trend. In this situation, the company may select a rating factor that follows the general pattern, rather than the indication. Volatility in the data may also result in wider confidence intervals around indications for certain levels of rating factors, such that there is more uncertainty in the output of the model. In this situation, the modeler might make a judgmental selection that acknowledges the pattern of the model outputs but does not match the indication exactly.

Beyond data volatility, there may also be business considerations and constraints that are incorporated into the selected model parameters. Limited insurer resources, such as infrastructure, technical capabilities, budget, and time, may make it impractical or impossible to implement structural changes to a pricing algorithm. The insurer may make selections that moderate negative impacts to certain subsets of its policyholders, cap the increases or decreases that policyholders may experience from a given rate change, or improve the company's ability to attract certain segments of customers away from competitors. Depending on the organizational structure of the company, there may be many different internal stakeholders reviewing the indicated model outputs and weighing in on the final selected rates. There may also be regulatory constraints prohibiting the use of certain rating factors in a model, limiting the overall change in rates that policyholders experience from one renewal to the next, or specifying the maximum and minimum premiums that insurers are allowed to

[6] A discussion of these concerns, with respect to four commonly used insurance rating factors, can be found in the 2022 CAS paper "Understanding Potential Influences of Racial Bias on P&C Insurance: Four Rating Factors Explored."

charge in each jurisdiction. The technical model may not directly incorporate such constraints, and thus what is implemented must deviate from the model to achieve compliance.

Finally, testing can include consideration of discounts and loads applied on top of the rating model selections. As mentioned in Part 1, Section 3, to the extent that expense and profit loads maintain the relative ordering of premiums, these may not introduce additional bias or potential discrimination into an insurer's rates. If the insurer has a complex system for determining expense and profit loads, that relative order could be validated as part of the testing process. Other discounts and loads may introduce additional disparities into the charged premiums. Affiliate marketing discounts are offered to potential customers based on their membership in a certain group (e.g., alumni of a certain university or employees of a certain company), and if societal systemic bias impacts the makeup of those groups' membership, the application of such a discount could introduce that bias into the insurer's charged premiums. Payment plans that offer policyholders a discount if they pay their premium in full at the beginning of the policy term rather than in monthly installments could also introduce a disparity in premiums between protected groups, if different classes have different likelihoods of selecting upfront payment options.

This paper has discussed a variety of tools that can help actuaries and insurance professionals to explore insurance pricing data and models in order to identify fairness concerns or impacts of bias, and mitigate issues that arise. As the insurance industry's definition of fairness continues to evolve, applying multiple testing methodologies to a variety of metrics (e.g., premiums versus loss ratios, model outputs versus final charged premiums, etc.) and experimenting with multiple mitigation techniques may provide valuable insights to the insurer and the industry as a whole.

References

Berk, R. A. 2008. "The Role of Race in Forecasts of Violent Crime." *Race and Social Problems* 1: 231–242.

Berk, R., H. Heidari, S. Jabbari, M. Kearns, and A. Roth. 2021. "Fairness in Criminal Justice Risk Assessments: The State of the Art." *Sociological Methods and Research 50* (1): 3–44.

Chibanda, K. F. 2022. *Defining Discrimination in Insurance.* CAS Research Paper Series on Race and Insurance Pricing. Arlington, VA: CAS. https://www.casact.org/sites/default/files/2022-03/Research-Paper_Defining_Discrimination_In_Insurance.pdf.

Dwork, C., M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. 2012. "Algorithmic Fairness through Awareness." *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.

Goldburd, M., A. Khare, D. Tevet, and D. Guller. 2019. *Generalized Linear Models for Insurance Rating*. CAS Monograph Series 5, 2nd ed. Arlington, VA: Casualty Actuarial Society. https://www.casact.org/sites/default/files/database/monographs_papers_05-goldburd-khare-tevet.pdf.

Griefer, N. 2023. "Matching Methods." Comprehensive R Archive Network. https://cran.r-project.org/web/packages/MatchIt/vignettes/matching-methods.html.

Head, M.L., Holman, L., Lanfear, R., Kahn, A.T., and M.D. Jennions. 2015. *The Extent and Consequences of P-hacking in Science*. PLoS Biology. https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002106

Ho, D., K. Imai, G. King, and E. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15 (3): 199–236. https://doi.org/10.1093/pan/mpl013.

Ho, D., K. Imai, G. King, and E. Stuart. 2011. "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference." *Journal of Statistical Software* 42 (8): 1–28. https://doi.org/10.18637/jss.v042.i08.

Kamishima, T., S. Akaho, and J. Sakuma. 2011. "Fairness-Aware Learning through a Regularization Approach." *Proceedings of the 3rd IEEE International Workshop on Privacy Aspects of Data Mining,* 643–650.

Lindholm, M., R. Richman, A. Tsanakas, and M. V. Wüthrich. 2022. "Discrimination-Free Insurance Pricing." *ASTIN Bulletin* 52 (1): 55–89. https://doi.org/10.1017/asb.2021.23.

Pedreschi, D., S. Ruggieri, and F. Turini. 2008. "Discrimination-Aware Data Mining." Paper presented at KDD 2008, Las Vegas, NV, August 24–27, 2008.

2021 CAS Race and Insurance Research Task Force. 2022. *Understanding Potential Influences of Racial Bias on P&C Insurance: Four Rating Factors Explored*. CAS Research Paper Series on Race and Insurance Pricing. Arlington, VA: CAS. https://www.casact.org/sites/default/files/2022-03/Research-Paper_Understanding_Potential_Influences.pdf.