Quantitative Social Science Methods, I, Lecture Notes: Model Evaluation

Gary King[1]
Institute for Quantitative Social Science
Harvard University

August 17, 2020

[1] GaryKing.org

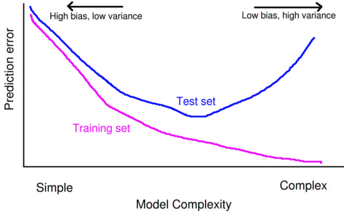
How Do You Know Which Model is Better?

Robust Standard Errors

A Better Way to Use Robust SEs: An Application

Evaluation by Out-of-Sample Forecast

- Your job: find the underlying (persistent) structure, not idiosyncratic features in your data
- Partition data: between training and test sets
- Fit model to training set; predict test set
- Compare to truth for average prediction and full distribution
- E.g.: for $\widehat{\Pr(y=1)} = 0.2$, 20% in test set should be 1s
- Best test sets: truly out-of-sample
- If world changes: model may fail anyway



See Trevor Hastie et al. 2001. *The Elements of Statistical Learning*, Springer, Fig 7.1

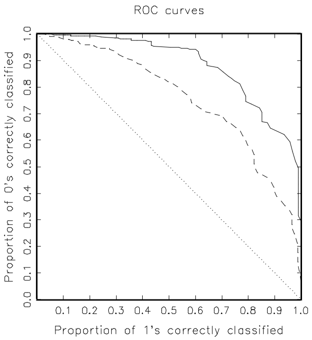
Other Methods of Evaluating Models

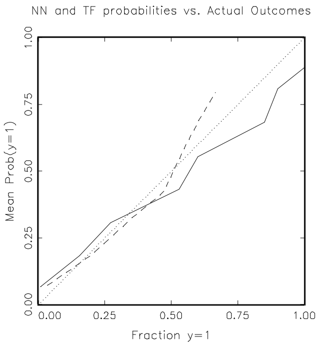
- Cross-validation
  - The idea: randomly select k observations as the "test set"; evaluate with rest of data as training; set aside another set of k observations; evaluate; Repeat; report performance averaged over subsets
  - Useful for small data sets; out-of-sample test sets are better
- Fit, in general: Look for all possible observable implications of a model, and compare to observations. (Think. Be creative here!)
- Fit: continuous variables
  - The usual regression diagnostics
  - E.G., plots of $e = y - \hat{y}$ by X, Y or $\hat{y}$
  - Check more than the means. E.g., plot e by $\hat{y}$ and draw a line at 0 and at $\pm 1, 2$ se's. 66%, 95% of the observations should fall between the lines.
  - For graphics:
    - transform bounded variables
    - transform heteroskedastic results
    - highlight key results; label everything

Binary Model Evaluation by ROC Curves

- Binary predictions require a normative decision
    - C: number of times more costly misclassifying 1 than 0
    - C must be chosen independently of the data
    - Justification: philosophy, policymaker survey, literature,...
    - Decision theory: choose $Y = 1$ when $\hat{\pi} > 1/(1 + C)$; 0 otherwise
        - If $C = 1$, predict $y = 1$ when $\hat{\pi} > 0.5$
        - If $C = 2$, predict $y = 1$ when $\hat{\pi} > 1/3$
    - Compute: (a) % of 1s correctly predicted; (b) % of 0s correctly predicted; (c) patterns in errors in different forecasts
- ROC (receiver-operator characteristic) curves
    - For every possible C: Compute %1s and %0s correctly predicted
    - Plot %1s by %0s correctly predicted
    - Overlay curves for several models
    - If one curve is above another the whole way, then it dominates, no matter your normative decision (about C)
    - Otherwise, one model is better than the other in only given specificed ranges of C

ROC Curve In Sample

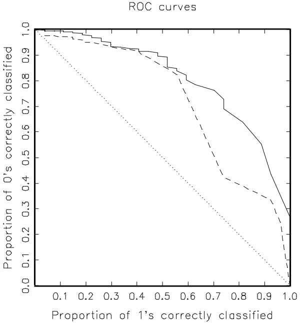


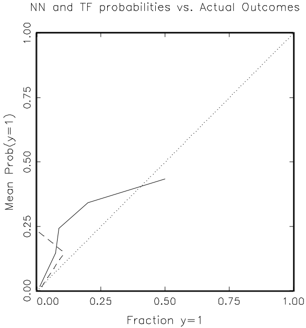


In-sample ROC on left

(from Gary King and Langche Zeng. "Improving Forecasts of State Failure," *World Politics*, 2001)

ROC Curve Out-of-Sample

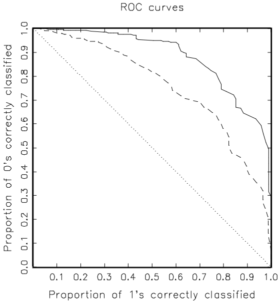


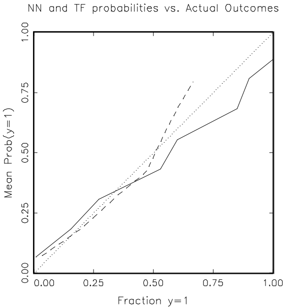


Out-of-sample ROC on left

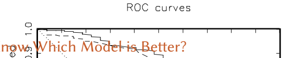
Calibration for Evaluating Binary Models

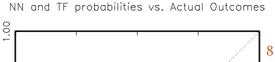
- Sort estimated probabilities into bins of say 0.1 width: $[0, 0.1)$, $[0.1, 0.2)$,..., $[0.9, 1]$.
- In each bin, compute: (a) mean predictions ($\approx$ 0.05, 0.15, etc.) and (b) the average fraction of 1s
- Plot (a) by (b) and look for systematic deviation from 45° line





In-sample calibration graph on right





How Do You Know Which Model is Better?

Robust Standard Errors

A Better Way to Use Robust SEs: An Application

Robust SEs

- Widely (mis)used: $> 141,000$ cites, $+ \approx 1000$/month
- Are: a way to estimate $V(\hat{\theta})$ with fewer assumptions
- Are Not: A way to estimate $V(\hat{\theta})$ without *any* assumptions
- Are: useful when only *some* assumptions are violated
- Are Not: A way to inoculate yourself from criticism
- When SEs and RSEs differ:
  - Best case
    - Some QOIs (β in regression): unbiased but inefficient
    - Other QOIs ($\Pr(Y > 0.6)$): biased
  - Worst case
    - Misspecification is more widespread
    - All QOIs are biased
- Are: A good test for misspecification

Regression Model Variance Specification

- Linear-normal regression model:
  1. $Y_i \sim N(\mu_i, \sigma^2)$ (systematic component)
  2. $\mu_i = X_i\beta$ (stochastic component)
  3. $Y_i \perp Y_j \mid X, \forall i \neq j$ (independence assumption)
- Equivalently: $\underset{n\times 1}{Y} \sim N(\underset{n\times k}{X}\underset{k\times 1}{\beta}, \underset{n\times n}{\Sigma})$, where $\Sigma = \sigma^2 I$; that is:

$$\Sigma = \begin{pmatrix} \sigma^2_{11} & \sigma^2_{12} & \cdots & \sigma^2_{1n} \\ \sigma^2_{21} & \sigma^2_{22} & \cdots & \sigma^2_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2_{n1} & \sigma^2_{n2} & \cdots & \sigma^2_{nn} \end{pmatrix} \begin{pmatrix} \sigma^2_{11} & \sigma^2_{12} & \cdots & \sigma^2_{1n} \\ \sigma^2_{21} & \sigma^2_{22} & \cdots & \sigma^2_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2_{n1} & \sigma^2_{n2} & \cdots & \sigma^2_{nn} \end{pmatrix} \begin{pmatrix} \sigma^2_{11} & 0 & \cdots \\ 0 & \sigma^2_{22} & \cdots \\ \vdots & \vdots & \ddots \\ 0 & 0 & \cdots \end{pmatrix}$$

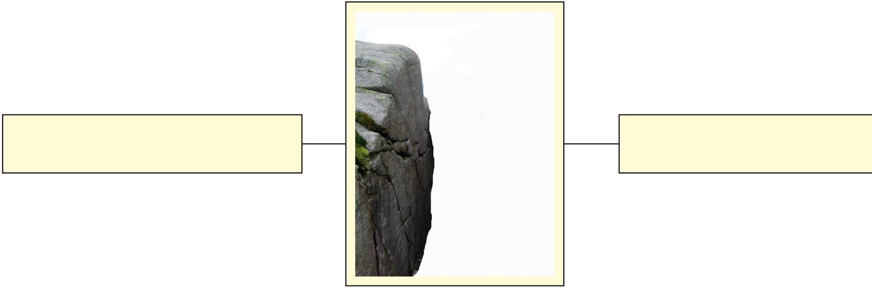
Variance matrix unconstrained Variance matrix unconstrained (includes covariances) Now assume independence (set covariances to zero) Still allows for heteroskedasticity Now assume independence and homoskedasticity Standard linear-normal regression model assumptions

What if $V(Y) \equiv \Sigma \neq \sigma^2 I$ and we run a regression?

- Coefficients $b = Q^{-1}X'y$ (with $Q = X'X$) unbiased: $E(b) = E(Q^{-1}X'y) = Q^{-1}X'E(y) = Q^{-1}X'X\beta = \beta$
- True variance: $V(b) = V(Q^{-1}X'y) = Q^{-1}X'V(y)XQ^{-1} = Q^{-1}X'\Sigma XQ^{-1} \neq \sigma^2 Q^{-1}$
- Usual estimate of $V(b) = \sigma^2 Q^{-1}$: biased
- Model sims are wrong! $\rightsquigarrow$ other QOIs are biased
- Estimating all the unknowns in Σ: seems hopeless
- Key Insight: Need to estimate fewer parameters than it seems
  - $V(b) = \underset{k\times k}{Q^{-1}} \underbrace{\underset{k\times n}{X'} \underset{(n\times n)}{\Sigma} \underset{n\times k}{X}}_{k\times k} \underset{k\times k}{Q^{-1}} = Q^{-1} \underset{k\times k}{G} Q^{-1}$, with $k \ll n$
  - Can estimate $V(b)$ by replacing σ_i^2 with e_i^2 in Σ
  - Generalizes to any MLE model
- Result: RSEs are statistically consistent
- But: The model, sims, QOI estimates are still wrong

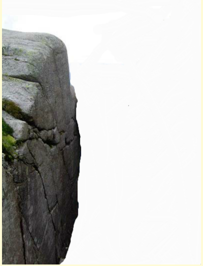
For RSEs to help: Everything has to be Juuuussttt Right

For RSEs to help: Everything has to be Juuuussttt Right



For RSEs to help: Everything has to be Juuuussttt Right

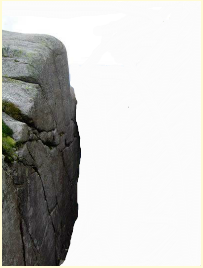
Model Misspecified



For RSEs to help: Everything has to be Juuuussttt Right

Model Misspecified

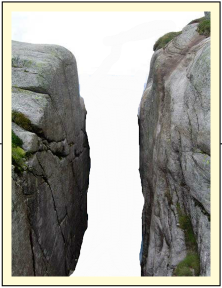
(point estimates biased)



For RSEs to help: Everything has to be Juuuussttt Right

Model Misspecified

(point estimates biased)

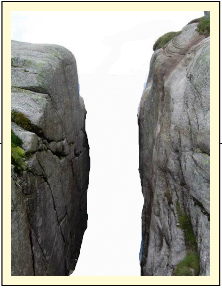


Model Correct

For RSEs to help: Everything has to be Juuuussttt Right

Model Misspecified

(point estimates biased)



Model Correct

(point estimates corre

For RSEs to help: Everything has to be Juuuussttt Right

Goldilocks Region

Model Misspecified

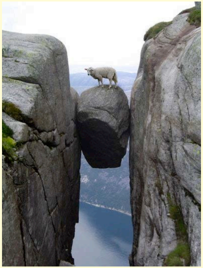
(point estimates biased)

Model Correct

(point estimates corre

For RSEs to help: Everything has to be Juuuussttt Right

Goldilocks Region



Model Misspecified

(point estimates biased)

Model Correct

(point estimates corre

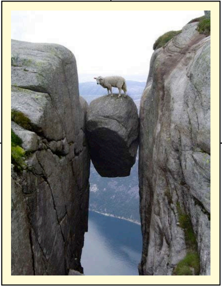
For RSEs to help: Everything has to be Juuuussttt Right

Biased just enough to make RSEs useful,

Goldilocks Region

Model Misspecified

(point estimates biased)



Model Correct

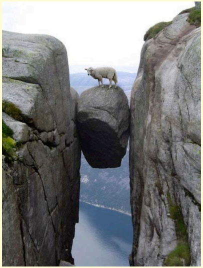
(point estimates corre

For RSEs to help: Everything has to be Juuuussttt Right

Biased just enough to make RSEs useful,

Goldilocks Region

but not so much as to bias everything else



Model Misspecified

(point estimates biased)

Model Correct

(point estimates corre

The Goldilocks Region is not Idyllic

- Only a few QOIs can be estimated
  - E.g., Y: Dem proportion of two-party vote
  - Can estimate: β
  - Can't estimate:
    - probability the Democrat wins
    - variation in the vote outcome
    - vote predictions with CIs
  - We don't know: the substantive meaning of our results. How big are they, really?
  - We can't check: whether model implications are realistic
- Parts of the model are wrong:
  why do we think the rest is right?
- If SE≠RSE, we should:
  find missspecification, fix model,
  rerun until SE=RSE

How Do You Know Which Model is Better?

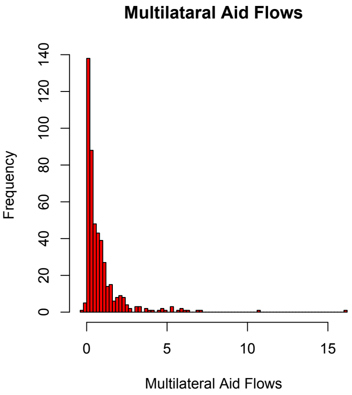
Robust Standard Errors

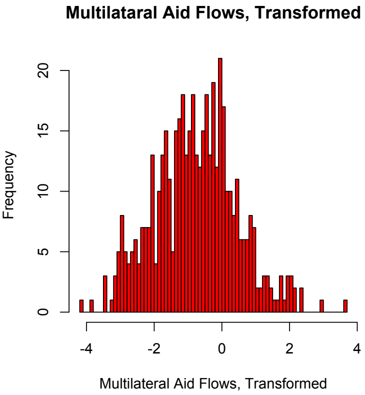
A Better Way to Use Robust SEs: An Application

Replication of Neumayer (ISQ, 2003)

- Claim: "Multilateral aid flows…exhibit a bias toward less populous countries."
- Method: Linear regression of multilateral aid flows on log-population, squared log-population, and control variables
- Result: Coefficient on log-population: -3.13
- Replication: Robust SE (0.72) twice classical SE (0.37)
  $\rightsquigarrow$ Clear evidence of misspecification
- Correction: Apply Box-Cox transformation of Y (like a log)
- New test: Robust SE (0.34) $\approx$ classical SE (0.32)
  $\rightsquigarrow$ No evidence of misspecification

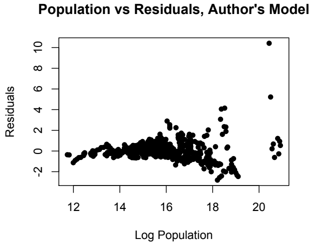
Box-Cox Transformation makes Y Normal

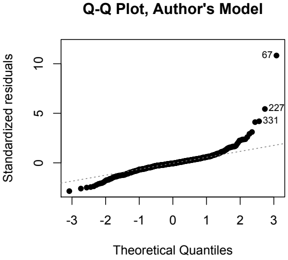


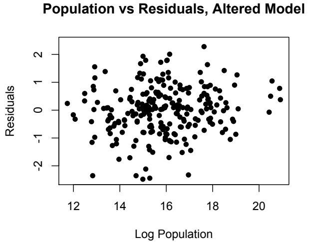


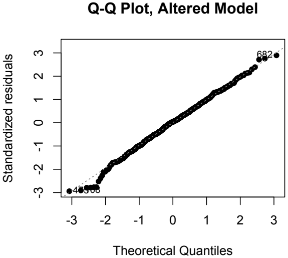
Quiz: Is this a good test of the new specification?

Transformation of Y: Removes Heteroskedasticty



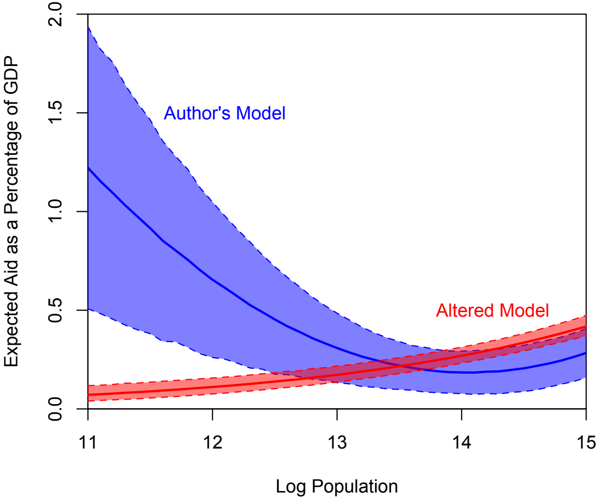






Results: Transformed Model, Opposite Results

Misspecification $\rightsquigarrow$ bias



Quiz: Why are the CIs smaller for the correctly specified model?