Data Mining
with Regression

Bob Stine
Dept of Statistics, Wharton School
University of Pennsylvania
Some Details

- Office hours
  - Let me know and we can meet at Newberry
  - stine@wharton.upenn.edu

- Class notes
  - http://www-stat.wharton.upenn.edu/~stine/mich/

- Data
  - Will post ANES and others on Z drive

- JMP software
  - Depends on your school
Topics for Today

• Review from last time
  • Any questions, comments?

• Growing regression models
  • Deciding which variables improve a model
  • Standard errors and significance

• Missing data

• Stepwise regression
Why use regression?

• Claim
  • Regression is capable of matching the predictive performance of black-box models
  • Just a question of having the right X’s

• Regression is familiar
  • Recognize then fix problems
  • Shares problems with black-boxes
    Opportunity to appreciate what happens in less familiar, more complex models with more flexible structure.

• Familiarity allows improvements
  • Patches in Foster and Stine 2004
Review ANES Example

• Start with simple regr, expand to multiple
  • Post FT Obama on Pre FT Obama
  • Add ‘Happy/Sad’ and ‘Care Who Wins’
  • Include interaction effect

• Visual exploration of model form
  • Show the effects of an interaction
  • What’s the interaction mean

• Calibration
  • Being right on average

• Tests and inference
  • Which terms are significant? What’s that mean?

avg(\gamma|\hat{\gamma})=\hat{\gamma}
Modeling Question

• How do we expand a regression model
  • Reach beyond obvious variables
  • Find subtle but important features

• Automate typical manual procedure
  • Iterative improvement
  • Try variable, diagnose, try another, diagnose…

• Computing allows more expansive search
  • Open modeling process to allow a surprise
  • Example: Include interactions
    transformations, combinations (e.g. ratios), bundles (e.g. prin comp)
  • Magnified scope also magnifies problems
Medical Example

• Numerical response

• Diagnosing severity of osteoporosis
  • Brittle bones due to loss of calcium
  • Leads to fractures and subsequent complications
  • Personal interest

• Response
  • X-ray measurement of bone density
  • Standardized to $N(0,1)$ for normal
  • Possible to avoid expense of x-ray, triage?

• Explanatory variables
  • Data set designed by committee doctors, biochemists, epidemiologists
Osteoporosis Data

- Sample of postmenopausal women
  - 1,232 women with 127 columns
  - Nursing homes in NE... Dependence? Bias?
  - Presence of missing data
  - Measurement error

- Marginal distributions
  - X-ray scores (zHip), weight, age...

ideal data?
Initial Osteo Model

- Simple regression
  - zHip on which variable?
  - How would you decide…
- Impact of weight

Interpretation?

pick largest correlation
consult science

- Interpretation?

- Interpretation?
Expanding Model

- What to add next?
  - Residual analysis
  - Add others and see what sticks

- Add them all?
  - Singularities imply redundant combinations
  - Summary of fit
    Impressive $R^2$ until you look at the sample size.
Missing Data

• Fit changes when add variables
  • Collinearity among explanatory variables
  • Different subsets of cases

• What to do about the missing cases
  • Exclude
    “Listwise deletion”
    “Pairwise deletion”
  • Impute. Fill them in, perhaps several times

• Imputation relies on big assumption
  Missing cases resemble those included.
  Real data is seldom (if ever) missing at random
Handle Missing Data

- Add another variable
  - Add indicator column for missing values
  - Fill the missing value with average of those seen

- Simple, reduced assumption approach
  - Expands the domain of the feature search
  - Allows missing cases to behave differently
  - Conservative evaluation of variable

- Part of the modeling process
  - Distinguish missing subsets only if predictive

- Categorical: not a problem
  - Missing form another category

Leads to complaints about lack of power
Example of Procedure

- Simple regression, missing at random
  - Conservative: unbiased estimate, inflated SE
  - $n=100$, $\beta_0=0$, $\beta_1=3$
  - 30% missing at random, $\beta_1=3$

![Graph showing data points and regression line]

<table>
<thead>
<tr>
<th></th>
<th>Est</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b_0$</td>
<td>-1.5</td>
<td>1.4</td>
</tr>
<tr>
<td>$b_1$</td>
<td>3.01</td>
<td>0.27</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Est</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b_0$</td>
<td>-0.25</td>
<td>1.0</td>
</tr>
<tr>
<td>$b_1$</td>
<td>3.05</td>
<td>0.17</td>
</tr>
</tbody>
</table>
Example of Procedure

- Simple regression, not missing at random
  - Conservative: unbiased estimate, inflated SE
  - \( n=100, \beta_0=0, \beta_1=3 \)
  - 30% missing follow steeper line

<table>
<thead>
<tr>
<th></th>
<th>Est</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( b_0 )</td>
<td>-0.02</td>
<td>2.6</td>
</tr>
<tr>
<td>( b_1 )</td>
<td>2.82</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Requires robust variance estimate
### Example from R

**Data frame with missing values**

```r
> example.df
  x1  x2  x3  lab  fac
1  1  NA -0.9532650  UVW  ABC
2  1   2 -2.8903951  UVW  ABC
3  1   3 -0.1693143  UVW  ABC
4  1  NA -0.8343432  UVW  ABC
5  NA   5 1.0919509  UVW  ABC
6  1  NA  1.3706193  UVW  ABC
7  1   7 -1.7155066  UVW  ABC
8  1   8  0.6355785  UVW  ABC
9  1   9  0.7014913  UVW <NA>
10 1  10  0.4994391  UVW <NA>
```

**Filled in data with added indicator columns**

```r
> fill.missing(example.df)
  x1  x2  x3  lab  fac  Miss.x1  Miss.x2
1  1  1  6.285714 -0.9532650  UVW  ABC       0       1
2  1  1  2.000000 -2.8903951  UVW  ABC       0       0
3  1  1  3.000000 -0.1693143  UVW  ABC       0       0
4  1  1  6.285714 -0.8343432  UVW  ABC       0       1
5  1  1  5.000000  1.0919509  UVW  ABC       1       0
6  1  1  6.285714  1.3706193  UVW  ABC       0       1
7  1  1  7.000000 -1.7155066  UVW  ABC       0       0
8  1  1  8.000000  0.6355785  UVW  ABC       0       0
9  1  1  9.000000  0.7014913  UVW Missing           0       0
10 1  1 10.000000  0.4994391  UVW Missing           0       0
```

No cheating: You don’t get to fill in the y’s!
Background of Procedure

• Been around for a long time
  • Well suited to data mining when need to search for predictive features

• Reference
  • Paul Allison’s Sage monograph on Missing Data (Sage # 136, 2002).

• For a critical view, see Jones, M. P. (1996)
  • J Amer. Statist. Assoc., 91, 222–230
  • He’s not too fond of this method, but he models missing data as missing at random.
Expanded Osteo Data

• Fill in missing data
  • Grows from 126 to 208 possible Xs

• Saturated model results
  • Full sample but so few significant effects

Still missing interactions

Do in R
Stepwise Regression

• Need a better approach
  • Cannot always fit the saturated model
  • Saturated model excludes transformations such as interactions that might be useful

• Mimic manual procedure
  • Find variable that improves the current model the most
  • Add it if the improvement is significant.

• Greedy search
  • Common in data mining with many possible X’s
  • One step ahead, not all possible models
  • Requires caution to use effectively
Stepwise Example

- Predict the stock market

- Response
  - Daily returns (essentially % change) in the S&P 500 stock market index through April 2014

- Goal
  - Predict returns in May and June using data from January through April

- Explanatory variables
  - 15 technical trading rules based on observed properties of the market
  - Designed to be easy to extrapolate
Results

- Model has quite a few X’s but is very predictive and highly statistically significant.

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>29</td>
<td>0.00424379</td>
<td>0.000146</td>
<td>14.1056</td>
<td>&lt;.0001*</td>
</tr>
<tr>
<td>Error</td>
<td>52</td>
<td>0.00053947</td>
<td>0.000010</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. Total</td>
<td>81</td>
<td>0.00478325</td>
<td></td>
<td></td>
<td>&lt;.0001*</td>
</tr>
</tbody>
</table>

Residuals diagnostics check out fine...
Predictions

- Plot of predictions with actual
- Fit anticipates turning points.
Evaluating the Model

• Compare claimed to actual performance
  • $R^2 = 89\%$ with RMSE = 0.0032
  • How well does it predict May and June?

• SD of prediction errors much larger than model claimed

±2 RMSE

What went wrong?
Forward Stepwise

• Allow all possible interactions, 135 possible
  • Start with 15 X’s
  • Add 15 squares of X’s
  • Add $\frac{15 \times 14}{2} = 105$ interactions
  • Principle of marginality?

• Forward search
  • Greedy search says to add most predictive
  • Problem is when to stop?

• Use statistical significance?
  • What threshold for the p-value?
  • Follow convention and set $\alpha = 0.05$ or larger?

Response surface in JMP
Explanation of Problem

• Examine the definition of the technical trading rules used in the model

• Why did the stepwise get this so wrong?
  • Problem is classic example of over-fitting
  • Tukey “Optimization capitalizes on chance”

• Problem is not with stepwise
  • Rather it lies with our use of classical statistics
  • $\alpha=0.05$ intended for one test, not 135
Over-Fitting

• Critical problem in data mining
  • Caused by an excess of potential explanatory variables (predictors)

• Claimed error steadily shrinks with size of the model

• “Over-confident”
  • Model claims to predict new cases better than it will.

• Challenge
  • Select predictors that produce a model that minimizes the prediction error without over-fitting.
Problem in Science

- Source of publication bias in journals
- Statistics rewards persistence
How to get it right?

- Three approaches
  - Avoid stepwise (and similar methods) altogether
  - Reserve a validation sample (cross-validation)
  - Be more choosy about what to add to model

- Bonferroni rule
  - Set the p-value based on the scope of the search
  - Searching 135 variables, so set the threshold to $0.05/135 \approx 0.00037$
  - Result of stepwise search?

Bonferroni gets it right…
Nothing is added to the model!
Take-Aways

• Missing data
  • Fill in with an added indicator for missingness

• Over-fitting
  • Model includes things that appear to predict the response but in fact do not

• Stepwise regression
  • Illustrative greedy search for features that mimics what we do manually when modeling
  • Expansive scope that includes interactions
  • Bonferroni: Set p-to-enter = 0.05/(# possible)
Assignment

• Missing data
  • What do you do with them now?

• Try doing stepwise regression with your own software.
  • Does your software offer robust variance estimates (aka White or Sandwich estimates)

• Take a look at the ANES data
Next Time

- Review of over-fitting
  - What it is and why it matters
  - Role of Bonferroni

- Other approaches to avoiding over-fitting
  - Model selection criteria: AIC, BIC, …
  - Cross-validation
  - Shrinkage and the lasso