Improving Logistic Regression Analyses

Center for Health Care Research & Policy - First Methods Seminar

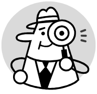
Thomas E. Love, PhD
TEL3@po.cwru.edu
August 31, 2001

Center for
Health Care
Research &
Policy

General Goals of Methods Seminar Series

- Discuss a variety of methodological issues of importance to health services researchers.
- Input: presenter to pose some key issues, discussion to identify current practices, interesting examples, gaps in understanding
- Output: short discussion of important issues, bibliography, material for grant proposals, teaching, etc. Web? Technical Report Series?

Goals of today's session



- Discuss two practical issues that often arise in using logistic regression models:
  1. How should we **check assumptions**? Diagnostics / consideration of goodness of fit and of outliers – identifying potential problems
  2. How should we develop **parsimonious models** for our response? Variable selection, best subsets vs. stepwise procedures, C_p-style criteria.
- When should we worry about all of this?

Part 1 – Diagnostics & Outliers

- Why/when should I care about outliers?
- Diagnostics as outlier identification
- Methods for studying outliers (residual, leverage and influence)
- What is current common practice?
- What do we know now, and can we do things more effectively?
- What are the open issues?

Hosmer, DV et al. (1991) The Importance of Assessing the Fit of Logistic Regression Models: A Case Study

- Demonstrates methods for assessing the fit, adverse consequences of failing to do so.
- Defines GOF assessment in two stages
  - Providing a summary measure of the errors
  - Examining the individual values of the errors
- Amer J Public Health (1985-1989)

1985-1989	#	Using Log Reg	GOF assessed
Articles	579	113 (20%)	6 (5%)
Briefs	379	23 (6%)	1 (4%)

Summary Measures of Goodness-of-Fit

- How well does the model fit the available data?
  - P values of individual coefficients
  - More than a dozen "R^2"-type summaries
  - Deviance and Pearson χ^2 statistics
  - Hosmer-Lemeshow Goodness-of-fit tests
  - Classification tables
  - Area under the ROC Curve

Diagnostic Tools for Logistic Regression Models

- Role of individual subjects in the model
- Measures of residual
- Measures of leverage
- Measures of influence
- Standards for evaluation
- Useful graphical analyses
- How to combine these measures?
- These tools are easy to get at in software.

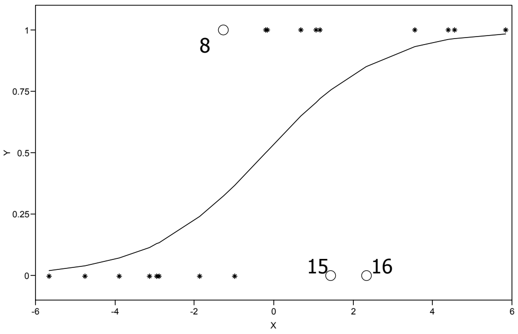
"Typical" Data Table

(Hosmer & Lemeshow, Table 5.13)

ID	x	y0
1	-5.65	0
2	-4.75	0
3	-3.89	0
4	-3.12	0
5	-2.93	0
6	-2.87	0
7	-1.85	0
8	-1.25	1
9	-0.97	0
10	-0.19	1

ID	x	y0
11	-0.15	1
12	0.69	1
13	1.07	1
14	1.18	1
15	1.45	0
16	2.33	0
17	3.57	1
18	4.41	1
19	4.57	1
20	5.85	1

"Typical" Data with Model



"Typical" Logistic Regression

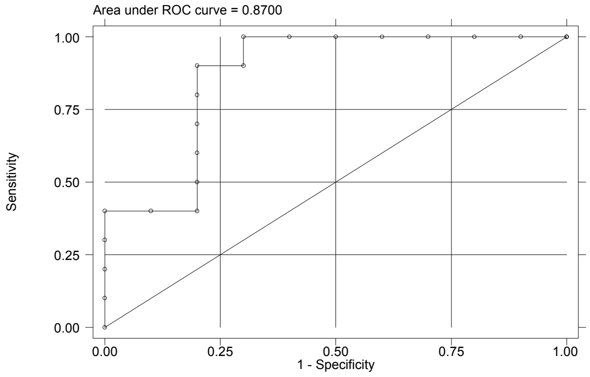
	Coef.	Std. Err.	Z	P>|z|
X	.693	.301	2.30	.002
Cons	.139	.614	0.23	.820

LR χ^2 = 10.98 on 1 DF, p = .001, OR = 1.999

Goodness of Fit Assessment

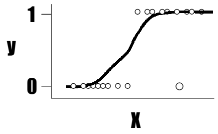
- Pearson χ^2 = 15.93 on 18 df, p = .597
- Hosmer-Lemeshow test χ^2 = 11.77 (8 df), p = .162
- Pseudo-R^2 is .396
- Classifies 8/10 "zeros" and 9/10 "ones" correctly.

"Typical" ROC Analysis



Logistic Regression Residuals

- A particular subject might be called to our attention because the model-based prediction of its outcome is not very close to that which is observed.
- Does this case have a large residual?



Good Model?

Measuring differences between observed and fitted values?

Pearson Residuals (Residual in Stata)

$$r(y_j, \hat{\pi}_j) = \frac{y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}$$

- Pearson χ^2 is the SS of these residuals:

$$X^2 = \sum_{j=1}^{J} r(y_j, \hat{\pi}_j)^2$$

Measuring differences between observed and fitted values?

Deviance Residuals (Deviance in Stata)

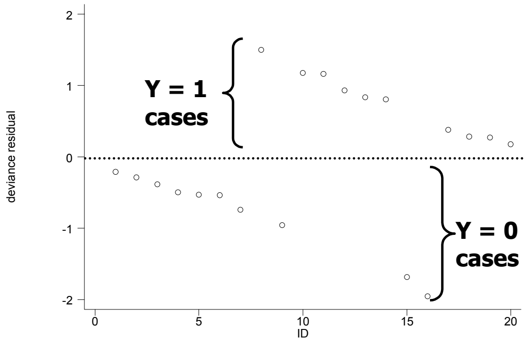
$$d(y_j, \hat{\pi}_j) = \pm\left\{2\left[y_j \ln\left(\frac{y_j}{m_j\hat{\pi}_j}\right) + (m_j - y_j)\ln\left(\frac{m_j - y_j}{m_j(1-\hat{\pi}_j)}\right)\right]\right\}$$

where the + or – is the same as the sign of $(y_j - m_j\hat{\pi}_j)$

- Deviance is the SS of these residuals:

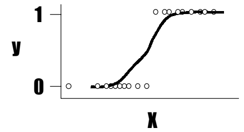
$$D = \sum_{j=1}^{J} d(y_j, \hat{\pi}_j)^2$$

Index Plot of Deviance Residuals for "Typical" Data



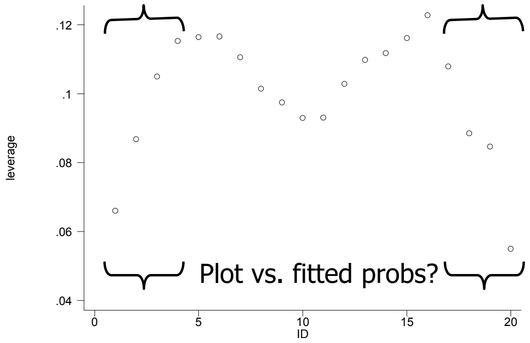
Measuring Leverage

- A subject may have a configuration of independent variable values that is especially unusual relative to the rest of the subjects.
- Leverage measures this "X-outlierness"

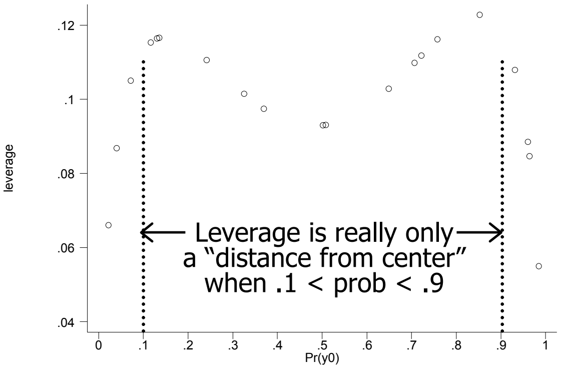


Good Model?

Index Plot of Leverages for "Typical" Data Set



"Typical" Data: Leverages vs. Fitted



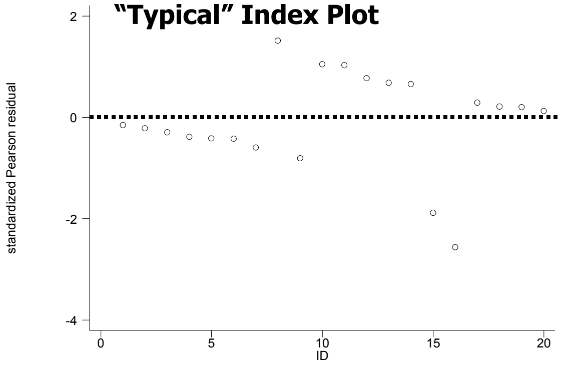
Standardized Pearson Residuals

- It turns out that Pearson residuals do not have variance equal to 1, unless we look at standardized Pearson residuals:

$$r_{sj} = \frac{r_j}{\sqrt{1-h_j}}$$

(Rstandard in Stata)

Standardized Pearson Residuals



Measuring Influence

- We'd like to identify subjects who have a large effect on the coefficients of the fitted model – if we removed this case, how would this change the model?
- If multiple cases have the same **x** pattern, we'll delete them all simultaneously here.
- The contribution of a single observation depends on both its residual and leverage – so will our measures.

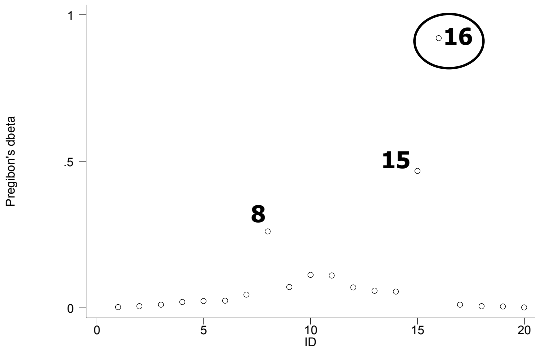
Influence on the Coefficients

- Delta-Beta Influence Statistic (Stata: Dbeta)
  - Standardized difference between estimated coefficient vector (β) using all cases and estimated β using all cases except those with covariates matching this observation.

$$\Delta\hat{\beta}_j \approx \frac{r_j^2 h_j}{\left(1-h_j\right)^2} = \frac{r_{sj}^2 h_j}{1-h_j}$$

Large values of this statistic identify covariate patterns that have large influence on the parameters.

"Typical" Data: Index Plot of $\Delta\hat{\beta}$



Logistic Regression Model using All 20 "Typical" Cases

	Coef.	Std. Err.	Z	P>|z|
X	.693	.301	2.30	.002
Cons	.139	.614	0.23	.820

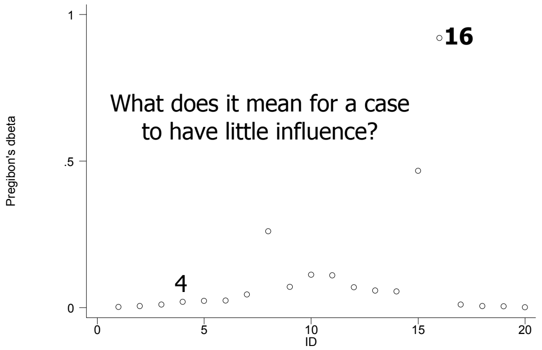
LR χ^2 = 10.98 on 1 DF, p = .001, OR = 1.999

Same Model without Case 16

	Coef.	Std. Err.	Z	P>|z|
X	1.065	.494	2.16	.031
Cons	.726	.802	0.90	.366

LR χ^2 = 14.42 on 1 DF, p = .0001, OR = 2.902

"Typical" Data: Index Plot of $\Delta\hat{\boldsymbol{\beta}}$



What does it mean for a case to have little influence?

Logistic Regression Model using All 20 "Typical" Cases

	Coef.	Std. Err.	Z	P>|z|
X	.693	.301	2.30	.002
Cons	.139	.614	0.23	.820

LR χ^2 = 10.98 on 1 DF, p = .001, OR = 1.999

Same Model without Case 4

	Coef.	Std. Err.	Z	P>|z|
X	.660	.300	2.20	.027
Cons	.179	.614	0.29	.771

LR χ^2 = 9.80 on 1 DF, p = .002, OR = 1.935

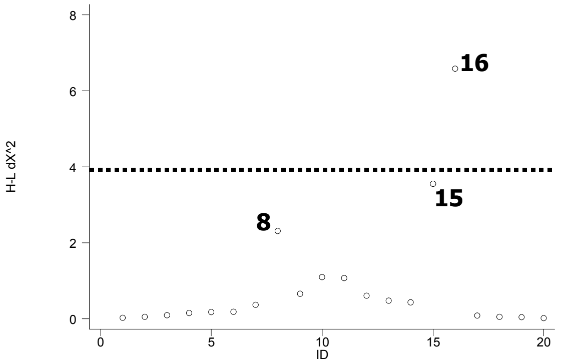
Influence on the Summary Statistics

- Delta-Chi-Square Statistic (Stata: Dx2)
  - Decrease in the value of the Pearson chi-square statistic due to the deletion of subjects with covariates $\mathbf{x}_j$

$$\Delta X_j^2 \approx \frac{r_j^2}{\left(1-h_j\right)} = r_{sj}^2$$

Large values of this statistic identify covariate patterns that are poorly fit.

"Typical" Data: Index Plot of ΔX^2



Influence on the Summary Statistics

- Delta-Deviance Statistic (Stata: Ddeviance)
  - Decrease in the value of the deviance statistic due to the deletion of subjects with covariates $\mathbf{x}_j$

$$\Delta D_j \approx d_j^2 + \frac{r_j^2 h_j}{\left(1-h_j\right)} \approx \frac{d_j^2}{1-h_j}$$

Large values of this statistic identify covariate patterns that are poorly fit.

Likely Values of the Diagnostics within Five Estimated Probability Regions

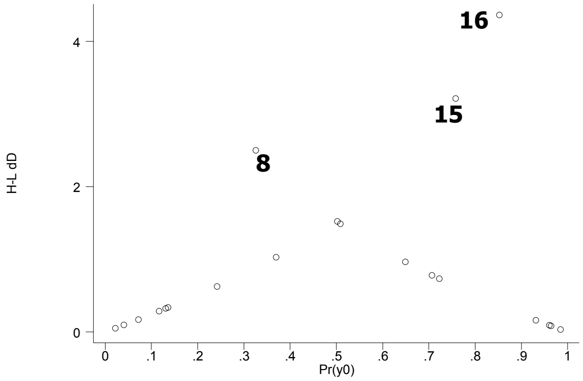
$\hat{\pi}$	ΔX^2	$\Delta\hat{\boldsymbol{\beta}}$	h
< .1	Large or Small	Small	Small
.1 to .3	Moderate	Large	Large
.3 to .7	Moderate to Small	Moderate	Moderate to Small
.7 to .9	Moderate	Large	Large
> .9	Large or Small	Small	Small

Core Plots for an Analysis of Diagnostics

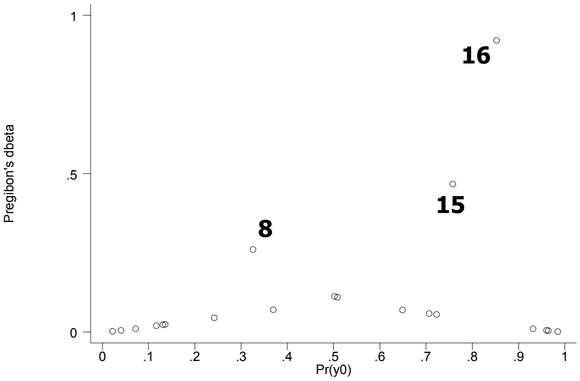
1. Plot ΔX_j^2 versus $\hat{\pi}_j$.
2. Plot ΔD_j versus $\hat{\pi}_j$.
3. Plot $\Delta\hat{\boldsymbol{\beta}}_j$ versus $\hat{\pi}_j$.
4. It's often useful to also plot these influence measures versus h_j, to directly assess the impact of leverage.
5. Plot ΔX_j^2 versus $\hat{\pi}_j$, sizing by $\Delta\hat{\boldsymbol{\beta}}_j$.

Hosmer & Lemeshow (2000) Applied Logistic Regression, 2nd ed., 176-177

"Typical" data: ΔD vs. fitted probs



"Typical" data: $\Delta\hat{\boldsymbol{\beta}}$ vs. fitted probs



How Do I Deal with More than One Outlying Observation?

- Sequential nature of usual outlier checks
- Should we delete pairs, or larger groups?
- If "all subsets" of observations cannot be feasibly studied, select m = $2k_{max}$ most suspect observations, then examine all subsets of these of size k or less.
- Eliminates much of the masking effect.

*Andrews & Pregibon (1978) JRSS B 40(1): 85-93.

Why aren't these ideas in common practice?

- Is it tough to compute this stuff?
- How have people presented these ideas in the context of a larger paper?
- Is it ever inappropriate to consider the impact of outlying values on the model?
- Are there situations where one doesn't care about robustness in this sense?

Lowenstein, DH et al. (2001) The Prehospital Treatment of Status Encephalitis (PHTSE) Study: Design and Methodology*

- Description of NIH – R01 to UCSF
- RCT using logistic reg. to estimate ttt effects and adjust for covariates
- "The fit of the logistic models will be assessed with the H-L GOF test and regression diagnostics. If there is a substantial lack of fit, techniques such as transformations of covariates will be used to improve the fit."

*Control Clin Trials 22: 290-309.

Schellenberg, JRMA et al. (2001) Effect of large-scale social marketing of insecticide-treated nets on child survival in rural Tanzania*

- Case-control study – model used to estimate effects of treated nets allowing for matching variables with case or control status as the outcome
- Indiv. effectiveness defined as 100(1-OR).
- Model robustness assessed by use of $\Delta\beta$ influence statistics, and fit was checked using H-L χ^2.

*Lancet 357: 1241-1247.

Krumholz, HM et al. (1997) Thrombolytic Therapy for Eligible Elderly Patients with Acute Myocardial Infarction*

- Retrospective cohort study – correlates of thrombolytic therapy use in elderly Medicare pts hospitalized with acute MI.
- STATA used throughout, Stepwise fit then…
  - Partial residual plots to evaluate potential problematic areas of fit in all models.
  - Goodness of fit χ^2 within deciles of probability.
  - Area under the ROC curve to evaluate discriminating power of each model.

*JAMA 277(21): 1683-1688.

Signorini, DF et al. (1999) Predicting survival using simple clinical variables: A case study in traumatic brain injury*

- Details univ and multiv analyses – final model contained age, GCS score, ISS, pupil score, and presence of haematoma on CT – sequential model ("forward selection")
- Two pts with influence 50% higher than body of the data, but retained.
- 3 pts with enormous residuals (died with predicted surv prob > .96) – "encouraging"
- Cross-valid. dropped ROC area .901 to .835

*J Neurol Neurosurg Psychiatry 66: 20-25.

When should / shouldn't I care about outliers?

- What is the purpose of your model?
  - Describe the nature of a relationship between X's and a binary response?
  - Obtain a predicted probability (propensity) for Y given a series of X's?
  - Adjust for confounding factors in a study?
- If a point has almost no influence on the results, there is little point in agonizing over how deviant it appears.

Diagnostics & Outliers Summary

- H&L GOF, ROC, etc. all provide some summary assessment of fit quality
- Residuals, Leverage and Influence measures all of value in studying individual covariate patterns.
- Set of generally recommended plots.
- Why/when should I care about outliers?
- Common practice? Open Issues?

Part 2: Choosing The "Best" Model

Parsimony and Variable Selection

Main Reference: Hosmer, DW & Lemeshow, S (2000) Applied Logistic Regression, 2nd ed., Wiley, Chapter 4.

Part 2 – Variable Selection

- Why/when should I care about parsimony?
- C_p and other summary statistics
- Best subsets methods for logistic regression – combating some stepwise flaws
- When should I automate the search?
- What is current common practice?
- How about external validation of the model?
- What are the open issues?

Motivations for Parsimony

- "Everything should be made as simple as possible, but no simpler." Einstein
- "All models are wrong, some models are useful." John Tukey
  - Select the variables that result in a "best" model within the problem's scientific context.
  - Results more likely to be numerically stable (avoid "overfitting") & easily generalized.
  - Smaller estimated standard errors
  - Less dependence of the model on these data
  - KISS

A Roadmap to Model-Building

Based on Hosmer & Lemeshow, pp. 92-99

1. Careful univariable analysis of each variable, using contingency tables and scatterplot smooths.
2. Select variables for the multivariable model – including all with univariable p value < 0.25 and all of clinical importance. Start with a "kitchen sink" model containing all of these.

A Roadmap to Model-Building

3. It could be that a collection of variables is useful where individually they appear unimportant. C_p-based selection methods help identify the "best" model in these settings.
4. Check measure assumptions – are discrete categories appropriate, are continuous variables linear in the logit?

A Roadmap to Model-Building

5. Check for interactions among the variables in the model. Should base inclusion on statistical and practical considerations. The result is the preliminary final model.
6. Check goodness of fit, and model assumption adequacy (as in part 1).

Stepwise Variable Selection for Prediction in LINEAR Regression

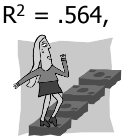
- Response: Monthly value-weighted returns for the 48 months in 1990-1993.
- Goal: Try to predict the returns for the 12 months of 1994.
- Ten predictors (X1 through X10) are available, in a response surface design so we have 65 different predictors, including squares and cross-products.

Foster, Stine & Waterman Business Analysis Using Regression, Springer.

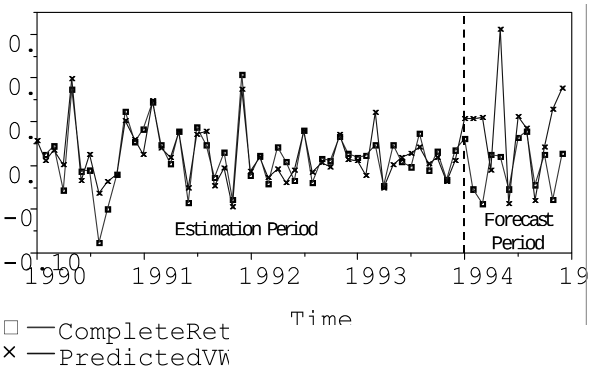
Final Preliminary Model: Start with Forward Selection (.25) then Backward Elimination (.05)

```
Parameter Estimat
Term Est[illegible]
Intercept [illegible]
X1   0.023[illegible]
X2   0.006[illegible]
X3  -0.00[illegible]
X5  -0.00[illegible]
X6   0.003[illegible]
X7   0.003[illegible]
X8   0.008[illegible]
X9  -0.00[illegible]
X10 -0.00[illegible]
X1*X[illegible]
X6*X[illegible]
X6*X[illegible]
X6*X[illegible]
X6*X[illegible]
X7*X[illegible]
X7*X[illegible]
X8*X[illegible]
X9*X[illegible]
X9*X[illegible]
X10*X[illegible]
X10*X[illegible]
```

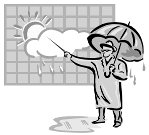
- All remaining terms are either significant ($p < .05$) or are part of a sig. interaction.
- Overall ANOVA F test $p = .0009$
- $R^2 = .763$, Adj $R^2 = .564$, RMSE = .023
- Conclusion?

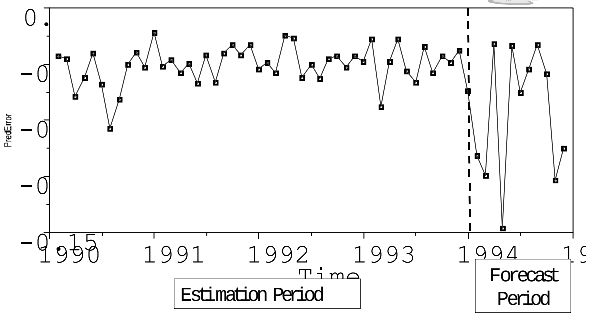


How Well Does this Predict?

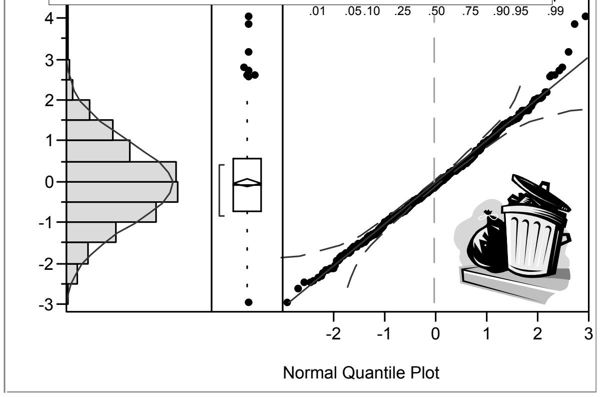


Prediction Errors over Time





The Predictors: X1 … X10



Stepwise Model-Building

- Can be helpful in suggesting possible models -- but does require thinking and judgment for proper use.
- Stepwise regression introduces biases, tends to over-estimate pseudo-R^2, ROC -- χ^2 statistics tend to be too large
- No guarantee that final model chosen is the "best" available choice by any criterion.

Armstrong, RW et al. (1998) Nasopharyngeal Carcinoma in Malaysian Chinese: Salted Fish and other Dietary Exposures*

- Case-control study: association of individual dietary components with NPC, adjusted for other dietary components
- Stepwise logistic reg. model selection
- Uses diagnostics as a post-selection check
  - "Distributions of X's among cases & controls"
  - "Model summary statistics and diagnostic plots"

*Int J Cancer 77: 228-235.

Price, KJ et al. (1998) Prognostic Indicators for Blood and Marrow Transplant Pts Admitted to an Intensive Care Unit

- Prospective study of 115 HSCT (hematopoietic stem cell transplantation) pts to identify correlates of survival.
- Used partial residual plots with LOWESS on simple logistic reg models first to determine transformations (S+ and StatXact)
- Then did stepwise BE to obtain two models (depending on inclusion of intubation)

*Am J Respir Crit Care Med 158: 876-884.

Best Subsets Procedure (without the equations)

- Obtain fitted values (estimated probabilities) for each subject based on the logistic regression of y (0/1) on x.
- Calculate the values z and v for each subject from the actual 0/1 values and fitted probabilities for each case. (just arithmetic)
- Now run a best subsets linear regression, with z as the dependent variable, case weights v and covariates x.

Criteria for "best" subsets

- R^2 and adjusted R^2 (don't do this).
- Mallows' C_p. Subset of q of p variables,

$$C_q = \frac{X^2 + \text{Wald}}{X^2/(n-p-1)} + 2(q+1) - n$$

- Models with C_q near q+1 are good choices.
- Best subsets picks those subsets with smallest C_q.
- Need to adjust SAS PROC LOGISTIC output (see Hosmer & Lemeshow, pp. 133-134)

Cross-Validation as a part of Logistic Regression Model-Building

- Suppose I have a response I want to fit to a predictor set, then validate the choice of model.
- How many observations (or what %) should I withhold?
- Should I withhold at random?
- What should I identify as an outlier before withholding?

What's the Right Order?

- What is the impact of removing outliers on stepwise logistic regression?
- On best subsets logistic regression?
- Can we implement new knowledge?
- What's the more logical order?
  - Outlier check, then (automated) paring down of variables, or vice versa?

3 More Krumholz HM et al Retrospective Cohort Studies using Logistic Regression Diagnostics and ROC after Stepwise Model Selection

- Krumholz, HM et al. (1998) Trends in the Quality of Care for Medicare Beneficiaries Admitted to the Hospital with Unstable Angina J Am Coll Cardiol 31(5): 957-963.
- Hierarchical logistic reg models found in:
  - Krumholz, HM et al. (1998) Prognostic Importance of Emotional Support for Elderly Patients Hospitalized with Heart Failure Circulation 97: 958-964. (Model selection earlier in the process.)
  - Vaccarino, V et al. (1998) Sex Differences in Mortality after Myocardial Infarction Arch Intern Med 158: 2054-2062. (Model selection came last.)

Normand, S-L et al. (1996) Using Admission Characteristics to Predict Short-Term Mortality From MI in Elderly Patients*

- Cohort study to develop a prediction model of 30d mortality to permit use of risk-adjusted rates as hospital quality measures
- Extensive discussion of perils of extreme values, missing data (assumed M at random).
- 3-phase process for model selection including multiple cross-validations and Monte Carlo procedure (with stepwise components)

*JAMA 275: 1322-1328.

A Model Isn't Useful Unless…

- It serves the purpose for which it was intended.
- It fits the available data reasonably well.
- It shows evidence of avoiding flukes and random behavior -- it displays statistical significance.
- It can be explained and assessed by you.
- It performs well when asked to predict results for new data.

Where Are We Now?

American Journal of Public Health
March - July 2001

	#	Logistic Regression Used	Goodness of Fit Assessed (beyond coefficient p values)	Model Selection of any form presented
Articles	43	22 (52%)	4 (18%)	9 (41%)
Briefs	37	16 (43%)	0 (0%)	4 (25%)

When should / shouldn't I care about variable selection?

- What is the purpose of your model?
  - Logistic regression to describe the nature of a relationship between a series of X's and a binary response?
  - Logistic regression to obtain a predicted probability (propensity) for Y given a series of X's?
  - Logistic regression to adjust for confounding factors in a study?

My "Biases"

1. In developing propensity models, variable selection (or even significance) isn't important, but you should care about the effect of outliers as they affect your scales.
2. Where you're looking at whether a treatment has an effect adjusting for covariates, you should care more about outliers and variable selection than is common practice.

Next Methods Seminar September 28

- Propensity Models – Charles Thomas [and others] (in part, this is to foreshadow our short course at the SMDM in San Diego)
- Later this Fall: Hierarchical Models? Power? SEM? Bootstrap? IRT?
- Suggestions to me (TEL3@po.cwru.edu) or Neal Dawson (nvd@po.cwru.edu)