CAS Ratemaking and
Product Management Seminar
March 10-11, 2009

GLM II: Basic Modeling Strategy

Bob Weishaar

Optimal Decisions Group

A LexisNexis Company

optimaldecisionsgroup
A LexisNexis® Company

ANTITRUST Notice

The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.

Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.

It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

Agenda

- Basic strategy
  - Understand the business problem
  - Know the data
  - Apply regression techniques and avoid pitfalls
- General modeling tips
  - Data cleaning
  - Variable screening and transformation
  - Variable selection
  - Model validation

Basic modeling strategy

- Achieving business objectives requires an understanding of:
  - Customer behavior (e.g., loss propensity, conversion, retention, response)
  - Impact of business decisions on customer behavior (e.g., price sensitivity)
- Customer behavior is explained by multiple predictors, and they are frequently correlated
- We need a multivariate solution

Basic modeling strategy (cont.)

- One-way analyses lead to “double-counting” effects
- Multivariate solutions handle correlations:
  - Loss ratios can be problematic: Current premiums? Intentional subsidies? Reusable?
  - Minimum bias procedures: no statistical framework, but allows non-linear forms
  - Classical regression: statistical framework; need more flexibility
  - GLM: provides framework; flexible (though still linear)
  - Some problems benefit from non-linear solutions

Basic modeling strategy (cont.)

- Need a solid understanding of:
  - Business problem
  - Data
  - Regression techniques / pitfalls
- The following slides highlight some examples
- Expertise in one area is not enough!

Understanding the business problem

- Forecasting
  - Absolute or relative values important?
  - Stable predictors?
- Impact analyses – intentionally changing predictors
- Lifetime value studies
  - Preference for "smooth" predictions for variables that change over time (e.g., driver/vehicle age, tenure)?
- Incorporating constraints and business decisions
  - IT system "stuck" with a 10% blue car discount, or
  - You "want" a 10% blue car discount

optimaldecisionsgroup
A LexisNexis® Company

Knowing the data

- Target definition:
  - Retention: transfers to other companies, reinstatements, company cancels, spinoffs, multiple policies per household
  - Conversion: company rejects, duplicates, when quotes are counted, default values
- Rolling up transactional data:
  - driver/vehicle/policy/household; timeframes; transitions
- Data source differences?
  - Accepted and non-accepted quotes from two different sources?

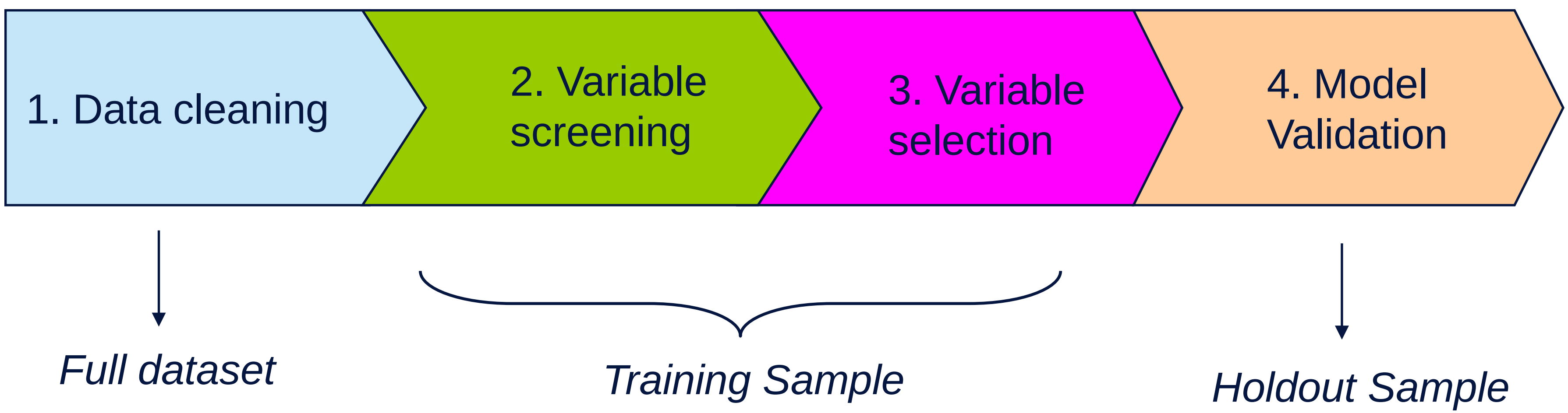
Knowing the data (cont.)

- External events – correlated with predictors?
  - Billing errors
  - Marketing initiatives
- Misinterpreting response due to lack of data
  - Attrition related to transitions – e.g., adding youthful driver
- A posteriori variables

Regression techniques and pitfalls

- Common link functions:
  - Identity for additive models
  - Log for multiplicative models
  - Logit for probabilities
- Common error distributions:
  - Poison for frequency / claim counts
  - Gamma for severity
  - Binomial for response
  - Tweedie for pure premium
- Weights and offsets:
  - e.g., one claim in 6 months versus four claims in two years:
  - Target with frequency 2, using time as weight (amount of info)
  - Target with claim counts 1 and 4 using time as offset (known effects)

General modeling tips



optimaldecisionsgroup
A LexisNexis® Company

General modeling tips

1. Data cleaning → 2. Variable screening → 3. Variable selection → 4. Model Validation

Create list of viable predictors

Create a list of variables and interactions which influence a customer's behavior

- Meet / discuss / interview peers to generate candidates
  - Pricing and underwriting variables
  - Recent or current policy transitions
  - Price change / competitive position
  - Other customer experiences – claims, billing, marketing, agent interaction, etc.
- Generate some standard descriptive statistics to assist, e.g. frequency distribution, density plots
- If the modeling data is coming from multiple sources, compare distributions

Remove nonsensical variables

1. Known to be consistently inaccurate, problematic, or irrelevant
2. Unvarying
3. Replicates of the same data
4. A posteriori variables

Missing Data: exclude or impute

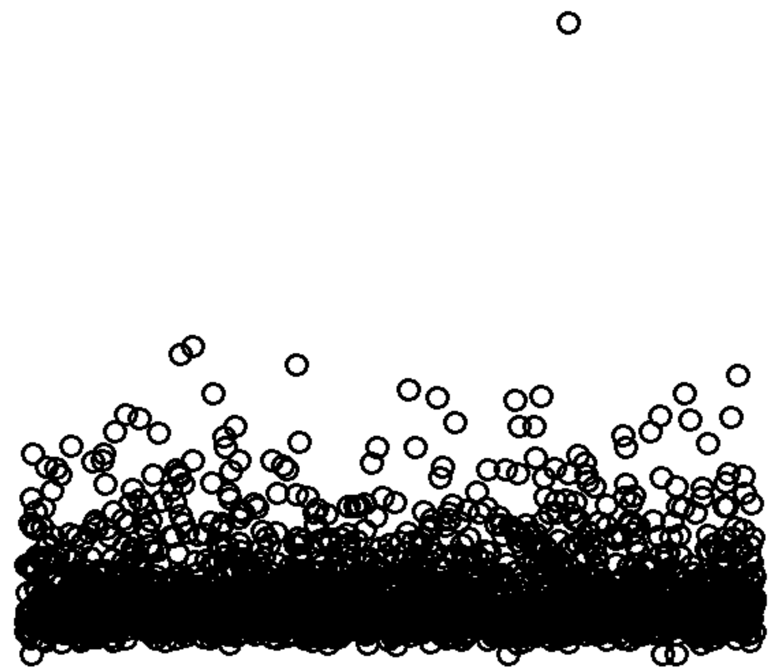
If variables are not populated for some observations:

1. Exclude the variables
2. Create a new level
3. Create a model to estimate the missing values:
    - Use the mean (or mode for a discrete variable), or
    - If particularly important, use a more detailed model:
        - A linear model adds nothing new
        - Consider a non-linear or tree design as appropriate

Find and remove outliers

- Look for outliers in each variable
  - Extreme / abnormal values
  - Small groups for discrete variables
- Solutions:
  - Remove rows
  - Censor or replace with mode
  - Remove entire column
  - Impute
- Also check for groups with abnormal values for the target (e.g., 0 frequency)

Outlier Example



General modeling tips

1. Data cleaning

2. Variable screening

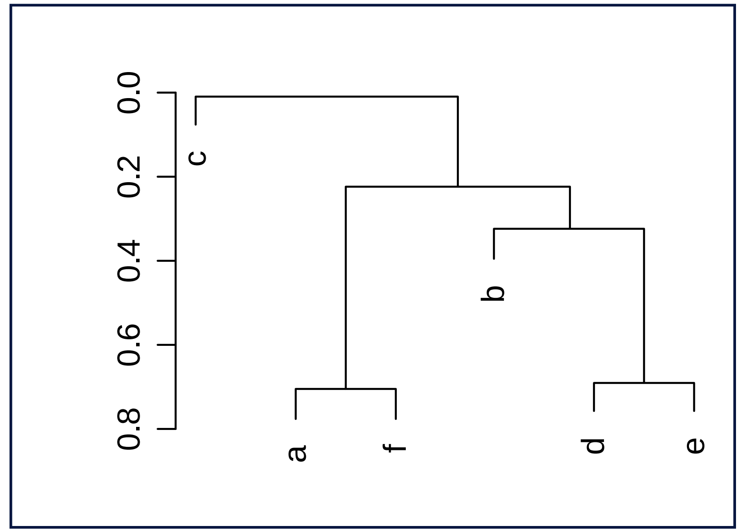
3. Variable selection

4. Model Validation

Check for strong correlation

- Look for significant correlation
  - Use knowledge of the data
  - Use R/SAS Varclus procedure
- Including correlated variables does not hinder the ability of a regression model to predict the response, however:
  - Difficult to interpret the model / use results
  - Problematic if relationships between the variables change over time

Varclus Example*



Consider removal of f & e

Consider removing variables and reducing data

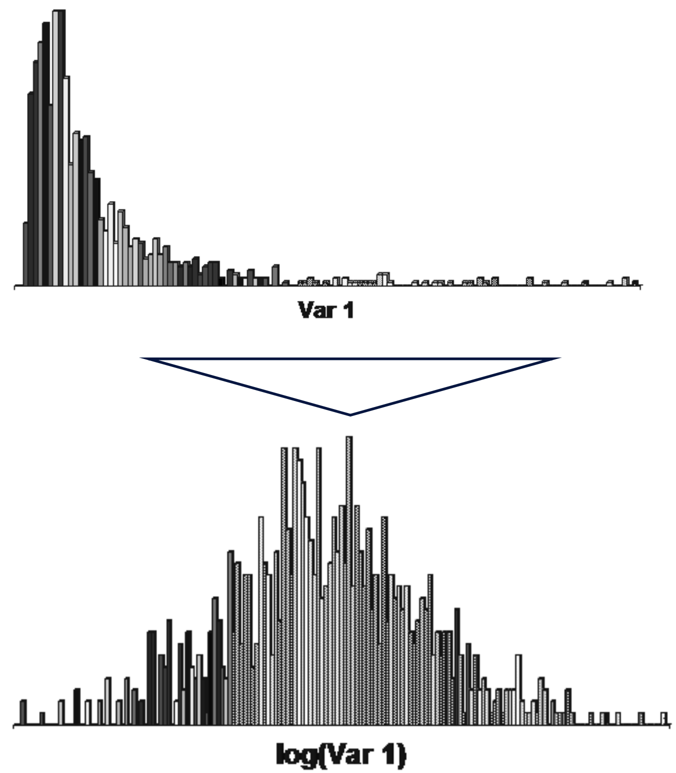
If significant correlation occurs between two or more variables:

- Eliminate variables / levels where their effects are included from other causal variables
- Perform manual orthogonalization
  - Volume ~ age + license years
  - Age and license years are highly correlated
  - Better parameterization: volume ~ age + age licensed
- Replace a cluster of variables with one score; e.g., a "Car Quality Score"

Perform initial transformations

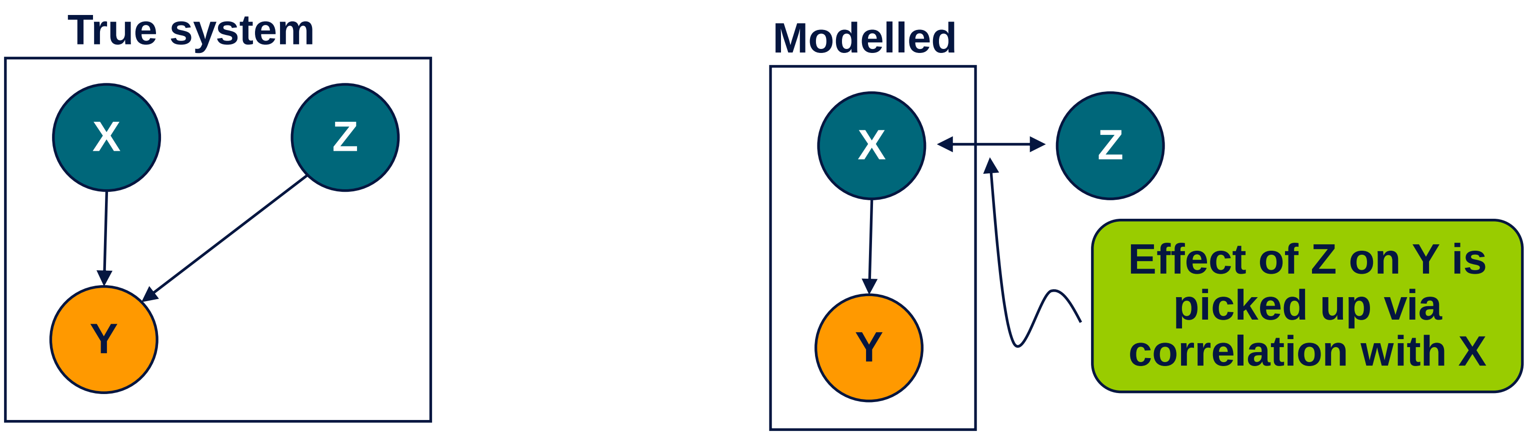
- Define each variable as categorical or continuous
  - Codes are categorical (e.g., territory)
  - Continuous with few values → categorical?
- For skewed continuous variables…
  - A logit transform for proportions?
  - A log transform for positive variables?
- Group small levels of categorical variables into conceptually contiguous levels
- After transformation, re-check for outliers

For example …



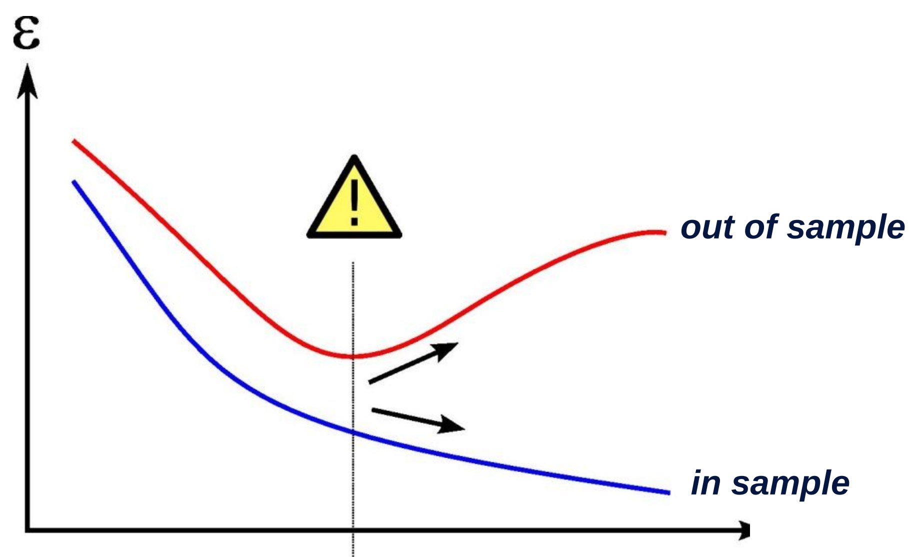
General modeling tips

Include enough degrees of freedom…



… but not too many!

Over-fitting manifests through reduced predictive ability out of sample *



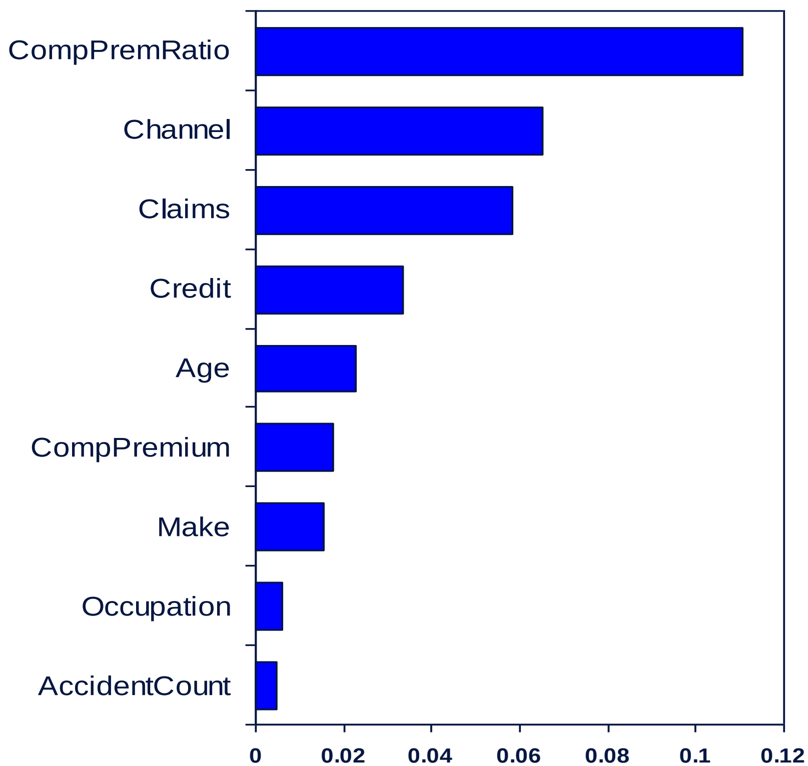
* Diagram Source : en.wikipedia.org/wiki/Overfitting

Causes of over-fitting

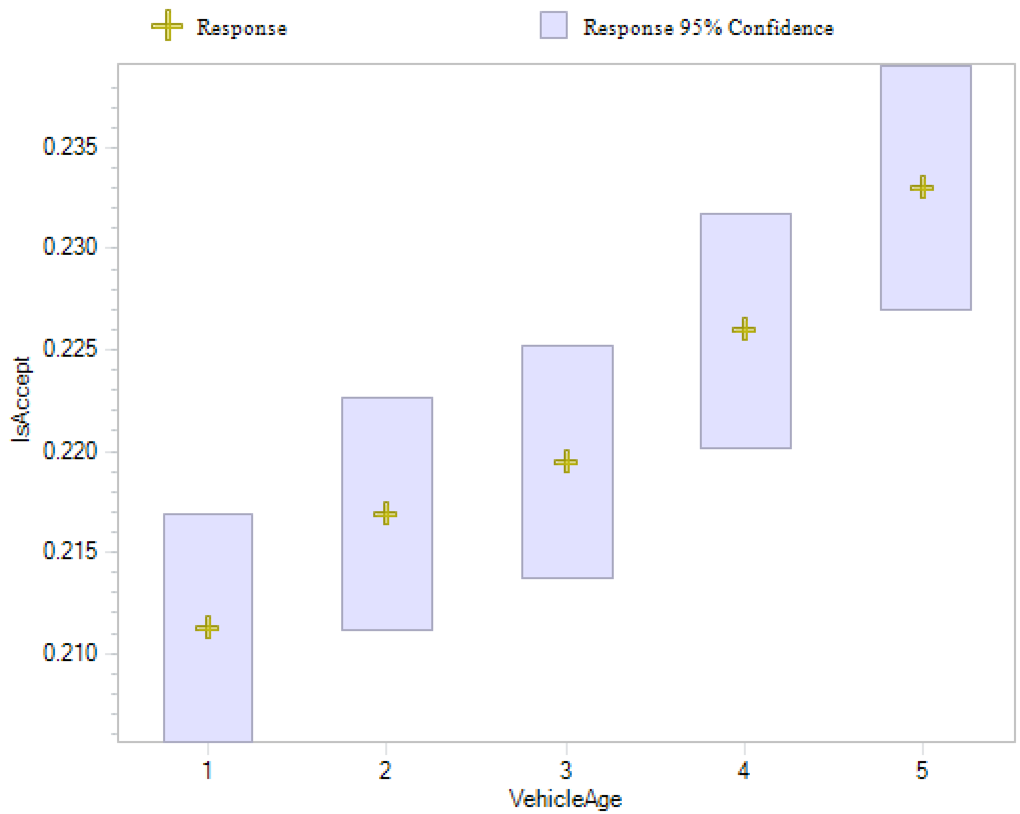
- Inclusion of irrelevant variables
- Use of too many degrees of freedom for some variables
- "Blind" variable selection strategies:
  - Full models which include all available predictors
  - Automated variable selection
- Transformations and groupings based on charts against the response, rather than an understanding of variables

Choose candidate predictor variables

Measure correlations with the response



Draw charts of each variable against the response

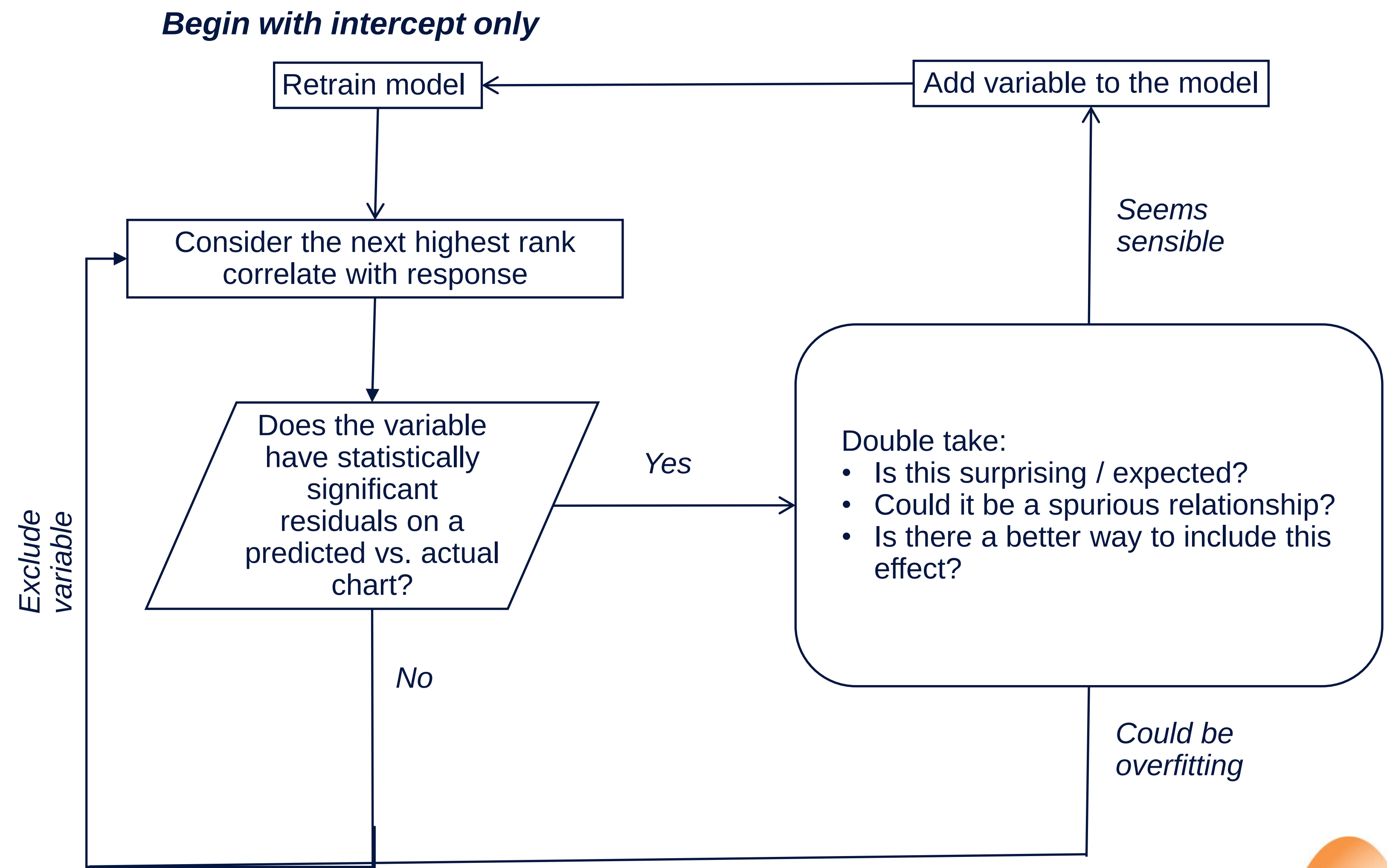


odg optimaldecisionsgroup A LexisNexis® Company

Set up transformations

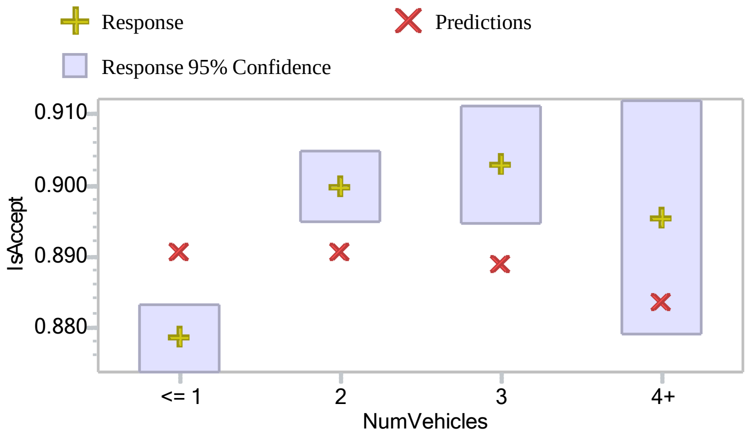
- Allocate more degrees of freedom to variables with strong relationship to the response
- Exclude variables with weak relationship to response
- For continuous variables:
  - Cubic spline transformations with more knot points for the stronger predictors
  - Single DF transformations for the weaker
- More groups for stronger categorical variables
- Also consider charts against the response to help conserve DF

Building the model

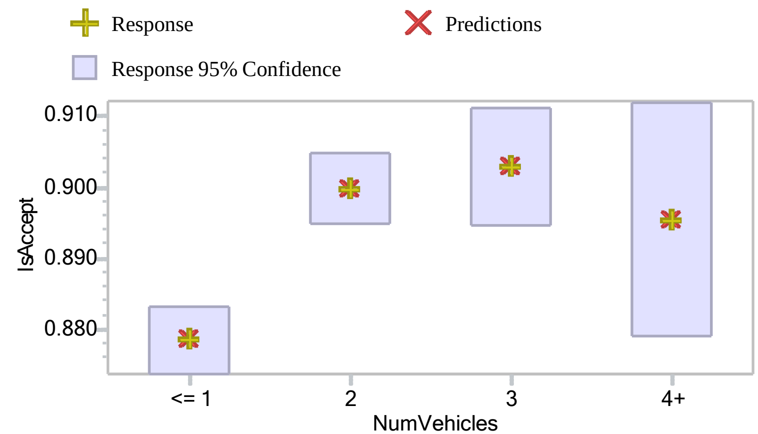


Choose predictor variables

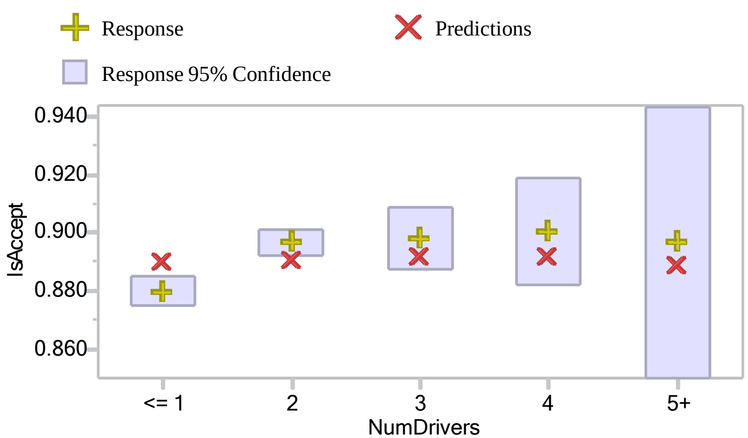
Add the most significant of correlated variables first…



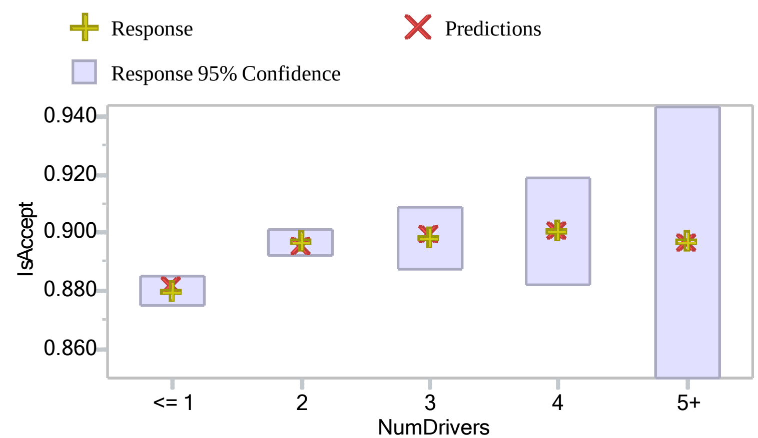
added to model



… correlated with



wait and see…

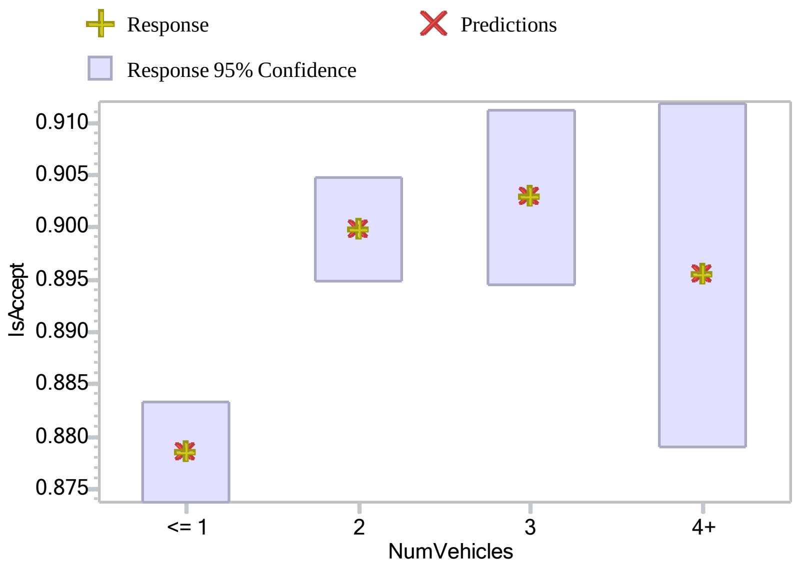


General modeling tips

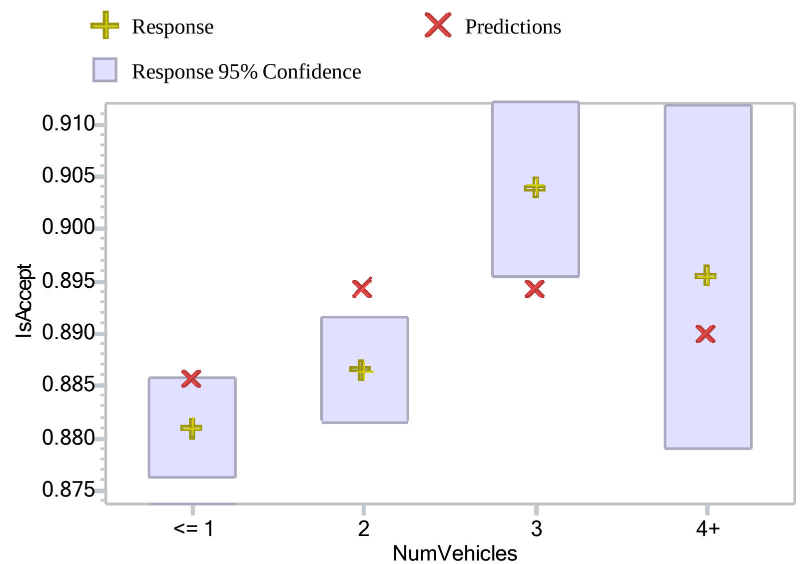
1. Data cleaning
2. Variable screening
3. Variable selection
4. Model validation

Check for signs of over-fitting

In-sample: looks good



Out-of-sample: more grouping may be necessary…



optimaldecisionsgroup
A LexisNexis® Company

Check for signs of over-fitting

- Where evidence of over-fitting is present, re-visit earlier decisions and err on side of simplicity:
  - Fewer variables, more data reduction
  - Group factors more aggressively, reduce degrees of freedom for continuous variables
- Be careful not to overuse the holdout sample

General modeling tips