Validating Models

Midwest Actuarial Forum

14 September 2012

Cincinnati, OH

Christopher J. Monsour, FCAS, MAAA

VP, Predictive Business Applications

Disclaimers, Copyright, etc.

The purpose of this presentation is to provide information, rather than advice or opinion. It is accurate to the best of the speaker's knowledge as of the date of the presentation. Accordingly, this presentation should not be viewed as a substitute for the guidance and recommendations of a retained professional. Any references to non-CNA Web sites are provided solely for convenience, and CNA disclaims any responsibility with respect to such Web sites.

To the extent this presentation contains any examples, please note that they are for illustrative purposes only and any similarity to actual individuals, entities, places or situations is unintentional and purely coincidental. In addition, any examples are not intended to establish any standards of care, to serve as legal advice appropriate for any particular factual situations, or to provide an acknowledgement that any given factual situation is covered under any CNA insurance policy. Please remember that only the relevant insurance policy can provide the actual terms, coverages, amounts, conditions and exclusions for an insured. All CNA products and services may not be available in all states and may be subject to change without notice.

CNA is a registered trademark of CNA Financial Corporation. Copyright © 2012 CNA. All rights reserved.

Knowledge Discovery?

“They think that intelligence is about noticing things that are relevant (detecting patterns); in a complex world, intelligence consists in ignoring things that are irrelevant (avoiding false patterns).”

- Nassim Taleb

“All models are wrong, but some are useful”

- George EP Box

What Is Model Validation?

- Many things
  - Verification of operational data feeds
  - Testing that what is implemented is what was intended
  - Does the model make sense to experts?
  - Does the model work on out-of-sample data

Which Models Must Be Validated?

ALL MODELS MUST BE VALIDATED

Why Do Models Need to be Validated on Out-of Sample Data?

- Ensure the model fits signal, not noise
- Obtain an **unbiased** measure of model effectiveness
- Obtain some assurance that the model generalizes
- In order to be a statistician, actuary, economist, etc., rather than a ….
- Avoid the Journal of Irreproducible Results

Assumptions Made by Models Rarely Hold— And May Make Bad Trade-Offs Even When They Do

- Many models (fixed effects models, including GLMs) give full credibility to any category you call out
  - Predictions may be less stable than you'd like
  - Complicated Answer: Mixed Effects or Hierarchical Models or Shrinkage
  - Low Budget Answer: Don't Overparameterize
- Heteroskedasticity (and Overdispersion)
  - Often lead to reported CIs and p-values that are smaller than reality
- Omitted Covariates
- Correlated Observations
  - CI and p-value optimism
  - Wrong weights!

Is Your Big Data Small?

- **Is** Your Data Big Data?
  - – How many claims?
  - – How many material claims?
  - – For one specific cause of loss?
  - – Are the claims independent?
  - – Are the outliers noise? Or are they the only important signal?

Is Your Big Data Small?

- **Are** Your Data Big Data?
  - Lots of variables
    - p > n ?   (probably not, but maybe p > 500)
  - Multiple testing—Handled appropriately?
    - Classical (Fisher or Neyman style) hypothesis testing
      - Requires Bonferroni or similar adjustment
    - False discovery rate (FDR)
      - Benjamini & Hochberg (1995)

Aside: Testing vs False Discovery Rate

	Accept Null	Reject Null
True Null	A	B
False Null	C	D

- Classical: Significance = B/(A+B); Power = D/(C+D)
  - Family-Wise Error Rate: Control Prob(B>0)
    - Bonferroni: Can do this by dividing desired probability by number of hypotheses
- False Discovery Rate: Control B/(B+D)

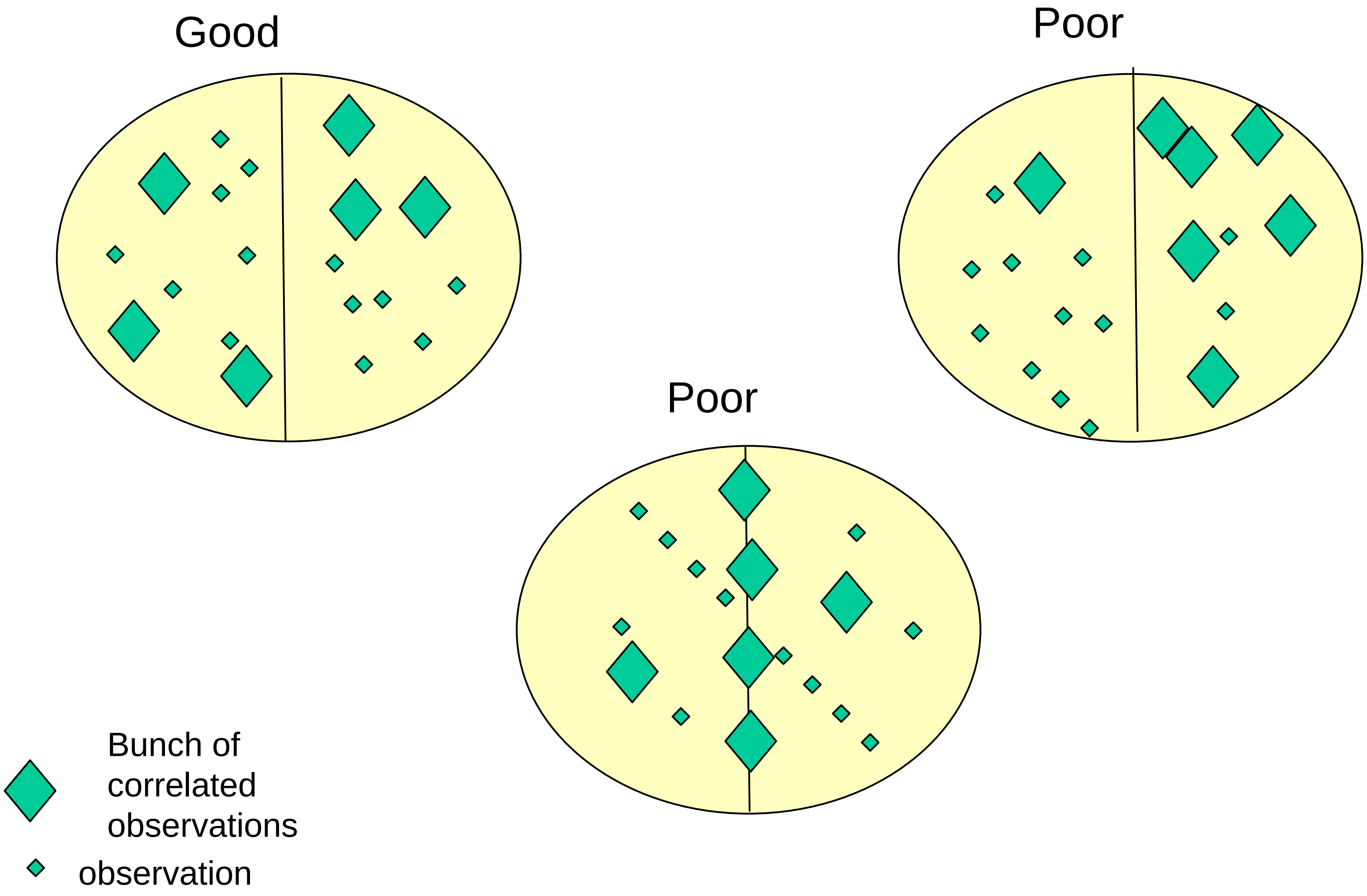
Data sunt omnia divisa in partes tres

- Some data must be held out until the end (call it “holdout”)
- But you want to look at out-of-sample predictions during the process
  - Split into “training” and “test”
  - Occasionally swap roles of training and test, or re-draw.

How to Split the Data?

- RANDOMLY
  - Necessary but not sufficient—what is the sampling frame?
- Want no correlation between the different pieces
  - What this means depends on the situation
    - Keep accident dates together for a wind severity model
    - Keep the same policy together across all years for complex policies
    - Keep each year or time period together if the economy is important in the model

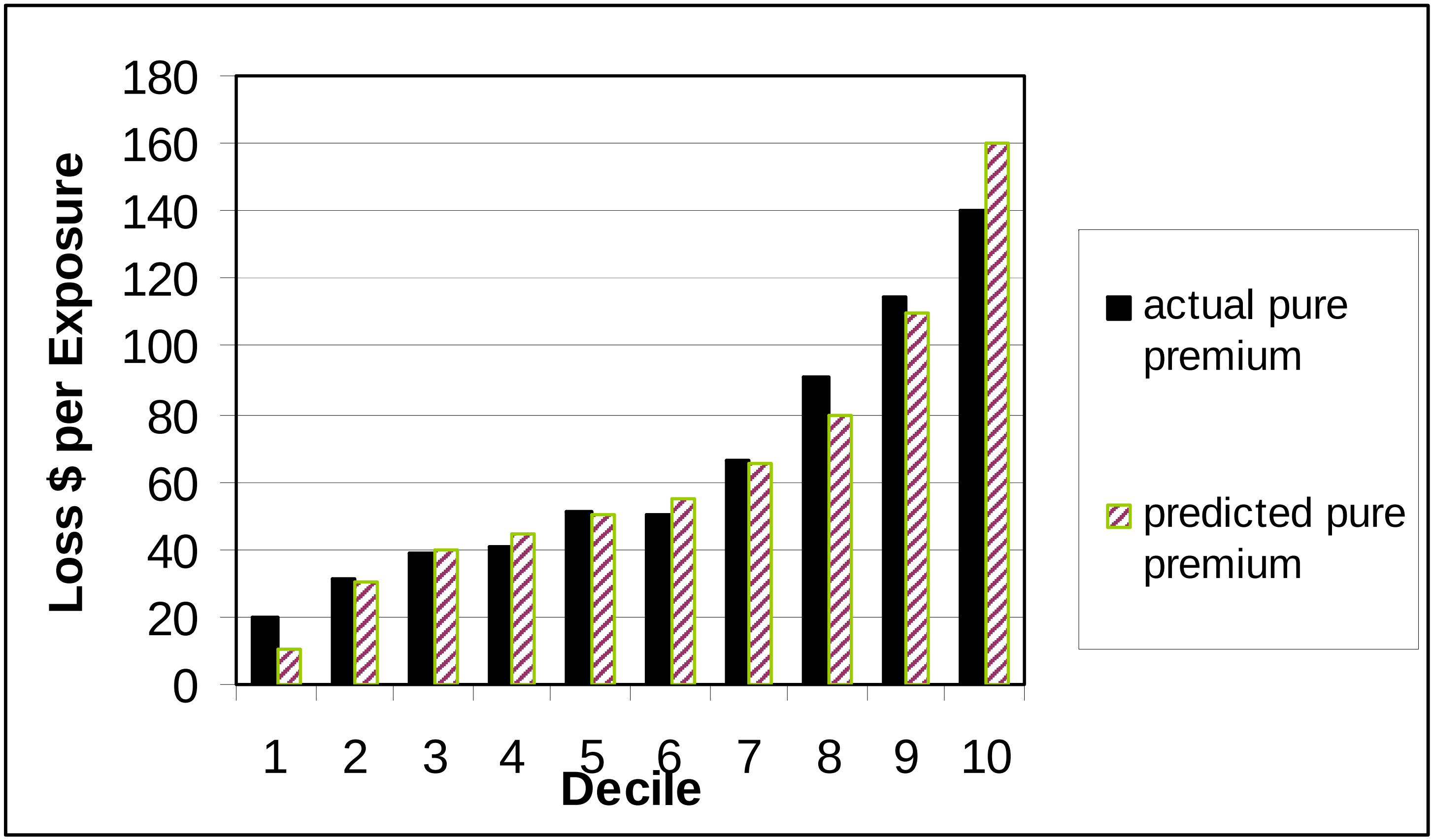
How to Split the Data?



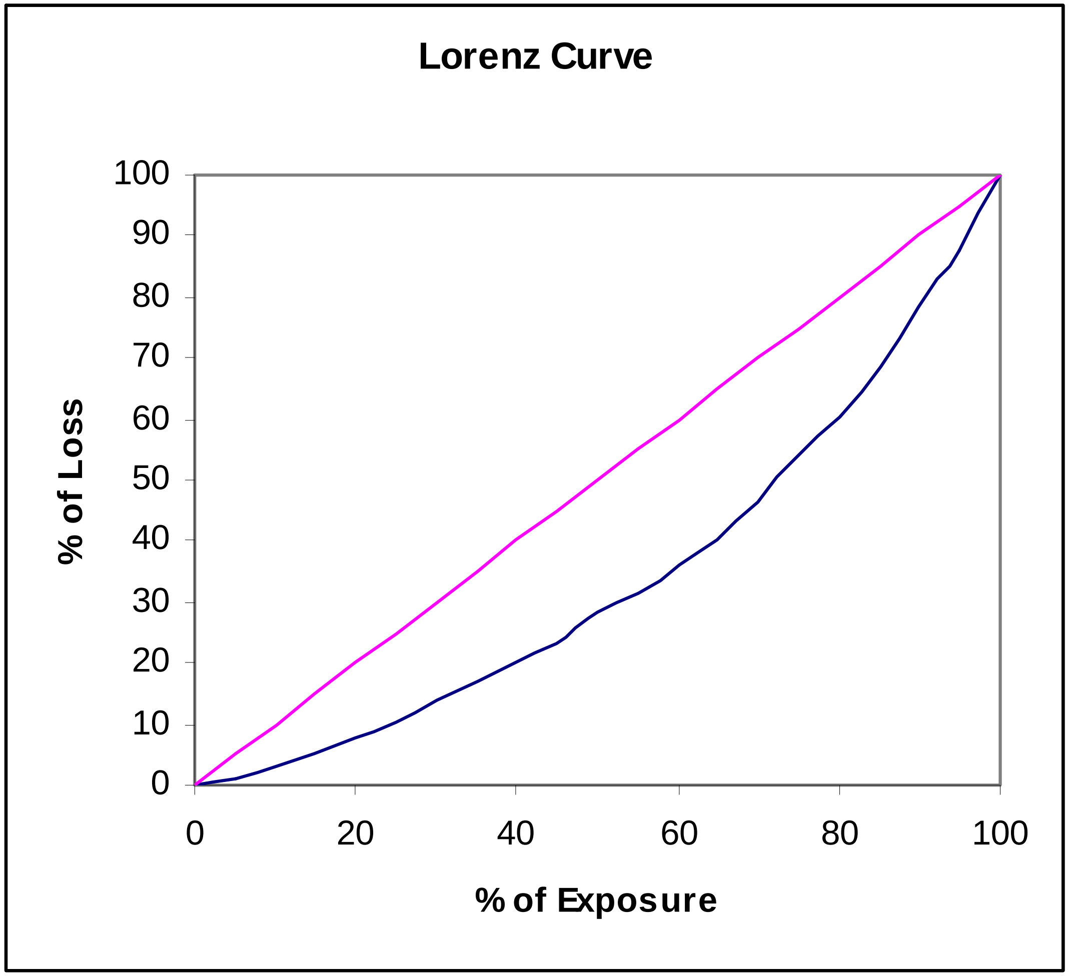
Overall Model Validation

- Numerical
  - Out-of-Sample Cost
    - Good if you know the cost function
  - Out-of-Sample Likelihood
    - Good for comparing models with different covariates but the same underlying structure
- Visual
  - "Decile" Chart
  - Lorenz Curve (with associated numerical measure…Gini coefficient)

"Decile" Chart



Lorenz Curve



Gini index = 31.6%

Overall Model Validation—Best Practice

- May measure a training model on test or *vice versa* while building
- At conclusion, combine training and test, refit, measure on holdout
  - This measurement is *the* measurement of Gini / lift / cost, etc.
- Then refit on all the data…but do NOT produce more decile charts, nor a new Gini index, etc.
- Also, while training and test may reside in the same dataset with an indicator variable, holdout data should be stored separately.
- The model can only be measured on holdout ONE TIME
- The Holdout should NOT be small (at least 30% of data)

Overall Model Validation—Best Practice

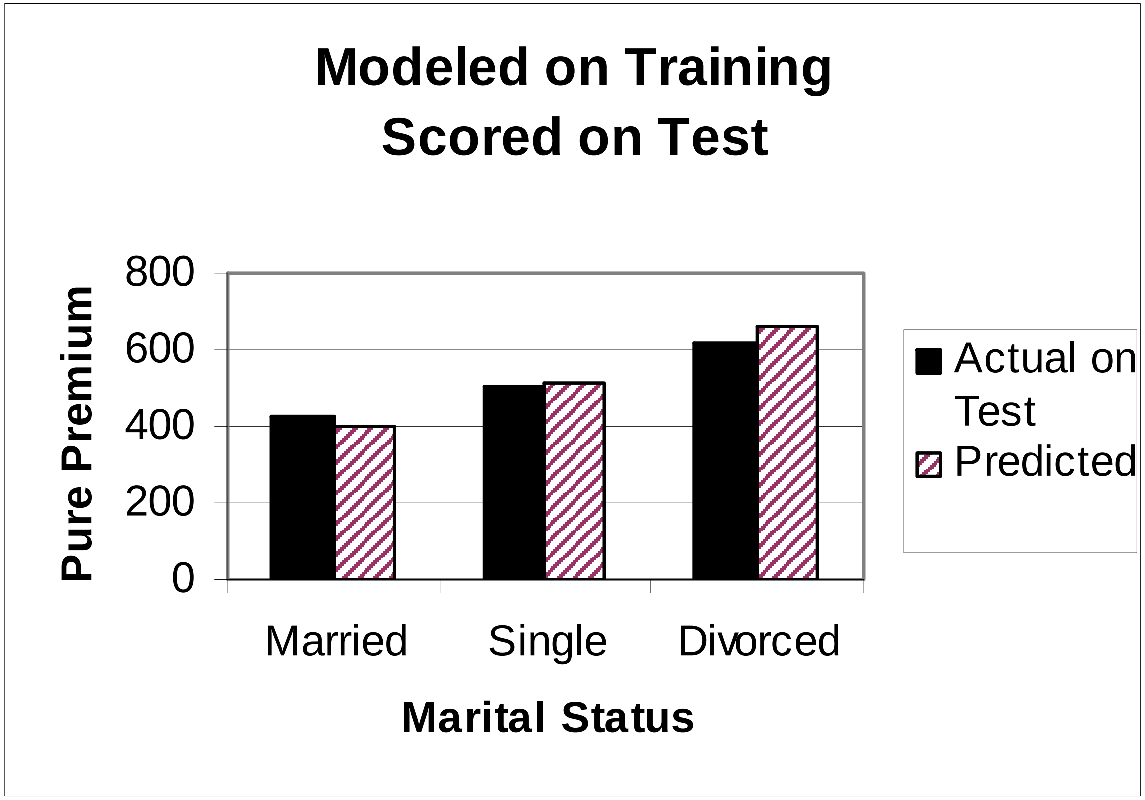
- “Exposure” may mean many things.
  - Exposure Used in the Rates
  - Exposure x class rate
  - Manual Premium
  - Policy-Year
  - Item-Year
  - Previous Model Prediction
- Maybe used only one of these to **build** your model, but any can play a role in validating
- In particular, can validate new model on old, and *vice versa*

Other Uses of Validation

- View generalization of predictions to unseen data (by subset)
- Consistency of predictions across subsets
- Estimating tuning parameters

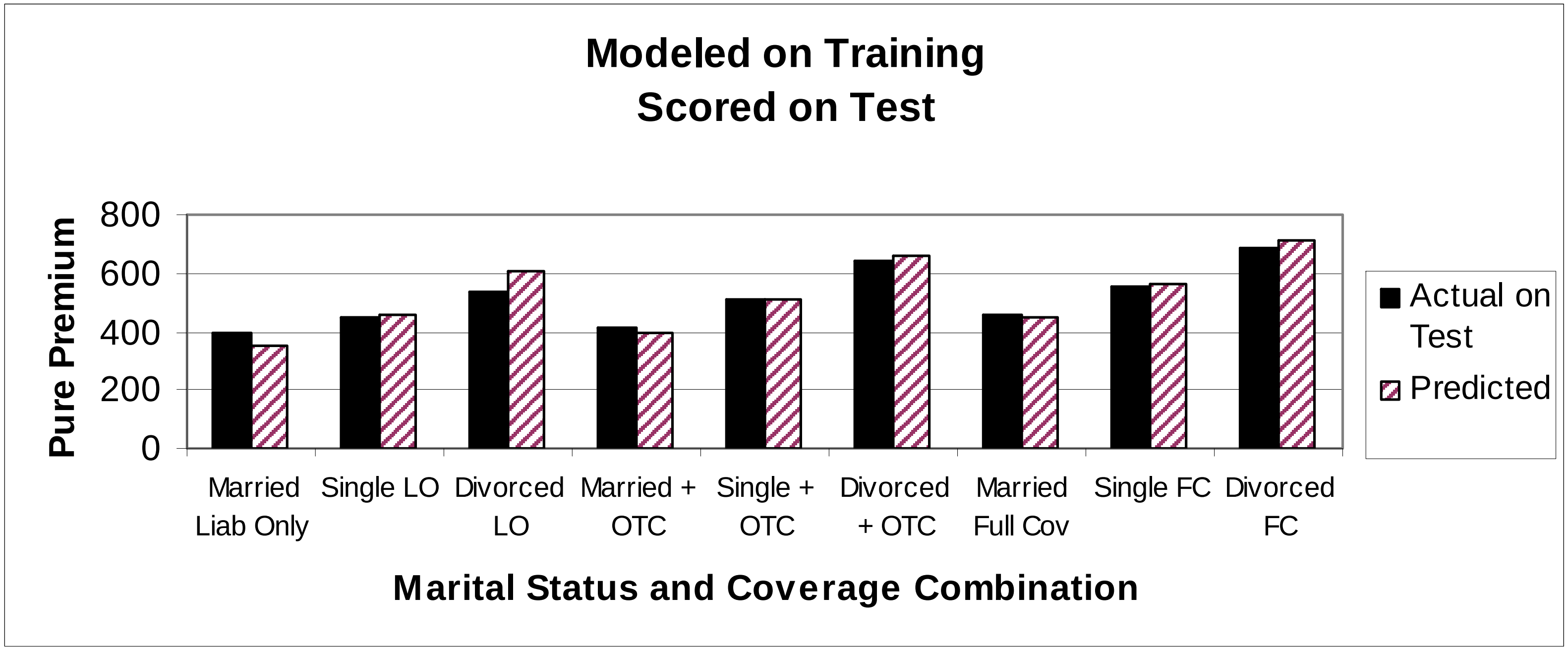
Generalization of Predictions to Unseen Data

ILLUSTRATION ONLY



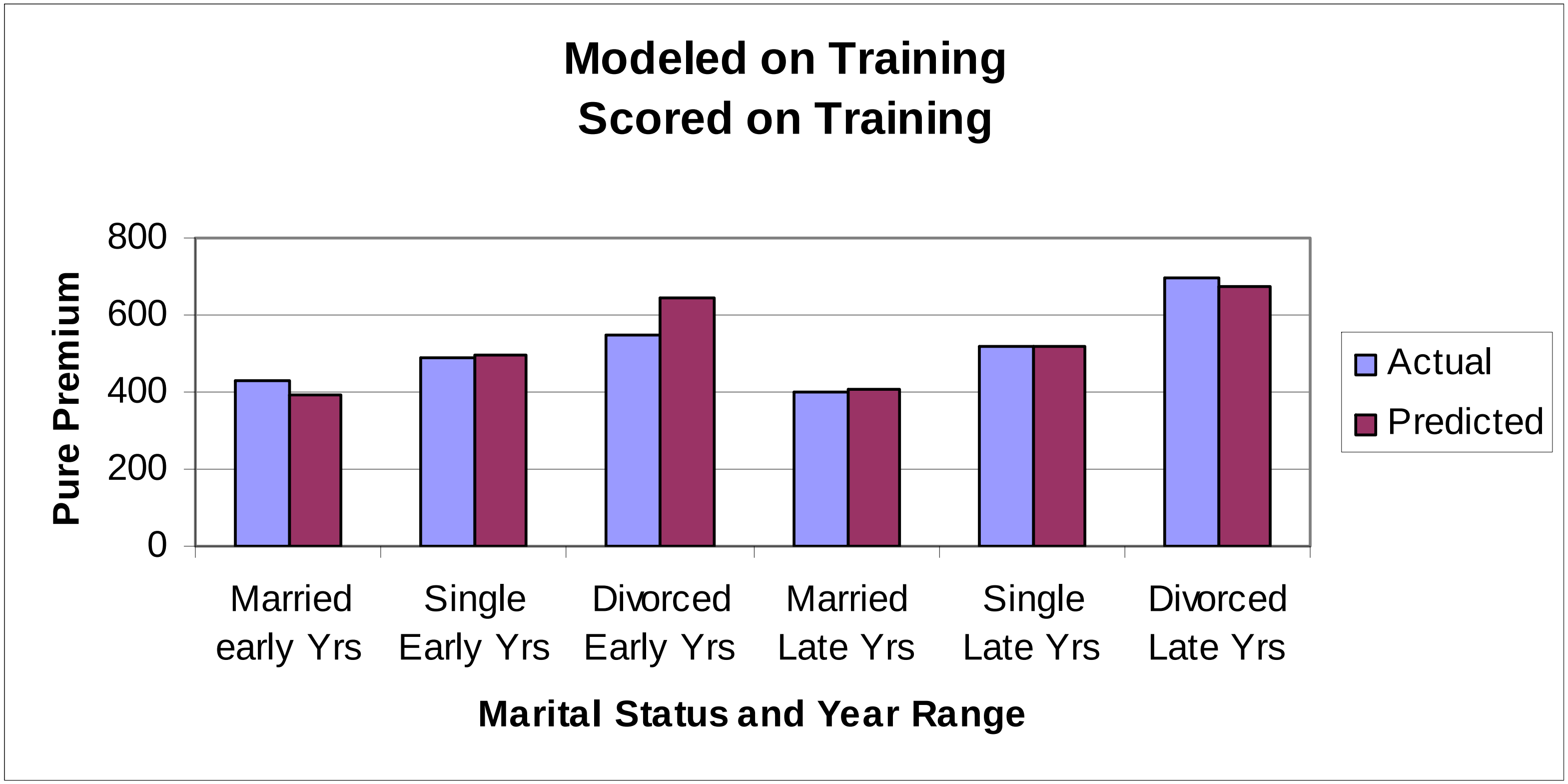
Generalization of Predictions to Unseen Data

ILLUSTRATION ONLY



Consistency Across Subsets

ILLUSTRATION ONLY



Many Types of Models Use Tuning Parameters

- Stepwise Regression (significance levels for entry and exit)
- Ridge Regression and Lasso (the constant factor in the parameter size penalty)
- Bühlmann Credibility (the ballast, K)
- Smoothing Splines (the constant factor in the roughness penalty)

Example of Tuning Parameter

- LASSO

Instead of minimizing the negative log-likelihood, we minimize:

$$-LL(\{\beta_i\}) + \lambda \Sigma_{i\neq 0}|\beta_i|$$

- Ridge

Same thing, but uses the squares of the parameters instead of their absolute values

Using Cross-Validation

- Instead of splitting non-holdout data into two pieces, split into N.
  - Example N = 5.
- Build a model on each 80%, and score on the other 20%
- Repeat for each 80%/20% split.
- Every data point will be out-of-sample once.
  - Compare its dependent variable value with the prediction for which it is out of sample
  - Get Gini or cost, or whatever measure you like

Using Cross-Validation

- Repeat for many values of the tuning parameter
- See what optimizes your cost function
- Often many models in the series come close to optimum

Coda: In-Sample Metrics

- In-sample, taking a more complex model always gives a better fit!
  - So must penalize somehow to have any hope
  - Akaike Information Criterion (AIC)
- Shrinkage techniques (optimize penalized likelihood to restrict model)
- Bootstrapping
- Significance testing
- False Discovery Rate techniques

Why Right-Sizing a Model is Hard

- The models have a good union
- Large Sample Asymptotics May Fail
  - Most stuff you use relies on this, e.g., likelihood ratio test (LRT)
- Violation of model assumptions
  - Distributional Assumptions
  - Heteroskedastic Errors
  - Measurement Errors

Confidence Intervals and p-values

- Need the right measure of dispersion
- Do you like your linear regression estimate of σ?
  - Your GLM estimate of ϕ? (By ML? Deviance? Pearson χ^2 ?)
- If you don't trust the estimate, might bootstrap your confidence intervals to try to calibrate your estimate.
- By no means assume that $\phi=1$ (in $\mathrm{Var}(y) = \phi\mu$) just because you are doing a Poisson regression

Akaike Information Criterion (AIC)

- AIC = -2*LL+2*p
  - Smaller is better
  - Problem: How much smaller is smaller enough?
    - If you want to hypothesis test, LRT instead?
    - LRT stricter for small difference in p
    - AIC stricter for large difference in p (crossover at $\Delta p=7$ for sig level of 5%)

AIC and the Dispersion Parameter

- "In theory" you estimate the dispersion parameter by maximum likelihood
  - Then you have the full likelihood function and applying AIC is no problem.
  - In practice, this can be a bad idea (gamma regression) or not possible (over-dispersion in Poisson regression)
- You could estimate the dispersion parameter from the deviance
- You could also estimate the dispersion parameter from the more saturated model and use it in the likelihood functions for both models in AIC
  - In linear regression case, this is called "Mallows' Cp"

Two Examples, n=50

- A: Straight parabola with normal errors:

$Y \sim X^2 + N(0,\sigma=1/4)$

where X is drawn from U([-2.5,2.5])

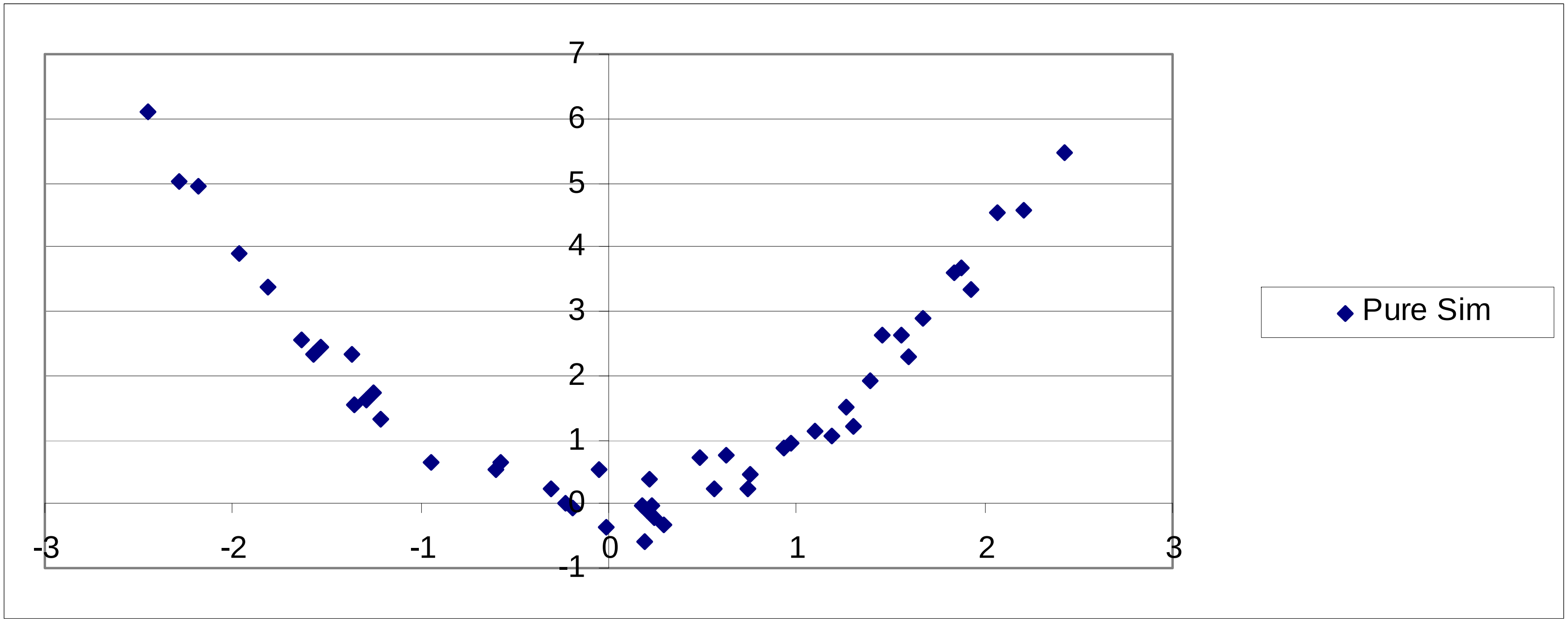
- B: With measurement error in X and a boundary on Y

$Y \sim \max(0,Q^2 + N(0,\sigma=1/4))$

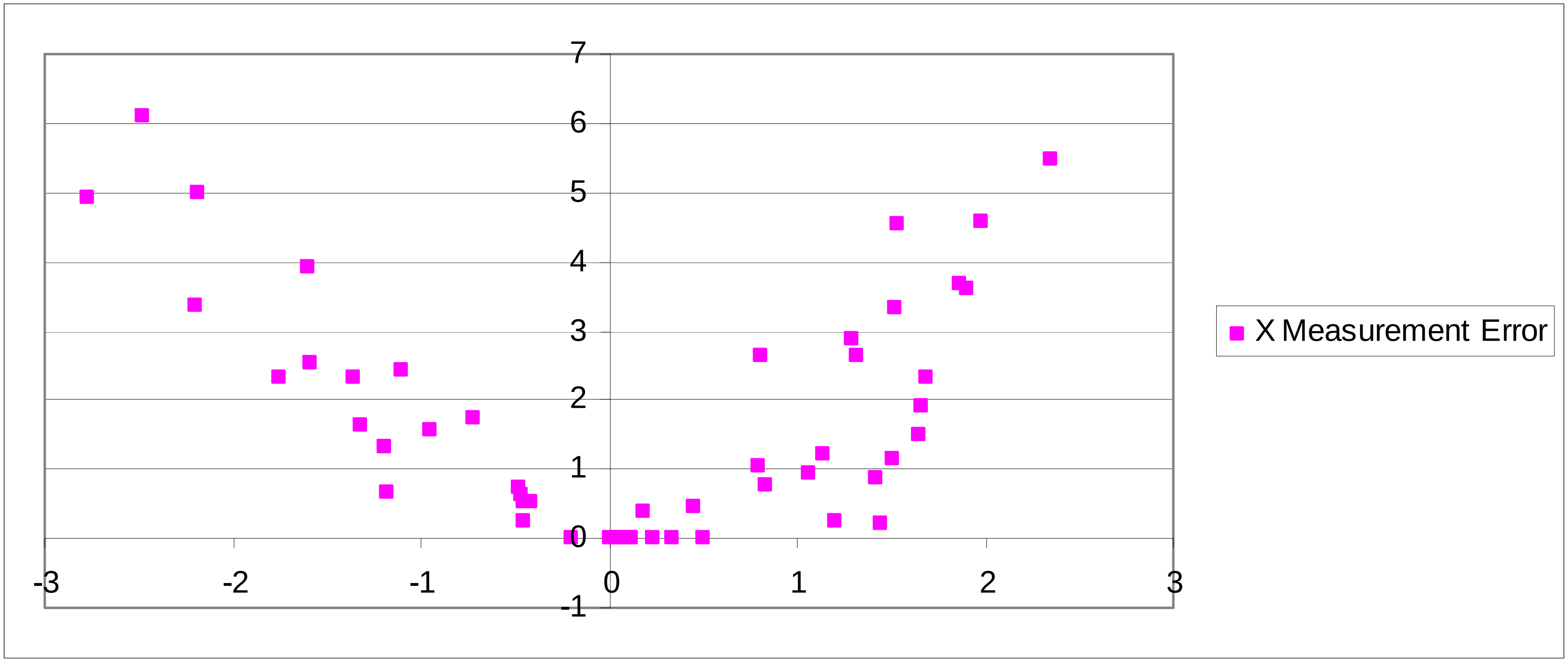
$X \sim Q + N(0, \sigma=0.35)$ where Q is drawn from U([-2.5,2.5])

- Try fitting linear regression on intercept, x^2
- And same regression with an x^4 term…test which is better

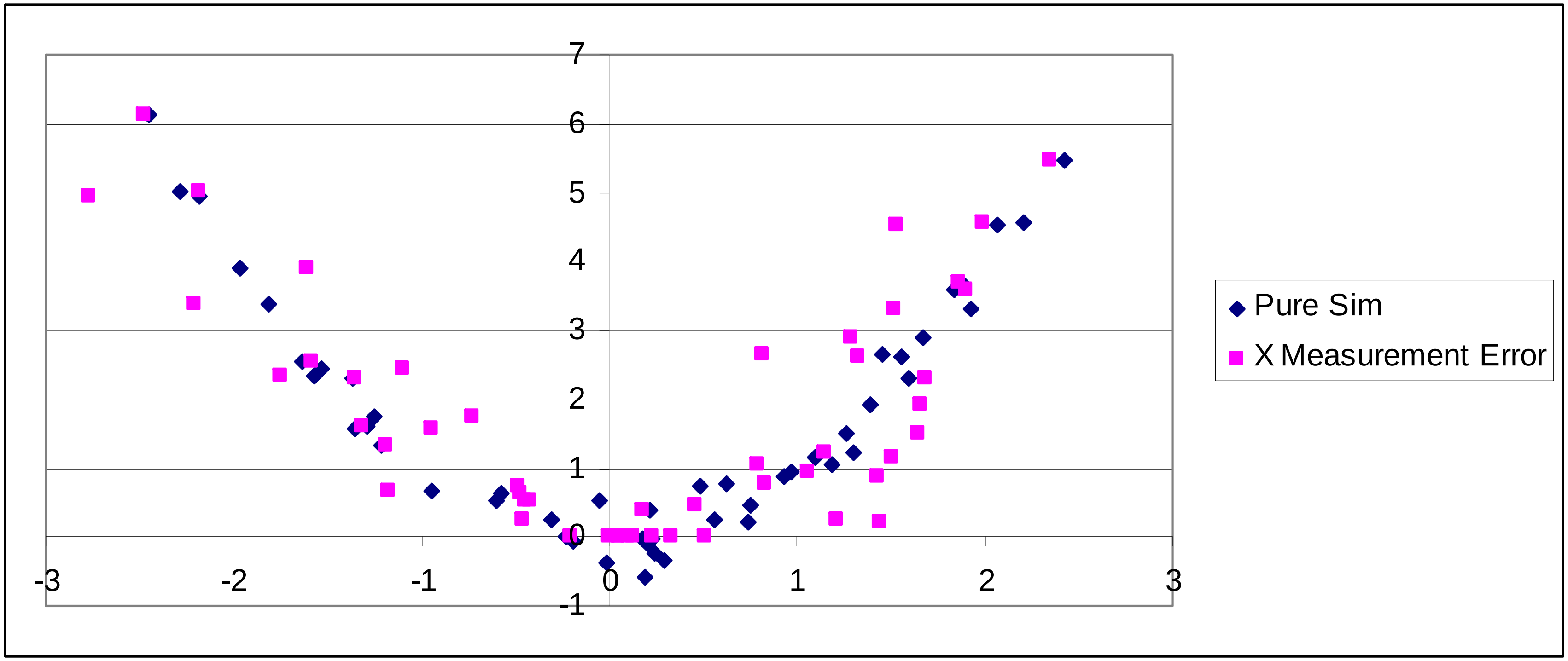
First Sample of Straight Simulation



First Sample of with X Measurement Error



First Sample of Each



We know we don’t need x^4, but…

How often do we end up including it in our model?

Generated 100 samples each with 50 observations

Bootstrap is based on 100 resamples

Tests (except AIC) are nominally at the 5% level

% of time x^4 included	Pure Sim	With Meas Errors
AIC	20	80
LRT	9	66
Wald χ^2	10	67
Bootstrap	6	47

Useful References

- Hastie, Tibshirani, and Friedman, *Elements of Statistical Learning* (lots of stuff on out-of-sample testing)
- Efron, *Large-Scale Inference* (for False Discovery Rate)
- Davison and Hinckley, *Bootstrap Methods and Their Applications*
- Bühlmann and van de Geer, *Statistics for High-Dimensional Data* (for Lasso)
- Bühlmann and Gisler, *A Course in Credibility Theory and Its Applications* (for credibility)
- Hastie and Tibshirani, *Generalized Additive Models* (for smoothing splines)