

# Lecture 3. Data Mining with Regression

---

## Questions from Previous Class?

**Quick summary of Lecture 2**

**Lab for next class**

---

## Key Topics

**Stepwise regression**

**Response surface**

**Over-fitting and the Bonferroni rule**

**Cross validation**

---

## Applications

### Financial modeling

- Can you look at a model and judge its adequacy without knowing how the form of the model was chosen?

### Osteoporosis

- What's important beyond age and weight?
- Do any of the biological markers, lab analyses help or are these unnecessary?  
If not needed, we can legitimately rely on self-reported data rather than require expensive lab results to diagnose presence of osteoporosis.

### Voting behavior in 2004, 2008

- Status of data files
  - Put 2004, 2008 JMP data on Z drive in stine directory
  - 2008 data file is a little raw
- Questions
  - What predicts the stated choice of voters?
  - Role of race and ethnicity in election? (coming evening lecture)

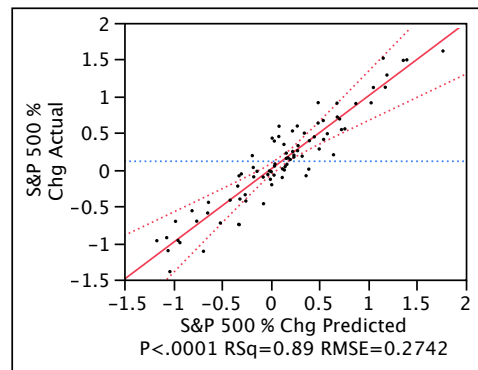
## Over-fitting

### Common problem

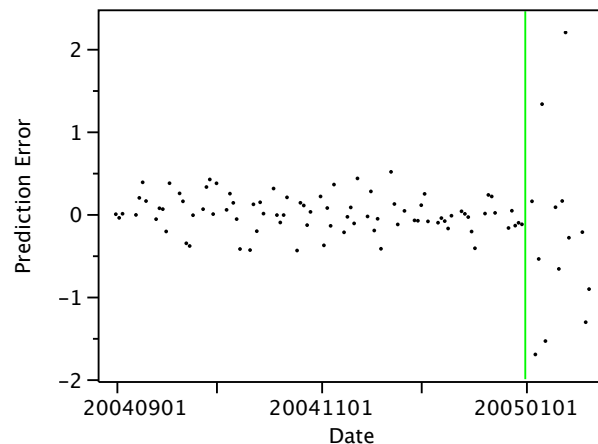
- Definition: Model does not perform so well as advertised due to selection bias caused by using the same data to pick the model and to estimate the fit.
- “Optimization capitalizes on chance.” (Tukey)
- Occurs in manual and automatic modeling

### Example: predicting the stock market

- Data on-line (stocks.jmp): market returns in 2004 and start of 2005
- Objective: predict 2005 based on what has happened at end of 2004
- Model included in data table is impressive (big  $R^2$ ,  $t$ 's are very large)



- Problem happens when use model for prediction



### Explanation of problem

- Look at the definitions of the explanatory variables!
- Model nonetheless looks great on paper, but is in fact a fiction produced by improperly using the algorithm known as stepwise regression.

## Stepwise Regression

### Algorithm to identify predictors

- Predictor  $\neq$  explanatory variable
- Algorithm picks variables as follows (forward stepwise)
  1. Initial model
  2. Add variable that boosts  $R^2$  the most
  3. Continue so long as p-value of next variable is larger than p-to-enter threshold
- Greedy search since it does not “look past” one step ahead

### Stock market example

- Initial model is empty (null model, which is the right answer)
- Stock of possible predictors built using JMP's response surface option. A response surface starting from 12 variables consists of  
original 12 variables + 12 squares +  $(12)(11)/2$  interactions = 90 choices
- Forward stepwise, with “promiscuous” p-to-enter set to 0.16 (called AIC)
- Backward stepwise, removing variables with p-value larger than 0.05.

### Cause of problem in forecasts

- First step picks the worst possible predictor, one that looks the most artificially statistically significant.
- Cure is to be sure that the variables chosen are real.

### Solution to problem

- Keep the algorithm and the large stock of possible variables, but use the algorithm more carefully in a way that acknowledges large number of possible predictors.
- Look at more variables then have to meet a higher standard for performance
  - a) Build the response surface from the initial stock of variables
  - b) Set initial search threshold p-to-enter =  $0.05/\#$  possible predictors
  - c) (Optional) Increase the threshold to  $q/\#$  possible predictors where  $q$  identifies the number of predictors identified in the first step. Continue doing increasing the threshold until it converges.

### Results for stock example

- 90 variables implies setting an initial threshold p-to-enter =  $0.05/90 \approx 0.00056$
- The initial step adds nothing. Why is this the right answer?

## Stepwise Regression

### Return to osteoporosis

- JMP cannot run the full response surface for these data (osteo\_big.jmp)
- Which model should we begin with? What's the appropriate initial model?  
Can we justify starting with variables such as Age, Weight already in the model rather than running them through the stepwise search? This is the role for theory.
- Define initial model starting with Age and Weight "locked in", then add others  
Response surface for all through H\_LOS (with missing, 117 columns), then just linear in the remaining 90 columns to accommodate JMP's limits on the size of the problem.  
Approximately  $117 + 117 + 117 * 116 / 2 + 90 = 7,110$  variables (very rough count)  
JMP software often adds a lot of variables that are not really needed, such as squares of dummy variables, interactions between mutually exclusive dummy variables, and between missing indicators and the filled-in variable.

### Stepwise results

- Baseline for comparison  
Saturated linear model, using all 208 columns.  
Lots of collinearity  
Fit obtains  $R^2 = 0.54$  (Adj  $R^2 = 0.45$ )
- Running in promiscuous mode: will the search ever stop?
- Stepwise search over linear terms      initial p-to-enter =  $0.05/208 = 0.00024$   
Finds 7 predictors (including Age and Weight), gets  $R^2 = 0.40$   
Weight, Fracture?, Race-2, Age, RhueArth, H\_LOS, OSTASE  
If use larger  $7/208$  threshold,  $R^2$  approaches the Adj  $R^2$  of saturated model but with a much more parsimonious selection of variables.
- Search over partial response surface      initial p-to-enter set to  $0.05/7110 \approx 0.000007$   
Center the interactions  
Hard for a variable to get into the model (|t| must be about 4)  
Search (after removing one irrelevant by collinearity) adds 3 predictors in addition to the initial model that has Age and Weight.
- Which model is the right one to use?

### Calibration

- Are these models calibrated?
- Do they predict as well as they claim to predict?



## Cross Validation

### Hold-back sample(s)

- Reserve a portion of the data to test the accuracy of the final model
- Randomly divide data into two or three subsets
  - (a) training sample    use these data to estimate the model
  - (b) tuning sample    Optional: some methods need these to pick model
  - (c) test sample        evaluate performance of model (eg, squared error)
- Repeat the procedure as often as you have time
- Claim: An algorithm like stepwise regression with the Bonferroni type p-values does not need to reserve a sample to pick the variables or test the model. The  $R^2$  you get is an unbiased estimate of how well the model will perform.

### Catch-22

- How much data to allocate for estimation versus testing?
- A large test sample gives a good estimate of accuracy, but leaves less for picking and estimating the model.
- A large estimation/tuning sample produces a better model, but one whose properties are less well understood.

### Cross-validation is optimistic

- Test sample is a random sample from the same population as the estimation sample.
- That's often not the case in real modeling (populations drift over time)

### Example with osteo data

- Randomly divide data in half (split-sample CV; formula in first column of osteo\_big)
- Fit models: saturated and stepwise
- Compare claims from fit to the training sample to accuracy on test sample.
- Saturated model
  - 197 predictors,  $R^2 = 0.63$ ,  $\text{adj } R^2 = 0.45$ ,  $\text{RMSE} = 0.95$
  - SD of prediction errors in test sample is inflated (extrapolation effects?)
- Stepwise model (make sure interactions are centered)
  - Cost of cross validation: cannot find many predictors
  - Start with Age, Weight: does not want to add any
  - Raise p-to-enter to  $1/7110 = 0.00014$  and it adds 2 or 3
  - 5 predictors,  $R^2 = 41\%$ ,  $\text{RMSE} = 0.99$ : similar accuracy in both.

---

## What Next

### Lab on Thursday in Newberry

#### Friday

- Improving stepwise, more practice examples

- Saturated model

196 predictors,  $R^2 = 0.62$ ,  $\text{adj } R^2 = 0.44$ ,  $\text{RMSE} = 0.97$

SD of prediction errors in test sample is inflated (extrapolation effects?)

- Stepwise model (make sure interactions are centered)

See cost of cross validation: cannot find so many predictors

Start with Age, Weight: does not want to add any, so raise p-to-enter to  $1/7110 = 0.00014$  and it adds 2 more predictors (both interactions)

4 predictors,  $R^2 = 38\%$ ,  $\text{RMSE} = 1.03$

Note the outlier: We have to adjust stepwise to handle these sparse problems.  
(Friday)