Bootstrap Methods in Regression

Questions

Have you had a chance to try any of this?

Any of the review questions?

Getting class notes from the web

Go to my web page
www-stat.wharton.upenn.edu/~stine/mich

Lecture notes are PDF files (Adobe Acrobat).

Updated daily (usually sometime after class) and will remain on the web for some time.

Software

"Script" files for R commands.

Try software while you are here.

Yet more Summer Program t-shirts

Overview

Calibration

Powerful idea of using the bootstrap to check itself.

Resampling a correlation

Correlation requires special methods
Its sampling distribution depends on the unknown population correlation.

Bootstrap does as well as special methods.

Simple regression

Model and assumptions

- Leverage, influence, diagnostics
- Animated simple regression
- Smoothing

Resampling in regression

Two methods of resampling

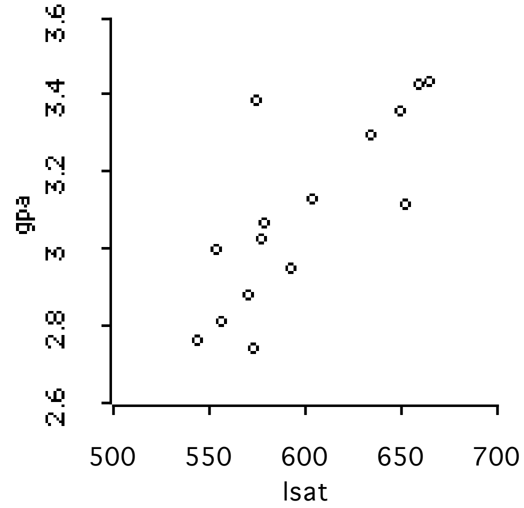
- Residual resampling (fixed X)
- Observation resampling (random X)

Picking a method of resampling

Inference for a Correlation

Classic bootstrap illustration

Efron’s law school data
  LSAT and GPA values for 15 law schools



How to make an inference for the correlation?
- What is the confidence interval?
- What is the population anyhow?

The sample correlation r = 0.776

New type of complexity

SE of average does not depend on □, but
SE of sample correlation depends on □.

Classical Inference for the Correlation

Fisher's z transform

The sample correlation is not normal, but Fisher's z-transform gives a statistic that is close to normal

$$z = f(r) = \tfrac{1}{2}\log\frac{1+r}{1-r}$$

This stat is roughly normal with mean f(ρ) and

$$SD = \frac{1}{\sqrt{n-3}}$$

Example with the law school data

Fisher's z transformation gives for the 90% confidence interval the range
[0.507, 0.907] = [.776-.269, .776+.131]

Fisher's interval is *not* of the usual form
[estimate ± 2 SE of estimate]
but instead is very asymmetric.

Why should the interval be asymmetric?

Bootstrapping the Correlation

How to resample?

Keep the data paired – resample observations
(What happens if you do not keep the pairing?)

Same basic resampling iteration

- Collect B bootstrap replications
- Repeatedly calculate the correlation for a large number of bootstrap samples

Raw calculations, one last time

Explore choice of # of bootstrap replications

Procedure

- Start with 50
- Add further bootstrap replications
- Compare the results as they accumulate

Observe that

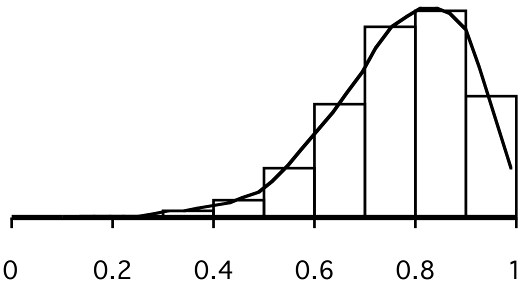
- SE “settles down” quickly but
- Lower limit of the CI is not stable until we have a large number of replications.

Correlation Results

Plot the bootstrap distribution

The bootstrap distribution is skewed

- clearly not normal
- has hard upper limit at 1
- foolish to use interval like r ± 2 SE(r)



Note: Fisher's transformation accommodates this special kind of asymmetry; the range of Fisher's z transform is not bounded.

Comparison of intervals

With 3000 replications:

90% bootstrap interval

[0.520, 0.943] = [.776 - .220, .776+.167]

Fisher's interval

[0.507 , 0.907] = [.776 - .269, .776+.131]

Both are skewed and within [-1,1] limits.

The bootstrap works without knowing Fisher's special transformation – or *assuming* normality.

Exploring the Bootstrap Distribution

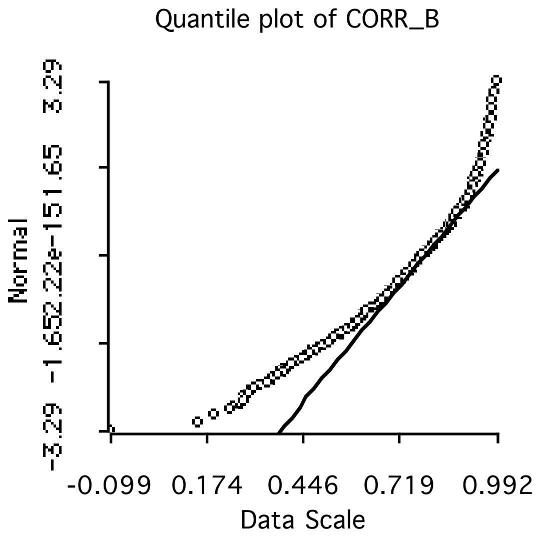
Resampled correlations are not normal

Kernel density estimates

These alternatives to histograms avoid binning the data, but require you to choose how much to smooth the data. You can explore these options using a *slider*.

Quantile plots

Shows how close to normality, focusing on the extremes rather than the center of the data.



Simple Regression Model

Assumptions for one-predictor model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

1. Independent observations
2. Equal variance $E(\varepsilon) = 0$, $Var(\varepsilon)=\sigma^2$
3. Normally distributed error terms

+ X is “fixed” OR perfectly measured, ind of ε

Data generating process

The “hot dog” model…

Least squares

Pick the line with the smallest sum of squared vertical deviations (residuals).

Least squares estimator (OLS) is “best”:
What does “best” mean in this context?

Issues

Linear? Is a line a good summary?
Really want ave(Y|X)

Outliers? What effects can these have?

Inference for the slope?

Diagnostics for Simple Regression

Examples

"Typical analysis" Law school data

- Small sample size (n=15)
- Let X=LSAT predict Y=GPA

"Unusual analysis" Voting in Florida

- Moderate sample size (n=67)
- Large outlier

Exploring a scatterplot

Animated sensitivity

Add OLS line to a scatterplot, then change the "mouse mode" to allow you to interactively drag points and watch the line shift.

Leave-one-out diagnostics

Fit the regression using the regression command to learn more about this important collection of regression diagnostics.

- Leverage (potential effect)
- Influence (changes if removed)
- Standardized residuals.

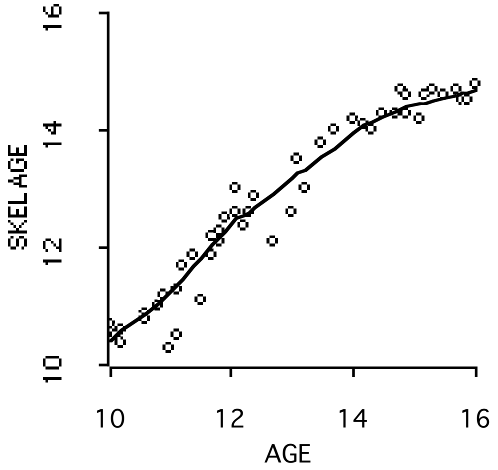
Linked diagnostic plots.

Fox *Regression Diagnostics*. Sage green mono.

Smoothing

Further example of smoothing

Skeletal age as measure of physical maturity.



Is the bend real?
This plot shows a “loess” smooth of the data.

Diagnostic procedure

Smooth curve based on local robust averaging should track the fitted model.

Use smoothing to detect curvature in residuals.

Bootstrapping a smoother

Visual inspection of fitted curves
Resample observations.

Want to know more?

Modern regression course.

Resampling in Regression

Two approaches

Generalize approaches to two-sample test
A two-sample test *is* a simple regression with a categorical (dummy variable) predictor.

Random X (observation resampling)
Resample <u>observations</u> as with correlation example or in one approach to the t-test.

Fixed X (experimental, residual resampling)
Resample <u>residuals</u> as follows

- Fit a model and compute residuals
- Generate BS data by
  Y* = (Fit) + (BS sample of OLS residuals)

Comparison

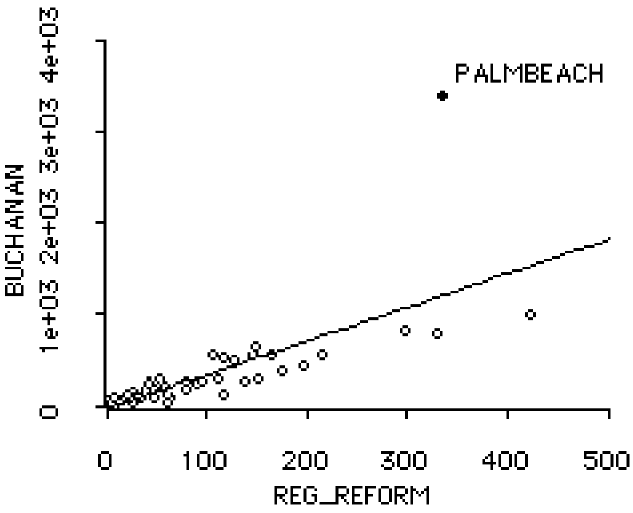
	Resample	
	Observations	Residuals
Model-dependent	No	Yes
Fixed design X	No	Yes
Maintains (X,Y) assoc.	Yes	No

Differences are most apparent
when something is "peculiar" about the regression model or data, e.g. a severe outlier.

Observation vs. Residual Resampling

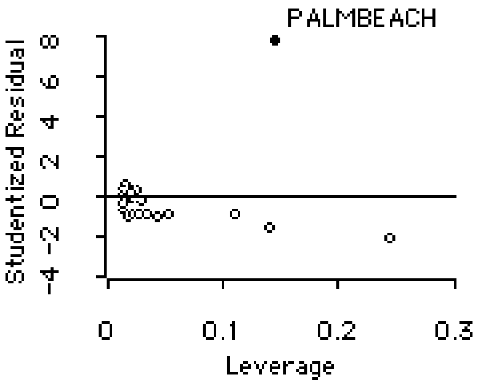
Florida 2000 US Presidential election results

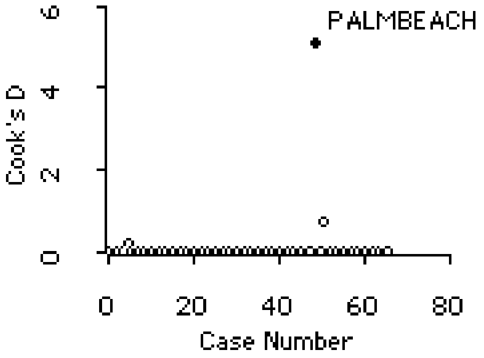
Data show by county
- number registered to Reform Party.
- number of votes received by Buchanan.



Slope estimate
b = 3.7 SE(b) = 0.41 (t □ 9)

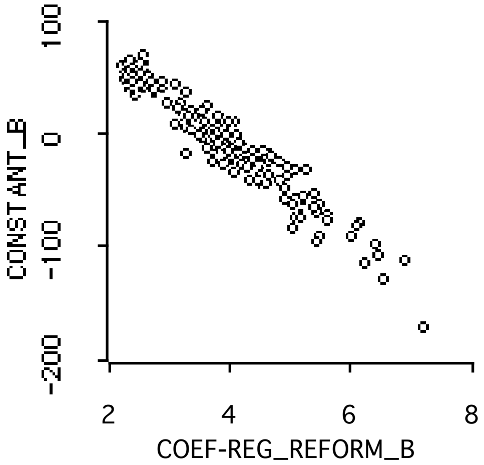
Palm Beach is not so leveraged, but is "influential"





Observation resampling

Sample counties as observations.



Replicates reminds of collinearity.
The slope and intercept are negatively correlated in a regression when X-bar>0.

SE* = 1.15 ... much larger than OLS claims

Residual resampling

Sample residuals of fitted model.

SE* = 0.37 ... about same as OLS claims.

Why different SE estimates? Random > Fixed SE

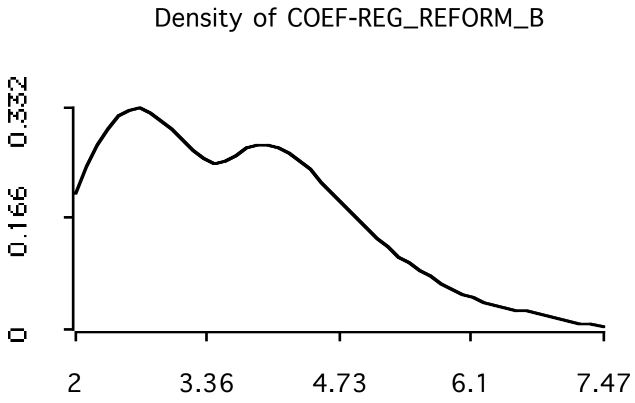
Is X fixed or is X not fixed?
Fixed usually gives a smaller estimate, Var(b|X) '≤' Var(b)

Resampling with Influential Values

Comparison of resampling methods

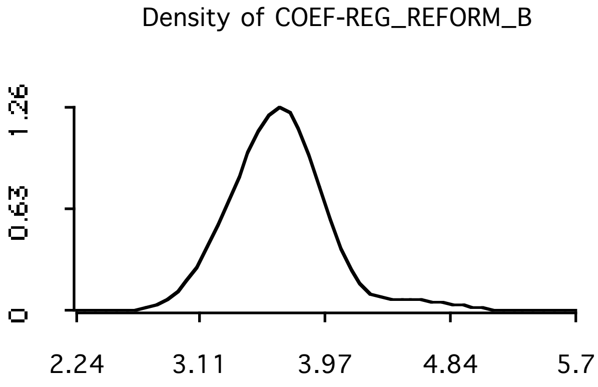
Observation resampling

Keeps Palm Beach residual at a leveraged location, leading to bimodal distribution.



Residual resampling

"Smears" the Palm Beach residual around, giving a "normal" BS distribution.



Extremely different impression of the accuracy of the fitted model. Which is right?

Which Method is Right?

Observation resampling

+ Does not assume so much of fitted model
  Example with unequal variance.
  Example with nonlinearity.

± Estimates unconditional variation of the slope rather than the conditional variation.

± Does not always agree with classical SE

– Not appropriate in Anova designs, patterned X's such as time trends (at least not without special care!)

– Slower to compute (less important now)

What would happen for "another sample"?

Would Palm Beach again be an outlier?

Would it again have a positive residual?

Seems that we might expect Palm Beach to be an outlier, and the direction of the residual also seems plausible.

Asymptotics (i.e., really big samples)

Asymtotic results

Describe what happens as the sample size gets larger and larger.

As the sample size grows (with other conditions),

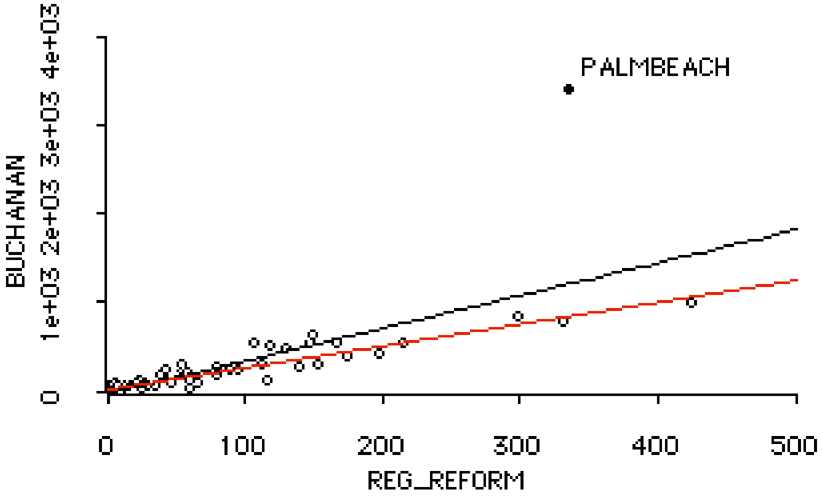
Random resampling and fixed X resampling methods become similar, assuming the model is correctly identified.

Relation to classical

Bootstrap SE* □ usual OLS formula for residual resampling as number of BS replications B –>

Robust Regression

Automatically adjusts for outliers



Comparison to OLS

OLS fit

Variable	Slope	Std Err	t-Ratio	p-value
Constant	1.5325	46.61	0.033	0.97
REG_REFORM	3.6867	0.41	9.019	0.00

R Squared 0.56 Sigma hat 301.9

Robust fit

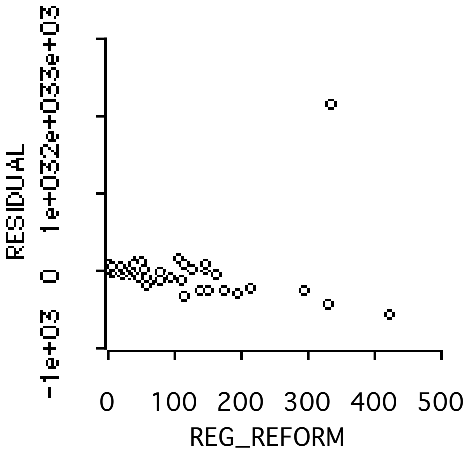
Robust Estimates (HUBER, c=1.345):

Variable	Slope	Std Err	t-Ratio	p-value
Constant	45.52	34.9	1.302	0.20
REG_REFORM	2.44	0.3	7.948	0.00

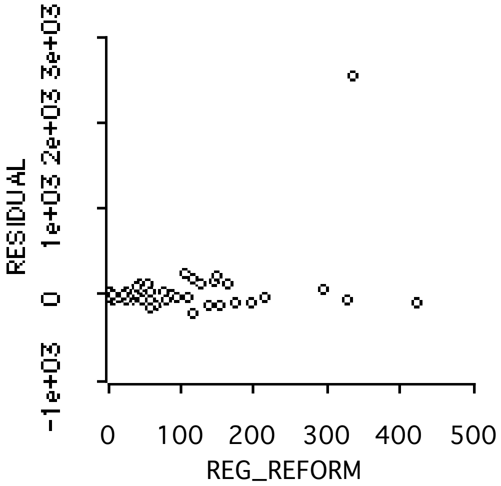
“R Squared” 0.86 Sigma hat 82.53

Size of outlier

OLS



Robust

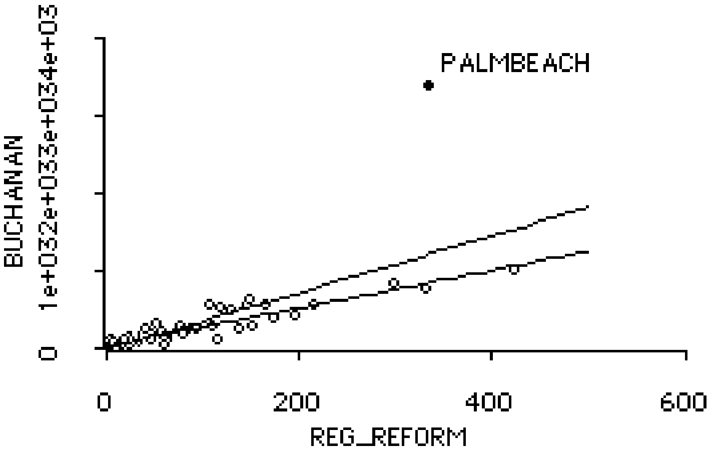


Outlier is larger, and more apparent relative to the scale of the fitted model…

OLS Residual SD ≈ 300

Robust Residual SD ≈ 80

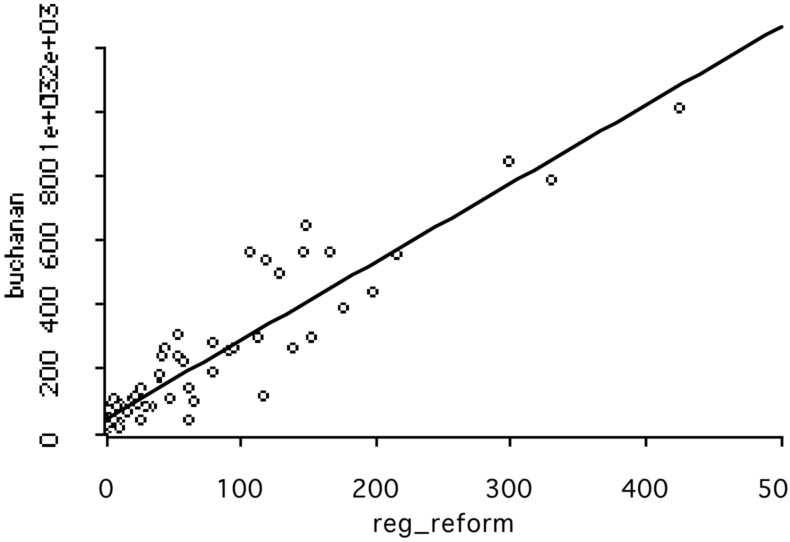
OLS fit without Palm Beach



OLS without Palm Beach

Very similar to robust regression.

Plot looks very different with Palm Beach removed from the data set.



OLS regression (n=66)

```
Least Squares Estimates for BUCHANAN :
Variable          Slope     Std Err     t-Ratio   p-value
Constant          50.28     12.98        3.873     0.00
REG_REFORM         2.44      0.12       20.180     0.00

R Squared: 0.86
Sigma hat: 83.3
```

Reasons to Bootstrap in Regression

Confidence intervals and SE's

Unless you are doing something special (or the data are unusual), the bootstrap typically gives you very similar SEs and confidence intervals.

So why bootstrap?

You learn more about regression.

Looking at the BS distributions helps you understand what's going on in the regression.

You can use methods other than least squares, methods that are less affected by outliers.

You can ask some more interesting questions.

SE is seldom all that we have interest in.

Inference for a robust regression

Simple questions can be hard to answer:
- Which X's to put into the equation?
- Where is the maximum of this fitted curve?

Things to Take Away

Bootstrap resampling in regression

Can be done in two ways, depending on the problem at hand

- residual resampling (fixed)
- observation resampling (random)

Properties of the bootstrap are related to leave-one-out diagnostics (leverage, influence)

NEXT TIME...

Special applications in regression.

Resampling in multiple regression.

Other issues in multiple regression

- missing data (just a little to say)
- measurement error (a little more)

Review Questions

What assumption is hardest to check, yet perhaps most important in regression?

The assumption is that the observations are independent of one another. Unless you have time series data, there are few graphical ways to spot the problem. You’ve got to know from the substance of the problem.

Do leverage and influence mean the same thing?

No, but they are related. An observation that is unusual in “X space” is leveraged. In simple regression, leveraged observations are at the extreme left and right edges of the plot. In contrast, influence refers to how the regression fit changes when an observation is removed from the fit. Heuristically,

Influence ≈ Leverage □ (Stud. Residual)

That is, to be influential requires leverage *and* a substantial residual.

How should you use the various regression diagnostic plots?

Residuals on fitted: lack of constant variance
StudRes. on leverage: source of influence
Residual density: normality

What would happen if we sampled X and Y separately when bootsrapping the correlation?

The "true" correlation in the BS samples would be zero. Since we would be independently associating values of X with values of Y, the resulting correlation would be zero; X and Y would by construction be independent.

How does residual resampling (fixed X) differ from observation resampling (random X)?

Residual resampling requires a "true" model in order to obtain the residuals which are resampled. Observation (or random) resampling does not. Residual resampling keeps the same X's in every bootstrap sample.

Which is larger?

Random resampling usually leads to a larger estimate of standard error (with enough bootstrap replications) since it allows for more sources of variation (from randomness in X's)

How does the bootstrap indicate bias?

The average of the BS replicates will differ from the observed value in the sample. For example, suppose the average of the bootstrap replicates is less than the original statistic. Since the original statistic plays the role of the population value, this implies that the original statistic is itself less than the real population value – and is thus biased.