TRAVELERS

Business Insurance Actuarial
Insight, Analytics, Solutions…
Advantage Travelers

Model Validation Techniques

Christopher Monsour, FCAS, MAAA

CAS Ratemaking and Product Management Seminar

March 11, 2009
Las Vegas

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.

What Models Need to Be Validated?

- **All** models need to be validated
- However, unlike many other statistical diagnostics,
  - **THE SAME CONCEPTS APPLY REGARDLESS OF THE TYPE OF MODEL**
- Examples of models:
  - Decision or Regression trees
  - Generalized Linear Models (GLMs)
  - Generalized Additive Models (GAMs)
  - Linear Discriminant Analysis
  - Ridge Regression
- You can apply the concepts you learn today to any of the above, plus just about any other type of predictive model you may fancy

The Many Meanings of Model Validation

- **Primary Meaning / Use—Quantifying Model Performance**
  - **How well can we expect this model to perform in the future**
  - **The only objective test is unseen data**
- Secondary Meanings/Uses—Using Similar Procedures for Other Goals
  - Looking at out-of-sample data during the modeling process to determine:
    - the "right" choice of predictor variables [feature selection], and/or
    - the "right" type of model, and/or
    - The "right" value of a tuning parameter
  - Looking at out-of-sample data to compare to performance on in-sample data
    - Question begged—What to do with an overfit model?

Important Caution
Data sunt omnia diuisa in partes tres

- The same out-of-sample data cannot serve both purposes above
  - If used in feature selection, then it influenced the model
  - So need additional out-of-sample data to quantify performance
- Data terminology
  - Data used in building the model (in-sample) are “training” data
  - Out-of-sample data used in guiding the modeling process are “test” data
  - Out-of-sample data against which the predictive power of the ultimately chosen model is tested are “validation” data or “holdout” data. It is sometimes important for this data to be out-of-time as well.
    - E.g., if you are modeling severities of homeowners losses, you don’t want claims from the same storms in the training/test data and the validation data

What is Mandatory?

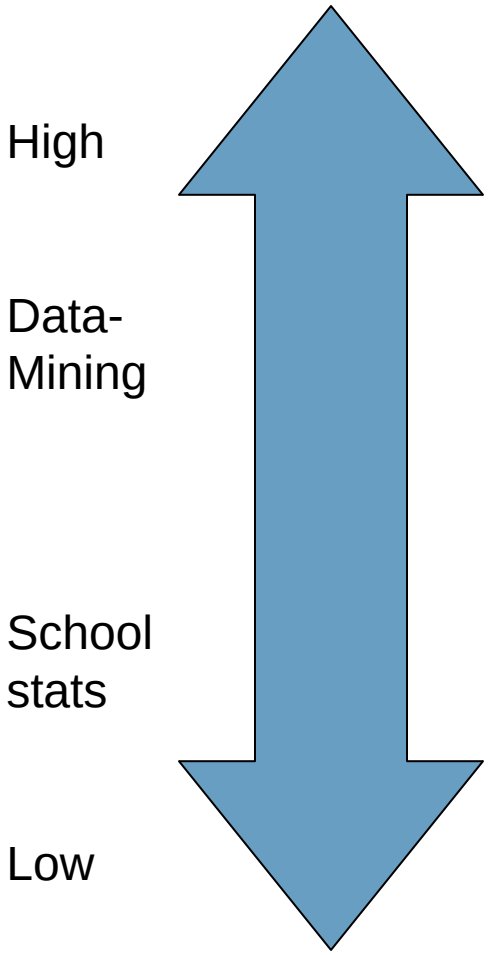
- Validation Must Involve Unseen Data in Some Way
  - If you think your data are rich enough to support the most complex model possible:
    - Then **USE** the most "complex" model possible
    - It's easy to implement….
    - It's a lookup table, because
      - Include enough predictors to distinguish any two observations
      - Include all the multi-way interactions
      - And your prediction will always be precisely what happened last time
  - A lookup table has "perfect" goodness-of-fit, but that doesn't make it a good model

Models and Tables—Example Lookup Table Construction vs Limit

Prop Pure Prem	Low	Med-Low	Med-High	High
Wood	25	2	42	67
Brick	4	54	21	34
Stone	3	13	23	33
Fire Res	7	7	14	13

Bailey knew we didn't want predictions that looked like this!

Danger Scale



Type of Search

Lookup Table

Neural Network

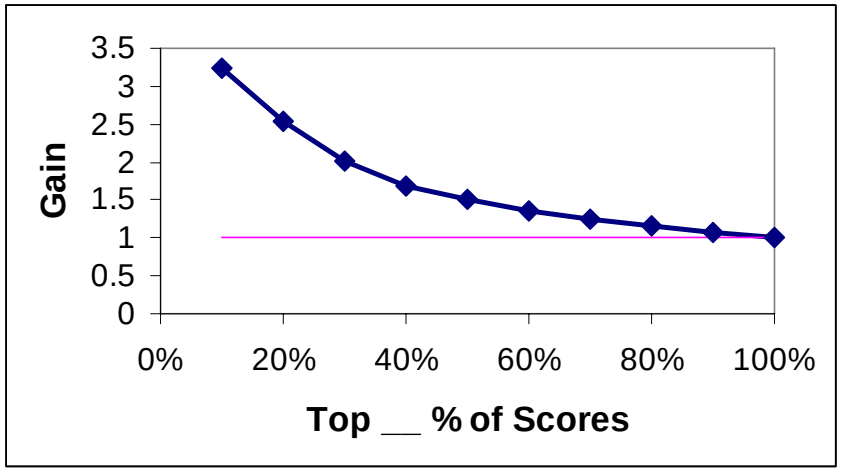
Adaptive methods: MARS, GAMs, splines

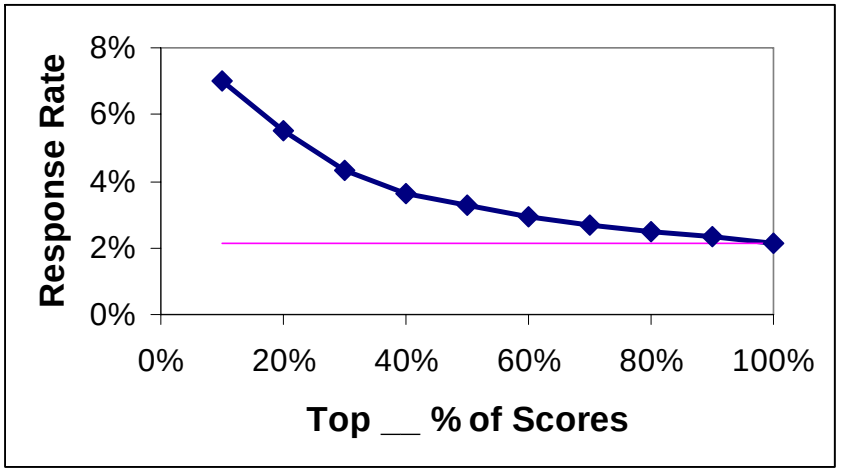
Automated feature searches (stepwise)

Search “by hand”

No search

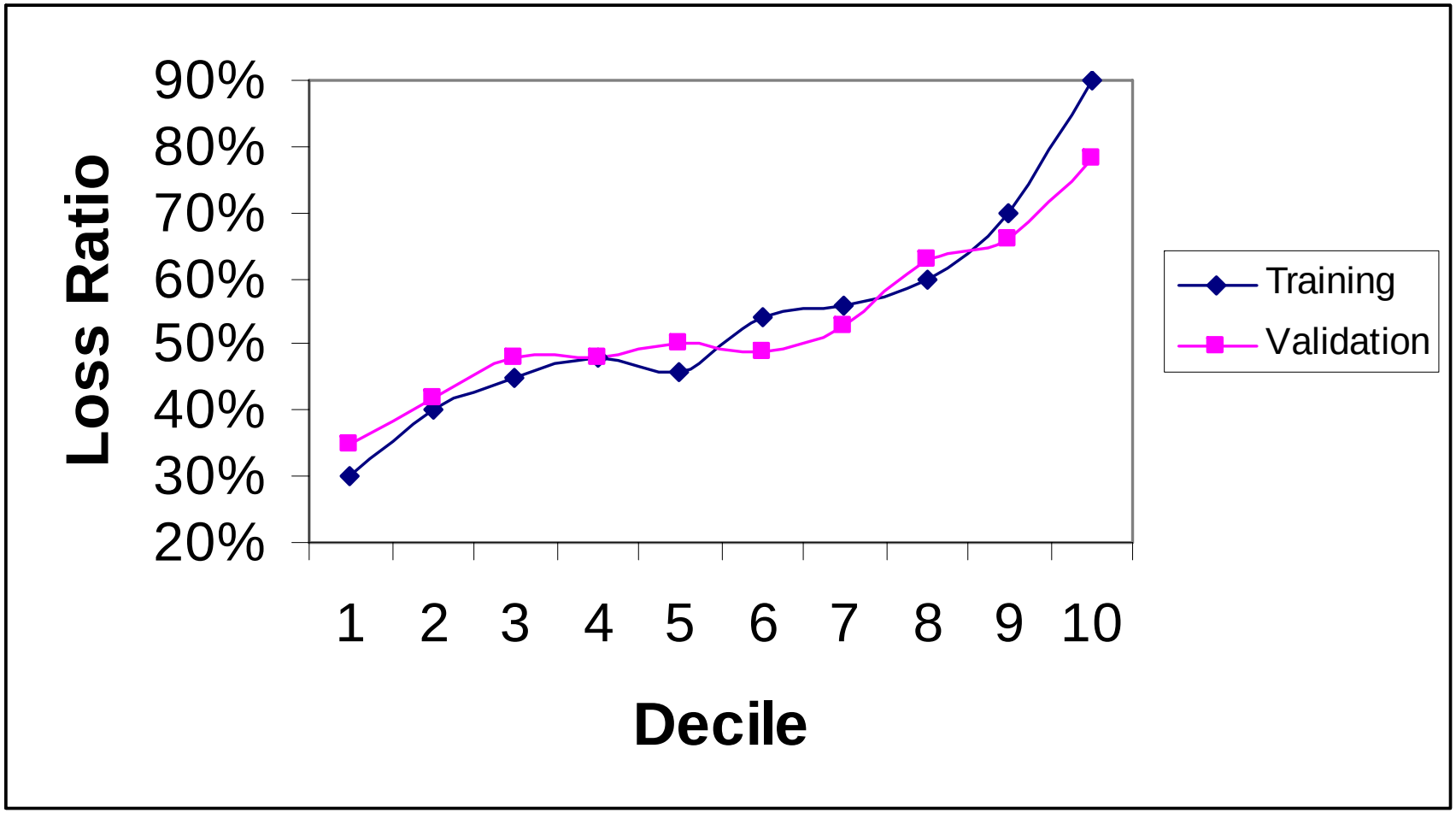
Gains Charts





- Classic chart to diagnose a response model
- Use the model to “score” the validation data
  - i.e., assign the modeled probability of a response
- SORT by the modeled probability from most likely to least likely
- Plot the actual response rate in the validation among the highest scoring x%
- Typically do NOT even plot the training data…goal is purely to show objective performance

Decile Charts



Decile Charts

- Again, you sort by predicted value
- You show actual value
- The validation line is key
  - This shows the actual predictive power of the model
- The discrepancy between the validation and actual lines is useful
  - In modeling (using test rather than final validation data), to diagnose overfit
  - In implementation: If implementing as a rating algorithm, discrepancies between the training line and validation line suggest "shrinking" extreme estimates
- Nothing "magical" about deciles: Use quintiles, vingtiles, whatever your data will support

The Many Meanings of Model Validation

- Primary Meaning—Quantifying Model Performance
  - How well can we expect this model to perform in the future
  - The only objective test is unseen data
- **Secondary Meanings—Using Similar Procedures for Other Goals**
  - **Looking at out-of-sample data during the modeling process to determine:**
    - **the "right" choice of predictor variables [feature selection], and/or**
    - **the "right" type of model, and/or**
    - **The "right" value of a tuning parameter**
  - Looking at out-of-sample data to compare to performance on in-sample data
    - Question begged—What to do with an overfit model?

Choice of Predictors

- Do NOT use validation data for this
  - Just training and test
- Divide dataset into training and test data
- Check that predictors still show up as significant if you model the test data
- Or divide the training data into many pieces
  - Say 5 pieces
  - Model each 1/5 (or each 4/5)
  - Only include predictors that were significant in at least 2 (or 3, or 4) of these 5 models

The Right Type of Model

- How to test regression tree against a GLM
  - Loglikelihood, AIC, BIC won't help
    - They aren't nested and don't nest in a common model
  - Even picking a cost function and measuring it on the in-sample data won't help
    - Because measuring on training data is "optimistic" and will be optimistic by different (unknowable) amounts for different models
  - Test against your cost function (squared error, median absolute deviation, misclassification cost, etc.) on validation data!
    - Objective test
    - If you cost function can be expressed in dollars, even better!
  - If you do this, you need another holdout dataset to get an objective measure of the chosen model!! (Winner's curse)

Compromises instead of Believing the Data 100%

- Using characteristics "in part"
  - Credibility, shrinkage methods (lasso, etc.), local regression estimates, smoothing splines
  - Uses less than a full degree of freedom for each parameter
- Even that will not get the degrees of freedom down enough
  - Some characteristics you shouldn't use at all
- **But, at the end of the day, how many degrees of freedom *should* you have?**
  - By what standard should we judge this?

The Right-sized Model

- Why not use only seen data but penalize the goodness-of-fit measure for the number of parameters and/or degrees of freedom?
  - The "information criteria", AIC, BIC, etc., do this
  - Limitations:
    - The number of parameters may be the wrong basis for the penalty
      - E.g., if using shrinkage techniques, like ridge regression, or credibility, or hierarchical or mixed models, the effective dfs may be much smaller than the number of parameters
      - Even if you have a good way to compute the effective degrees of freedom, that doesn't penalize for the size of the search…

        If you have 20 features, the "best" 8 feature model implies a search of 125,970 models; "an" 8 feature model implies a search of *1* model.

The Right-sized Model

- Why not use significance tests to decide whether to include a variable:
  - Well, of course, one does, but…
  - Raw significance tests also do nothing to adjust for the size of the search
  - Tests that are directionally correct may not function correctly in absolute terms when modeling assumptions are violated (i.e., always)
    - For example, widths of confidence intervals are very sensitive to the scale parameter in most GLMs
    - But the scale parameter has to be estimated from the data and may not itself be very certain
    - Other sources of problems include unmodeled heteroskedasticity

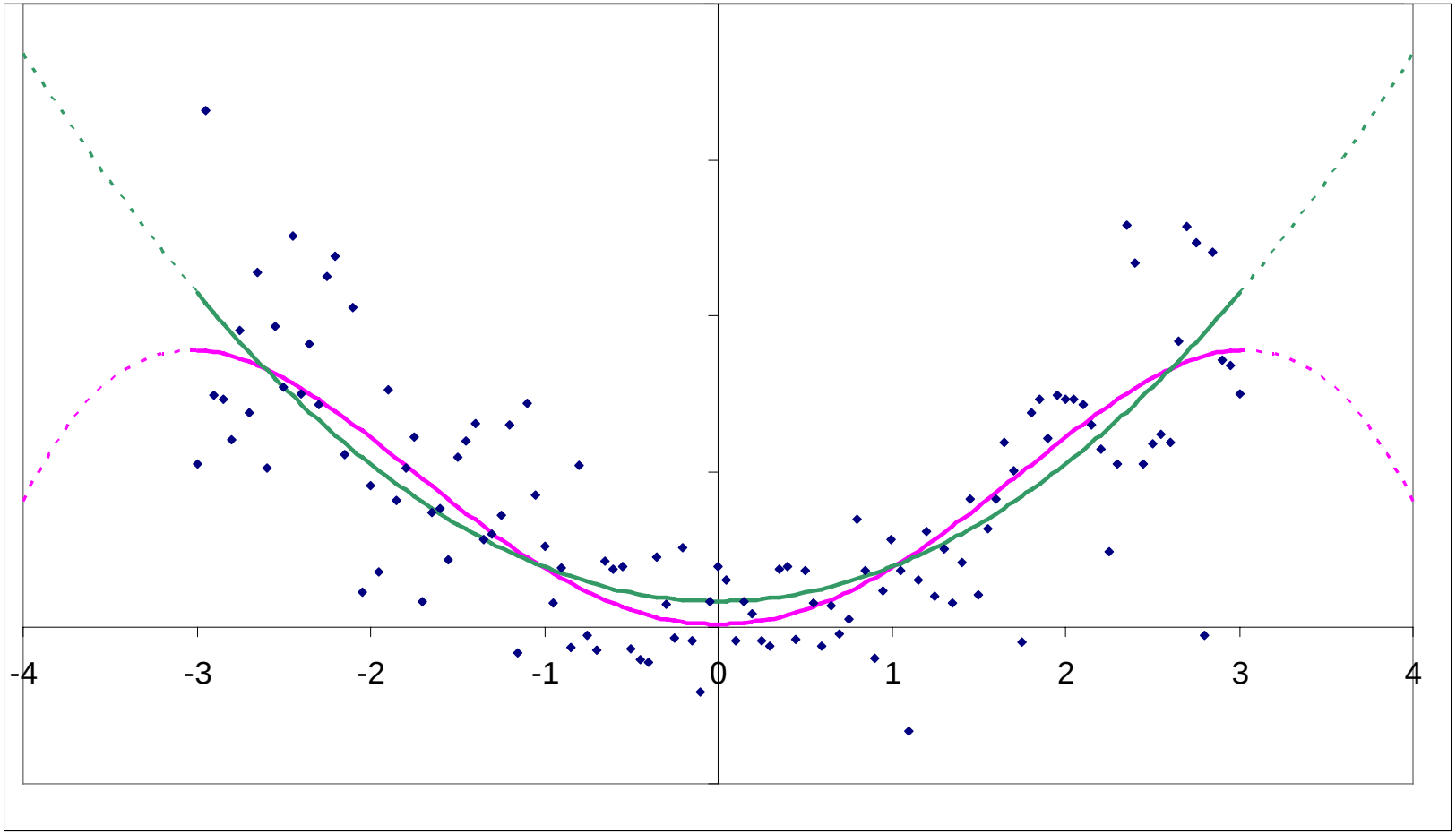
Cross-Validation

- Divide the data into N pieces
  - N=5 or 10 typical; N=2 convenient if hurried
- Run the model on each 4/5 or each 9/10
  - This results in N models, each on a high percentage of the data
  - Each datapoint has been left out in building exactly 1 model
  - Compare each actual observed value to the value predicted by the model that didn't see it
    - Use this to compute goodness-of-fit (squared error, misclassification rate, etc.)
    - Use this to compare models of varying complexity
      - Fewer or more predictors
      - Different values of a tuning parameter (e.g., K in a Bayesian credibility setup)

Cross-Validation

- In the data mining and machine learning community, often used to do the objective validation of model power
  - Only works because the model-building process is entirely automated
    - Each 9/10 model and the model on the entire dataset are built without knowledge of the other 10 models
      - Not just the fitting of parameters is independent
      - So is the choice of variables, indeed the entire process
      - If the process was open to CART or MARS before looking at the first 9/10, the process needs to re-make even that decision 10 more times
    - This is not possible when human beings are part of the modeling process

$A+Bx^2$ or $A+Bx^2+Cx^4$?



Parameter Estimates

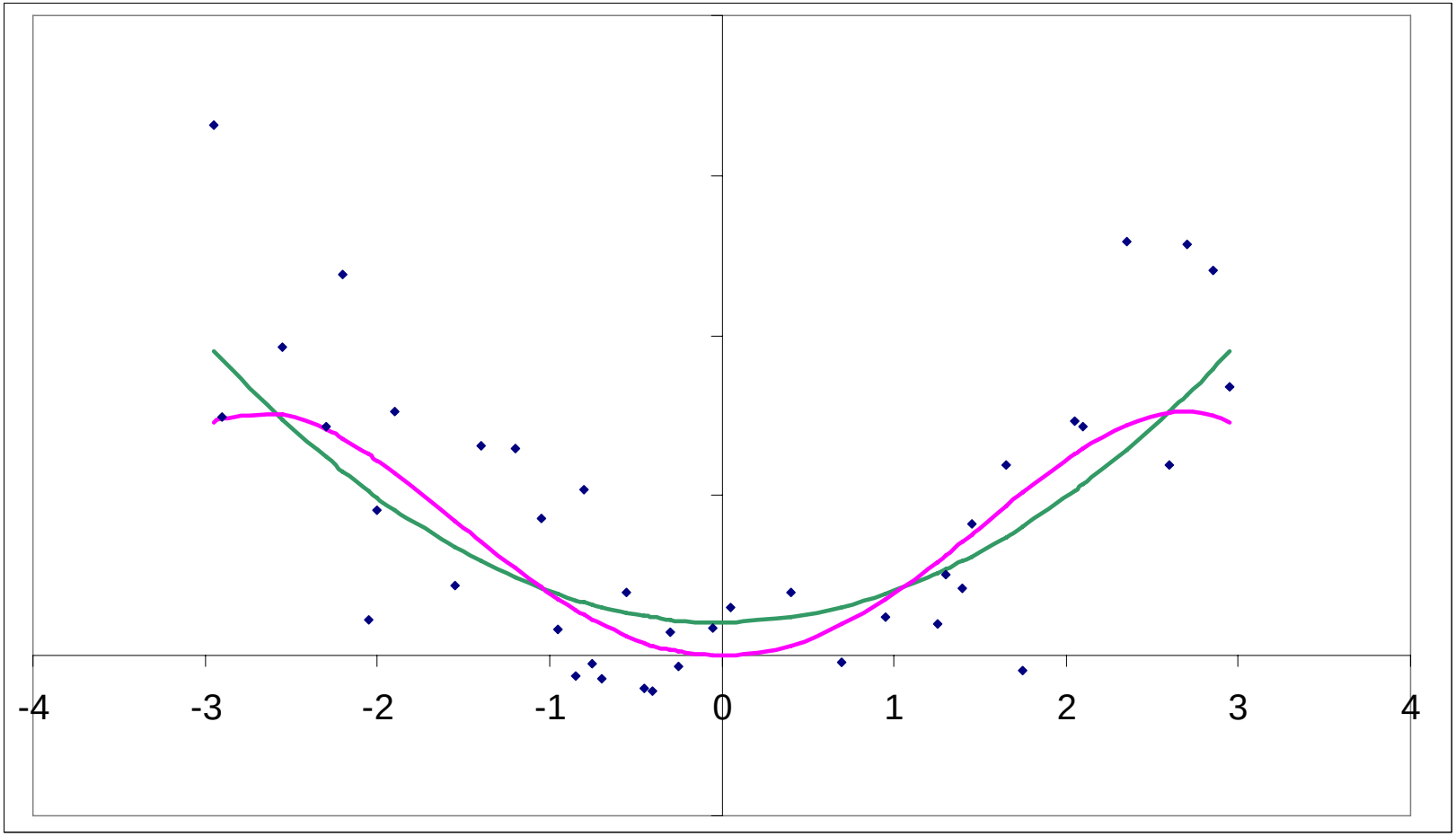
	Estimate	Std Dev	p-value
$A+Bx^2$			
A	0.68	0.28	0.015
B	0.88	0.07	2.26E-37
$A+Bx^2+Cx^4$			
A	0.09	0.34	0.803
B	1.53	0.23	5.08E-11
C	-0.083	0.029	0.003

So you need an x^4 term, right?

Hypothesis Test is OK?

- Or do you need an x^4 term?
- Do the errors look identically distributed?
  - Or are the data heteroskedastic?
- So assumptions are violated that may severely impact the hypothesis tests
- Let's look at another comparison of the two models:

1 / 3 of the data vs models fit on other 2 / 3



This is done 3 times, once for each 1/3

Cross-Validation to the Rescue

Sum of Squared Errors

	$A+Bx^2$	$A+Bx^2+Cx^4$
Full Model	505.6	471.3
3-Fold Cross-Validation	564.2	565.6

Mean Squared Error

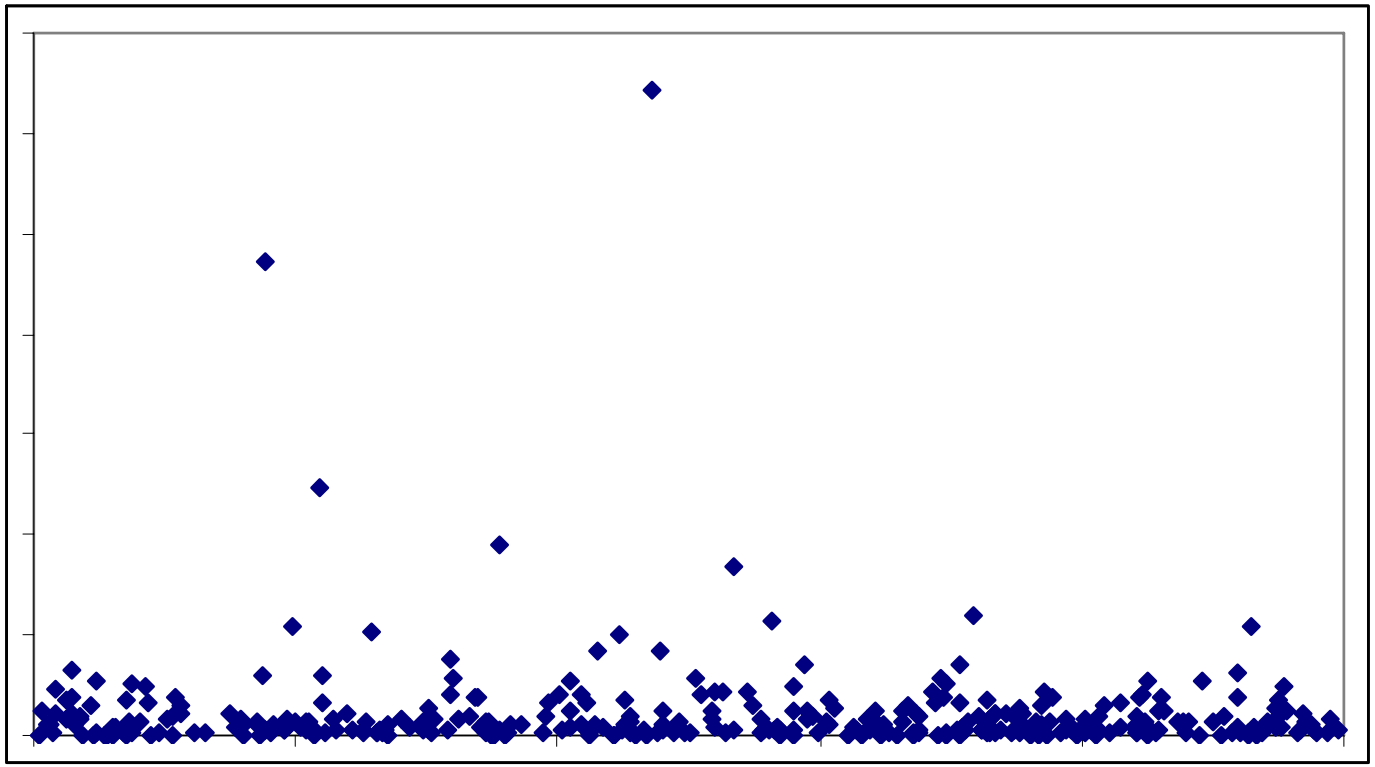
	$A+Bx^2$	$A+Bx^2+Cx^4$
Full Model	4.18	3.90
3-Fold Cross-Validation	4.66	4.67

Note the optimism of the full model (in-sample) errors

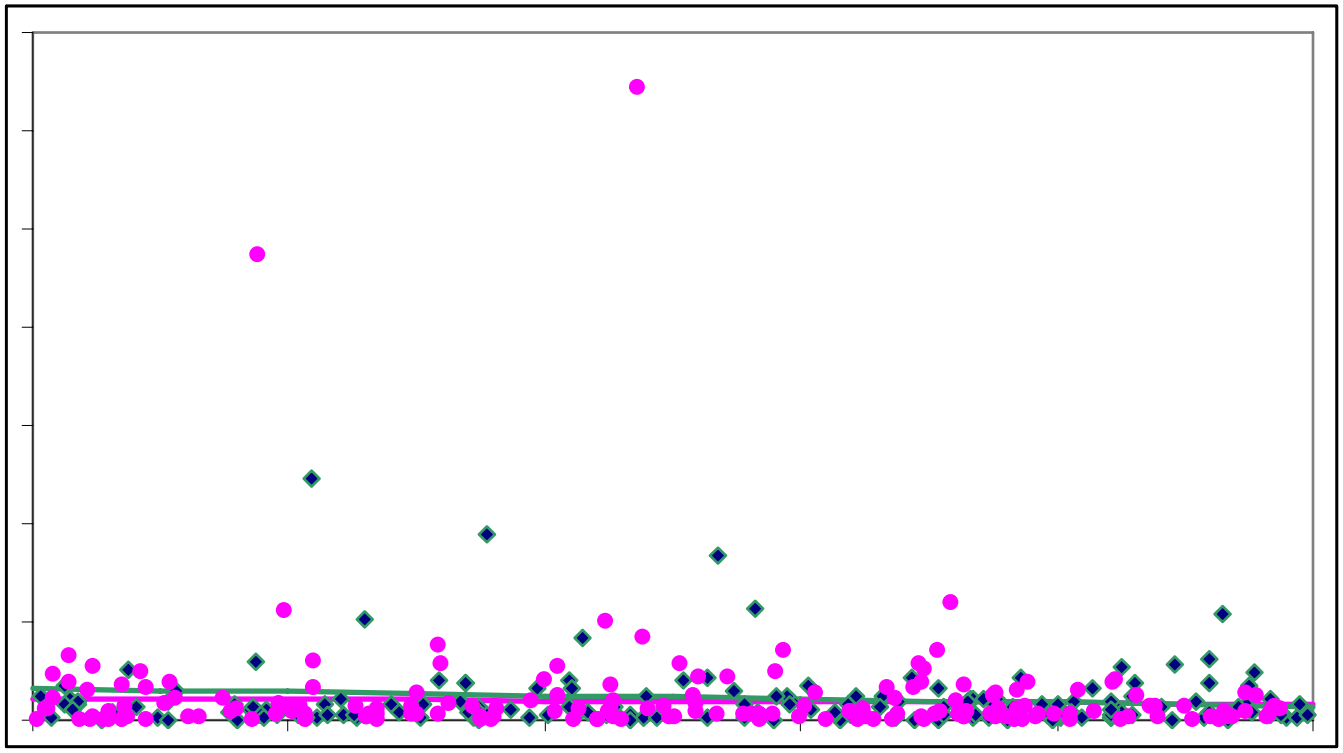
Note that the 4^{th} power term is completely unnecessary

I generated the data from a quadratic (and added heteroskedastic errors)

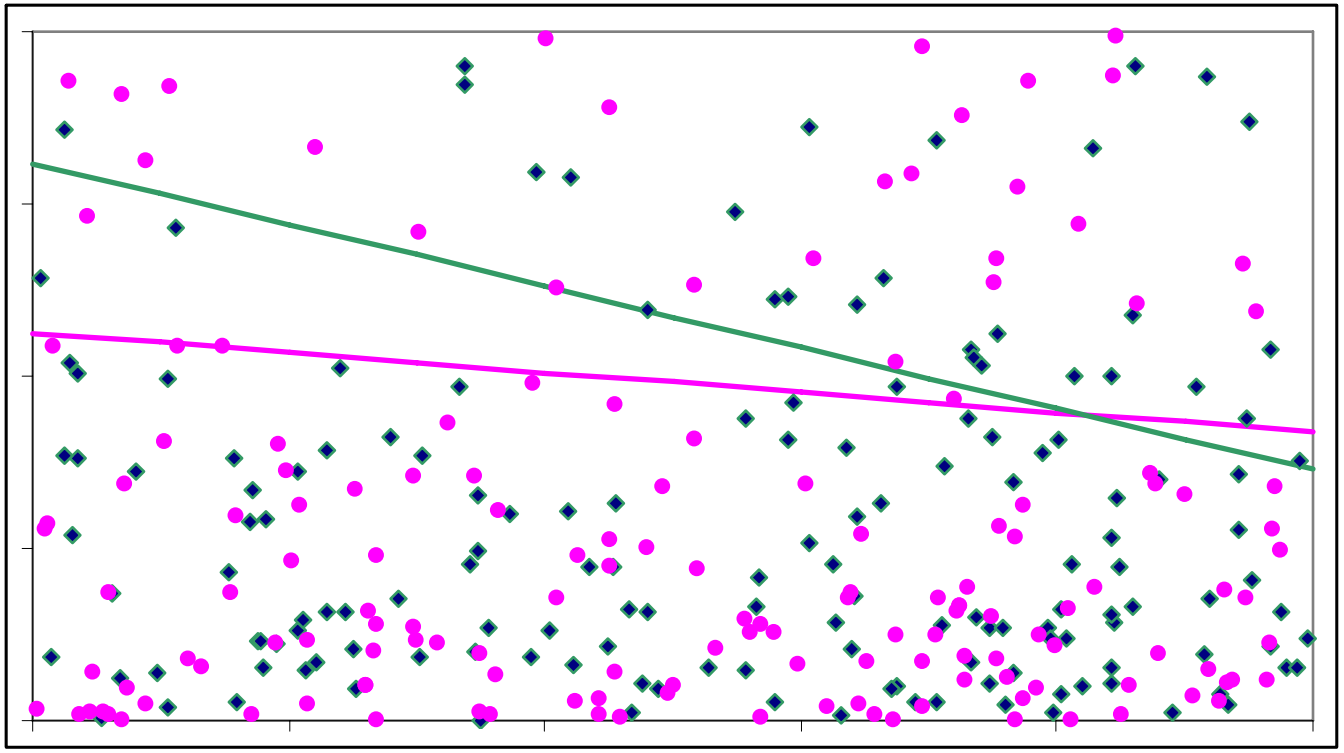
Simplified Version of Property-Casualty Insurance Data



Split into two and added lines



Zooming on the lines



Lesson of the above

- Insurance data often has a few observations that are outliers
  - But we can't throw them out because they are the observations that matter most
- Therefore:
  - Choose your model carefully (linear regression without transforming y should be obviously bad idea with the above)
  - Remember that you don't have as much information as the size of your dataset might indicate
  - Remember that you can overcome optimism in classical confidence intervals using cross-validation

Testing on Seen vs Unseen Data

In-sample tests

- Must adjust for "degrees of freedom"
- Many tests oriented toward inferential power
- Tests sensitive to fussy statistical assumptions
- May need deep statistical knowledge to interpret
- Difficult to present results to management
- May require adjustments if observations are correlated

Out-of-sample tests

- No need to adjust for degrees of freedom
- Tests typically oriented toward predictive power
- Tests purely empirical; only simple assumptions involved
- Have commonsense interpretations
- With modest effort, usually presentable
- In some cases, may need to be an out-of-time as well as out-of-sample test

Model Validation Today

- Model validation is a serious topic
- Regulators require some financial institutions to have a separate department that validates, for example, consumer creditworthiness models
- Should there be an actuarial standard of practice addressing validation of statistical models
  - Topics such a standard might address
    - When is out-of-time validation rather than just out-of-sample validation critical?
    - What steps should be taken to ensure knowledge of the validation data has not crept into the model-building process?
      - For instance, split off the validation data before or after EDA?
      - Splitting it too early makes balancing to control-totals difficult

Some References

Cross-Validation and Out-of-Sample Validation:

- Breiman, Friedman, Olshen, and Stone, *Classification and Regression Trees*
- Hastie, Tibshirani, and Friedman, *The Elements of Statistical Learning*
- Domingos, “The Role of Occam's Razor in Knowledge Discovery”, Data Mining and Knowledge Discovery, 3, 409-425

Re-sampling more generally:

- Davison and Hinckley, *Bootstrap Methods and Their Applications*
- Wolter, *Introduction to Variance Estimation*