

# Lecture 6. Improving Stepwise Regression

---

## Questions from Previous Classes?

---

### Topics

#### Review cross validation

#### Improving stepwise:

**standard errors**

**calibration**

**experts, auctions, and strategies**

---

### Cross Validation

#### Split data

- Training data used to build, estimate model
- Test data used to judge accuracy of model

#### Over-fitting

- Fits better on the training sample than on the test sample

#### Limitations

- Reduce the amount of data available for estimation
- Optimistic since insures that the test and training data are identical

#### Examples

- Experience with HR data (population drift), osteoporosis score (new population)
  - ANES 2008
-

## Improving Stepwise Regression

### Incremental improvements

- Estimating the change in the residual sum of squares
- Sandwich estimator for the standard error (aka, White estimator)
- Calculation of p-value from alternative distribution

### Major improvements

- Different way to search for variables
- Different criterion to evaluate the p-value

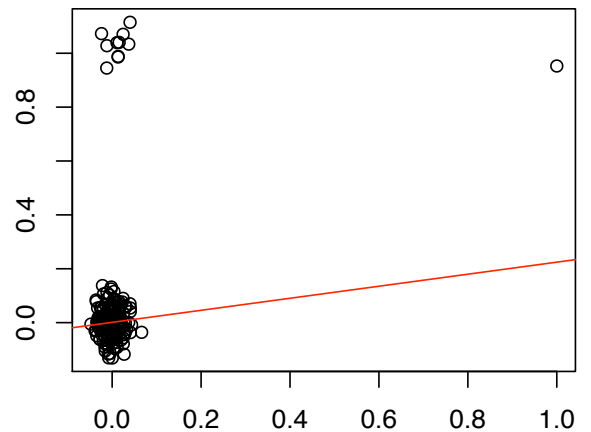
### Examples

- Do examples outside of JMP since require rearranging some calculations.
- You can do calculations inside JMP, but I've not learned how to program it.

## Illustrative Problem

### Outlying leverage point

- Often produced by interactions
- Simple example with sparse data
  - 10,000 cases have  $x = 1$
  - 1 case has  $x = 1$
- Rare event, highly leveraged
  - 1/1,000 chance for  $y = 1$
- Estimate of the statistical significance?
  - Estimated slope is  $b_1 \approx 0.22$



### Estimates of statistical significance

- Common sense
  - 1/1000 in body of data, so 1 for 1 for outlier means  $p = 0.001$
- Regression
  - $t = 15$  with p-value  $< .00000001$

## Incremental Improvements

### Calculation of the change in fit

- Standard approach is to add the variable to the current model, then use the residuals after adding the variable to estimate the standard error of the model.
- Alternative approach is to use the residuals assuming the null hypothesis  $\beta=0$  for judging the quality of the fit. Rather than use the residuals after adding the variable, use the residuals you had before adding the variable.
- Explanation comes from the expression for the SE of a slope in regression
- Small effect in this example:  $t$  falls from 15.2 to 15

### Sandwich estimator

- Robust estimate of scale (also makes robust to heteroscedasticity)
- Use residuals of model in expression for the SE of the slope in place of the errors and taking expectations
- Name comes from the expression for the calculations
- Effect in this example is huge (from  $SE = 0.015$  to  $0.16$ )  
Similar residual if combine with the prior stage residuals ( $SE = 0.20$ )
- Implied  $t$ -statistic reduced from  $t=33.3$  to about  $0.22/0.16 \approx 1.375$ . Not significant!

### Calculation of p-value

- $p$ -value calculation presumes normal model for the sampling distribution of the estimator.

Derived from central limit theorem, which presumes roughly comparable influence on the estimator.

- Estimator with rare events and leverage violates conditions for the CLT

Even with 10,000 observations, estimator is not normally distributed. Some cases have much more contribution to the fit. Sandwich estimator handles this effect.

- If not using the sandwich estimator, then need this adjustment

Don't know, so place a bound on the size of the  $p$ -value derived from a result in math stat known as Bennett's inequality. Messy.

No matter what, assuming still independence, how big might the  $p$ -value be? Calculation in this example bounds the  $p$ -value at about  $1/100$ . Not significant if this variable is found by a search.

### Discussion

- Loss of power versus finding significant effects.
- If not implemented (as in JMP), then watch for isolated cases that introduce this effect. Leverage plots show you these.

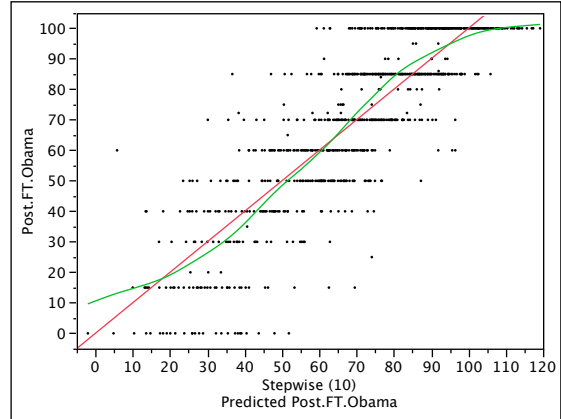
## Bigger Changes I: Auto-calibration

### Fit from stepwise

- Validates well, but not entirely calibrated
- Plot shows calibration in training sample.
- Polynomial matches for sigmoidal curve.
- Check calibration in validation too

### Bounded response

- Response (post-election FT Obama) is bounded between 0 and 100
- Like predicting a 0/1 variable using linear regression
- Regression line based on fitted model produces predictions that are larger than 100 and less than 0.



### Solution: auto-calibrate

- After each added variable, check whether latest model is calibrated.
- Optionally add functions of current predictions to calibrate the model.
- Current application fits a polynomial using the current predictions from the model as the calibration variables (easier to do than a spline and can test like others).

## Bigger Changes II: Experts and Auctions

### How do you search for variables

- Like a computer, try them all?
- Try a little of this, then depending on what happens, something else?

### Expert

- Sequential search strategy. Search adapts to the current state of the model and the current collection of possible explanatory variables.
- Examples from genetics. Do the same with voting?
- Substantive experts versus parasitic experts

### Auction

- Multiple experts, so how to decide which one to use?
- Code runs an auction, with  $\alpha$  level as the currency.
- Experts bid some portion of  $\alpha$  wealth for the right to recommend a variable.
  - If p-value of suggested variable  $<$  bid, model adds the variable and the expert gains more  $\alpha$  for future bids.
  - If p-value of suggested variable  $>$  bid, variable is not added and the expert loses the wager.
- Adaptive choice of experts. Those that pick useful variables grow in wealth so can bid more. Those that do not gradually run out of  $\alpha$  wealth and cannot win a bid.

### Theory

- Can prove that the procedure does not over-fit (at least under some idealized conditions that are not quite the same as in regression modeling).
- Foster and Stine (2008), JRSS.

