Topic 16 - Other Remedies

STAT 525 - Fall 2013

Outline

- Ridge Regression
- Robust Regression
- Regression Trees
- Piecewise Linear Model
- Bootstrapping

Ridge Regression

- Modification of least squares that addresses the multicollinearity problem
- Recall least squares suffers because $(\mathbf{X}'\mathbf{X})$ is almost singular thereby resulting in unbiased but highly unstable parameter estimates
- Ridge regression considers the bias-variance tradeoff and allows for slight bias in hopes of a dramatic improvement in variance
- Imposes a ridge constraint

$$\text{minimize} \sum (y_i - Z_i\beta)^2 \text{ s.t. } \sum \beta_j^2 \leq t$$

Convention: Y is centered and Z is standardized ($\mu = 0, \sigma = 1$)

Ridge Regression

- Can express ridge contraint in terms of finding β to minimize the penalized residual sum of squares

$$(Y - Z\beta)'(Y - Z\beta) + c\sum \beta_j^2$$

- Above solution for β same as considering the correlation transformation so the normal equations are given by $\mathbf{r_{XX}b} = \mathbf{r_{YX}}$. Since $\mathbf{r_{XX}}$ difficult to invert, we add a bias constant, c.

$$\mathbf{b^R} = (\mathbf{r_{XX}} + \mathbf{cI})^{-1}\mathbf{r_{YX}}$$

- Note: LASSO is a variation on this approach

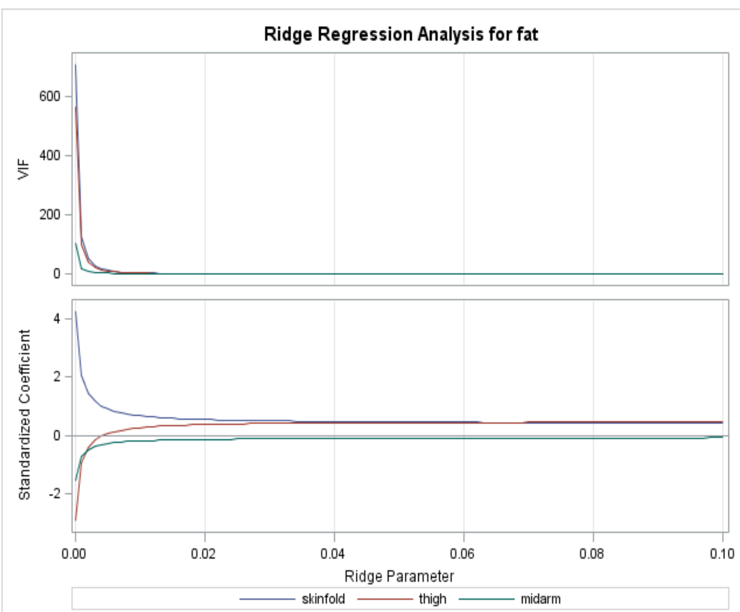
$$\text{minimize} \sum (y_i - Z_i\beta)^2 \text{ s.t. } \sum |\beta_j| \leq t$$

Choice of c

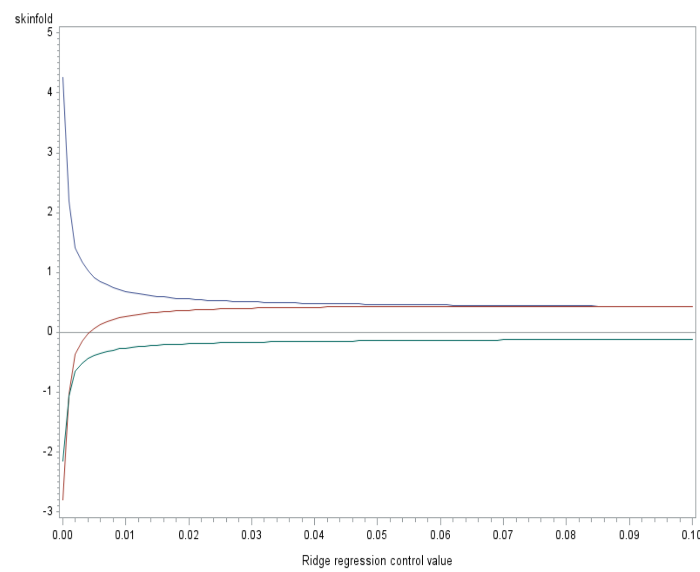
- Key to approach is choice of c, called the tuning or shrinkage parameter
- Common to use the *ridge trace* and VIF's
  - Ridge trace: simultaneous plot of $p-1$ parameter estimates for different values of $c \geq 0$. Curves may fluctuate widely when c close to zero but eventually stabilize and slowly converge to 0.
  - VIF's tend to fall quickly as c moves away from zero and then change only moderately after that
- Choose c where things tend to "stabilize" or better yet, use cross-validation
- SAS has option ridge=c;

SAS Commands

```
data a1;
   infile 'U:\.www\datasets525\Ch07ta01.txt';
   input skinfold thigh midarm fat;

proc print data=a1;
run;

proc reg data=a1 outest=b;
   model fat=skinfold thigh midarm /ridge=0 to .1 by .001;
run;

symbol1 v='' i=sm5 l=1;
symbol2 v='' i=sm5 l=2;
symbol3 v='' i=sm5 l=3;
proc gplot;
 plot (skinfold thigh midarm)*_ridge_ / overlay vref=0;
run;
quit;
```

Ridge Plot



Robust Regression

- Want procedure that is not sensitive to outliers
- Focus on parameters which minimizes
  - sum of absolute values of residuals (LAR)
  - median of the squares of residuals (LMS)
- Could also consider iterating through weighted LS where the residual value is used to determine the weight (IRLS)
- See pages 439-449 for more details
- Both robust and ridge regression are limited by more difficult assessments of precision (i.e., standard errors). Bootstrapping is often used.

Nonparametric Regression

- Helpful in exploring the nature of the response function
- i=sm## is one such approach
- All version have some sort of smoothing
- See pages 449-453 for more details
- Interesting theory with much research in both frequentist and Bayesian approaches

Regression Trees

- Very powerful nonparametric regression approach
- Standard approach in area of “data mining”
- Basically partition the X space into rectangles
- Predicted value is mean of responses in rectangle

Growing a Regression Tree

- Goal is to minimize SSE
- First split the data into two regions
- Find regions such that they minimize

$$SSE = SSE(R_1) + SSE(R_2)$$

- Next split one of the current regions and repeat
- Number of splits based on “cost criteria”
- Trade off between minimizing SSE and complexity

Piecewise Linear Regression

- At some points or points, the slope of the relationship changes
- If points known, can build into standard regression framework
- If points unknown, becomes a much more difficult problem
  - How many change points?
  - Location of change points?

Example

- Looking at the price of a manufactured unit versus the size of the lot in which the unit was produced
- Have $n = 60$ cases
- Assume cost relationship different after the lot size is larger than some unknown C_x because some operating efficiencies kick in (e.g., bulk raw materials, storage)
- Will consider two piecewise linear models

$$\text{Model 1: } E(\text{Cost}) = \beta_0 + \beta_1 \text{lotsize} + \beta_2 \max(0, \text{lotsize} - C_x)$$

$$\text{Model 2: } E(\text{Cost}) = \begin{cases} \beta_0 + \beta_1 \text{lotsize} & \text{lotsize} \leq C_x \\ \beta_0 + \beta_1 C_x & \text{lotsize} > C_x \end{cases}$$

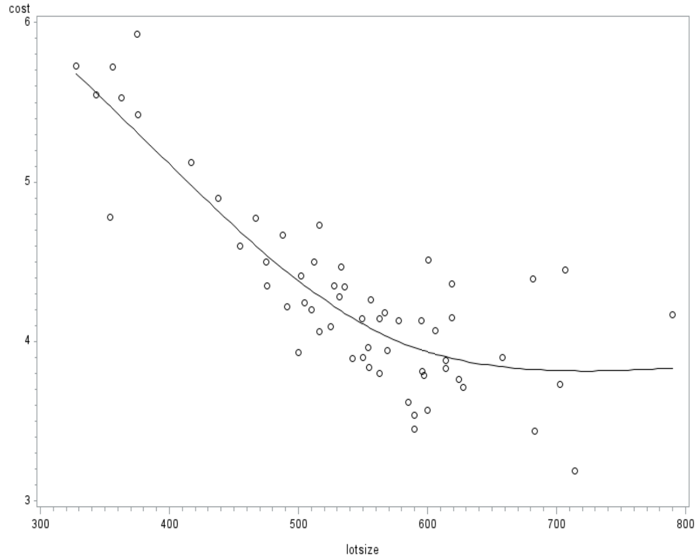
Model Parameters

- Model 1 will contain
  - Intercept
  - Changepoint value C_x
  - Slope for $X \leq C_x$
  - Slope for $X > C_x$
- Model 2 will contain
  - Intercept
  - Changepoint value C_x
  - Slope for $X \leq C_x$
- Consider fixed values of C_x and estimate remaining parameters using least squares
- Choose C_x which results in smallest MSE

SAS Commands

```
data a1;
 infile 'U:\.www\datasets525\manufacturing.txt' dlm='09'x;
 input case lotsize cost;                  **Tab delimited**

symbol1 v=circle i=sm70 c=black;
proc sort data=a1; by lotsize;             **Generate scatterplot**
proc gplot data=a1; plot cost*lotsize;

data changept; set a1;
   do changept = 450 to 650 by 5;
     if lotsize le changept then do;
       cslope=0; lotsize1=lotsize;         **Create data set**
     end;
     if lotsize gt changept then do;
       cslope=lotsize-changept; lotsize1=changept;
     end;
     grp=(lotsize <= changept);
     output;
   end;
```

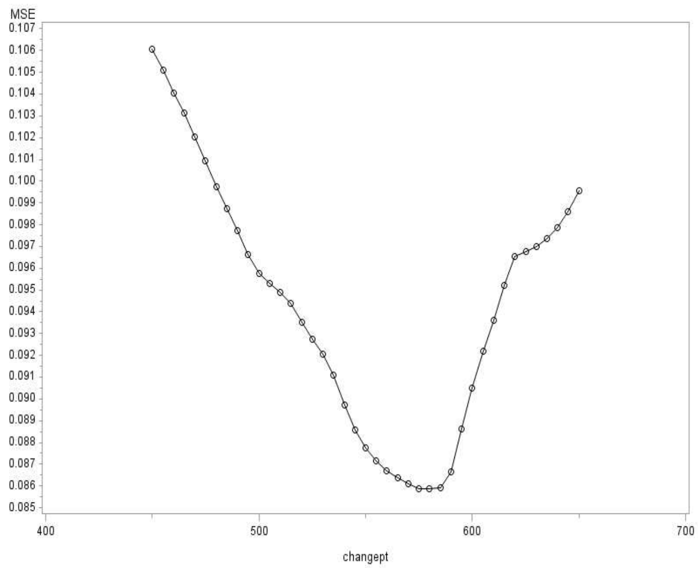
Scatterplot



SAS Commands Model 1

```
**Fit model 1 with different changepoints;
proc sort data=changept; by changept;
ods html close;                        ***Turns off html output;
proc reg data=changept;
   model cost=lotsize cslope;
   by changept;
   ods output FitStatistics=b1;
run;
ods html;                               ***Turns on html output;

data b2;
 set b1;
 if Label1='Root MSE';
 MSE = cvalue1*cvalue1;

symbol1 i=join;
proc gplot data=b2;
 plot MSE*changept;
run;
```

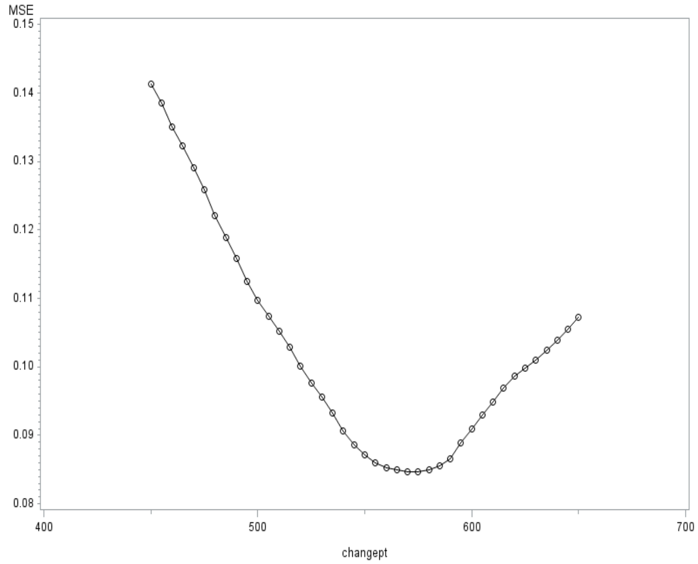

Profile plot Model I



SAS Commands Model 2

```
**Fit model 2 with different changepoints;
proc sort data=changept; by changept;
ods html close;
proc reg data=changept;
   model cost=lotsize1;
   by changept;
   ods output FitStatistics=b3;
run;
ods html;

data b4;
 set b3;
 if Label1='Root MSE';
 MSE = cvalue1*cvalue1;

symbol1 i=join;
proc gplot data=b4;
 plot MSE*changept;
run;
```

Profile plot Model 2



Output - Model 1 ($C_x = 580$)

Analysis of Variance

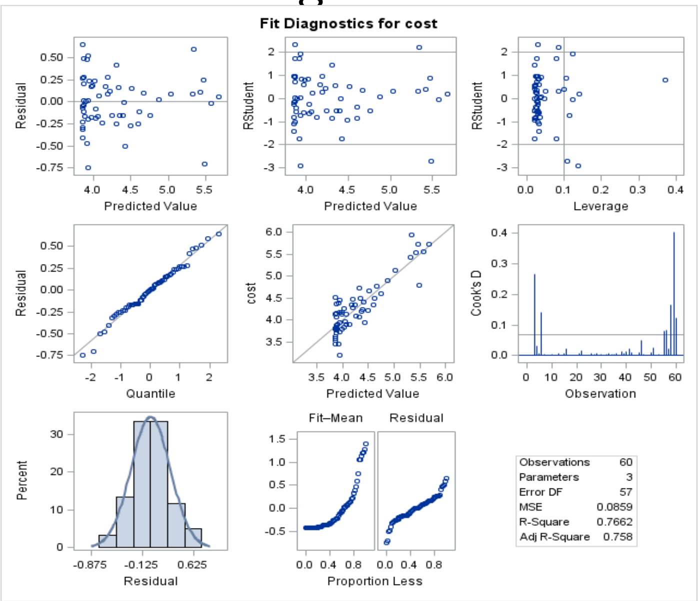
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	16.03966	8.01983	93.40	<.0001
Error	57	4.89413	0.08586		
Corrected Total	59	20.93379			

Root MSE	0.29302	R-Square	0.7662
Dependent Mean	4.28367	Adj R-Sq	0.7580
Coeff Var	6.84045		

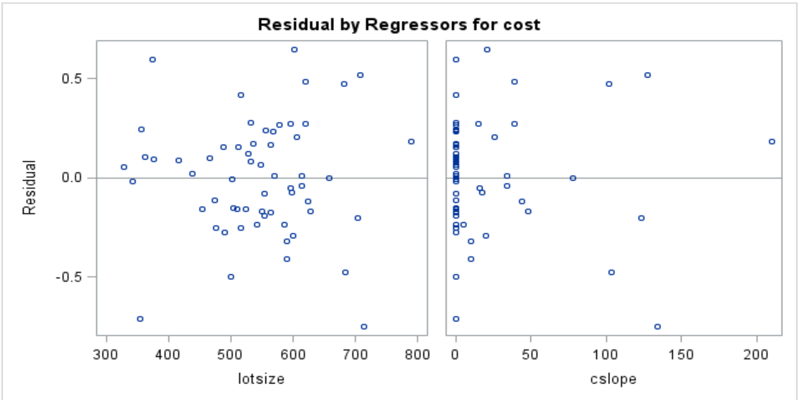
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	8.05373	0.29058	27.72	<.0001
lotsize	1	-0.00725	0.000566	-12.80	<.0001
cslope	1	0.00790	0.00131	6.03	<.0001

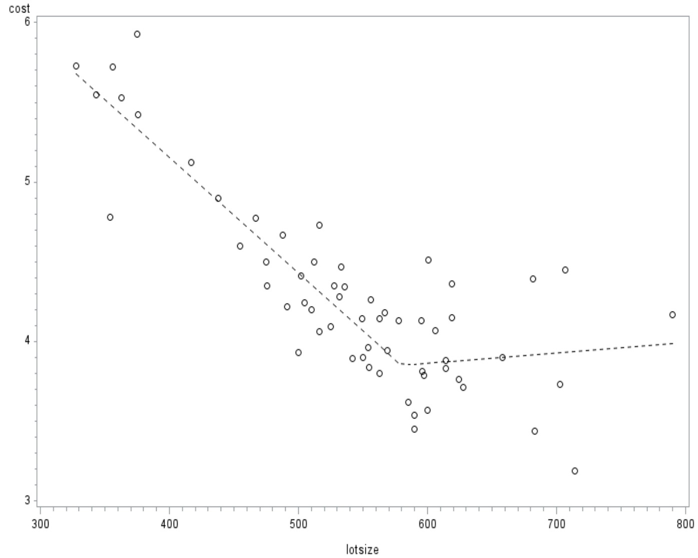
Diagnostics



Residual Plot



Fitted "Curve"



Output - Model 2 ($C_x = 570$)

Analysis of Variance

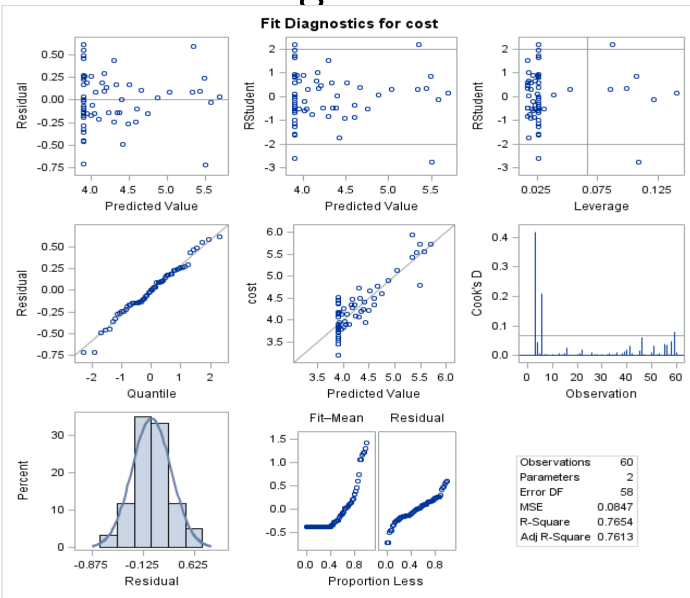
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	16.02217	16.02217	189.20	<.0001
Error	58	4.91162	0.08468		
Corrected Total	59	20.93379			

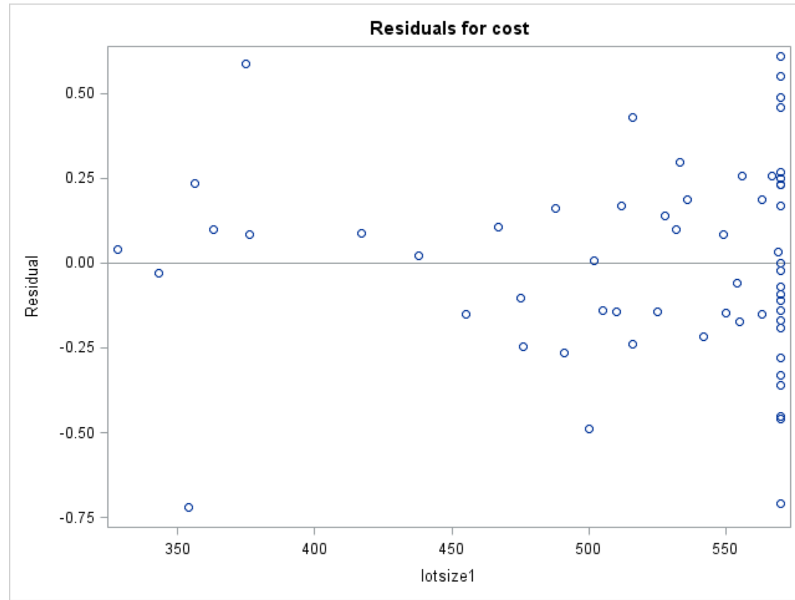
Root MSE	0.29100	R-Square	0.7654
Dependent Mean	4.28367	Adj R-Sq	0.7613
Coeff Var	6.79333		

Parameter Estimates

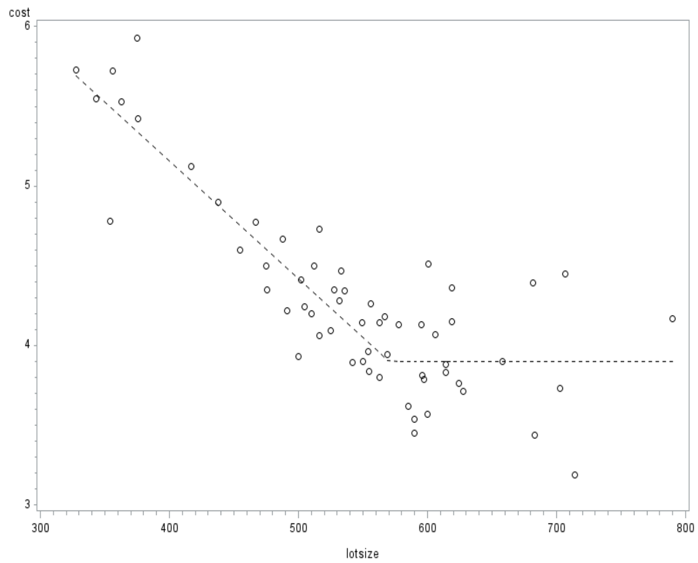
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	8.11879	0.28134	28.86	<.0001
lotsize1	1	-0.00740	0.000538	-13.76	<.0001

Diagnostics





Fitted “Curve”



Summary

- Model 2 fits the data a little better for chosen C_x
- Assumes constant variance about regression line. SAS template provides code to allow variance about each segment of the regression line to vary.
- Current SEs do not take into account uncertainty in C_x
- Could consider Bayesian approach with prior on C_x
- Could consider also consider bootstrapping to quantify uncertainty

Bootstrap

- **Very** important theoretical development that has had a major impact on applied statistics
- Uses resampling to approximate the necessary sampling distribution
- Sample <u>with</u> replacement from the observed data or residuals to generate “new” data sets of same size
- Analyze each data set to get the distribution of interest
- CI based on the quantiles of this sampling distribution if not too skewed or biased. Alternative methods available when this occurs.

Example - KNNL Problem 4.12

- Regression with no intercept
- Small data set and some concerns regarding assumptions
- Let’s compare CI for the slope given by Proc Reg (assumes Normal errors) and that provided by bootstrapping

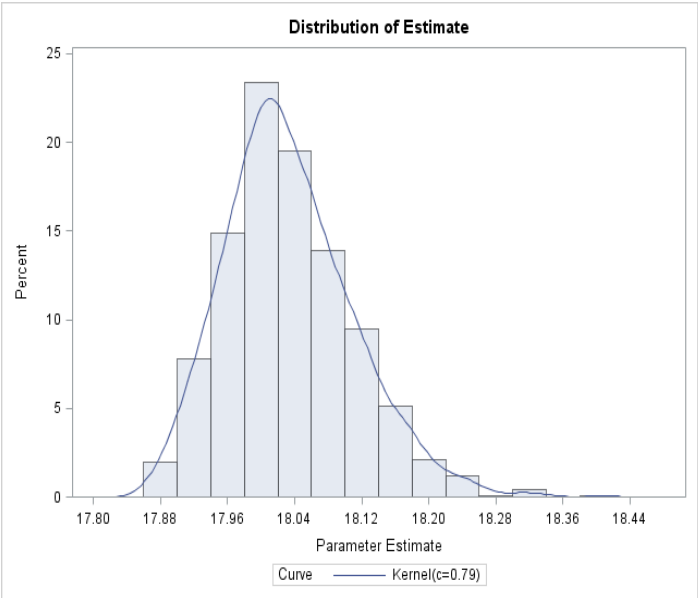
```
--------------------------------------------------------------------
Output from Proc Reg
                     Parameter       Standard
Variable     DF       Estimate          Error     t Value    Pr > |t|
x             1       18.02830        0.07948      226.82      <.0001
Variable     DF       95% Confidence Limits
x             1       17.85336       18.20325
--------------------------------------------------------------------
```

SAS Commands

```
** First we create a data set that contains 1000 copies of the original
   data set and associated fitted values from the Proc Reg analysis;
data pred; set a2;
  do sample=1 to 1000; output;  keep sample x y pred; end;
proc sort data=pred; by sample;

** Next, we randomly sample (with replacement) the residuals, again
   creating 1000 copies;
proc surveyselect data=a2 method=urs sampsize=12 rep=1000 outhits out=res;
   id res;
run;

** Now we merge them and add the residuals to the fitted values;
data new;
 merge pred res;
 ynew = pred + res;
run;
```


SAS Commands

```
* Perform regression on each sample data set and store parameter
  estimate results in a dataset called parm.  The ods html turns
  off the output going into the Results window.;

ods html close;
proc reg; model ynew=x / noint;
 by sample; ods output ParameterEstimates=parm;
ods html;

* Generate histogram and approximate the density;

proc univariate noprint data=parm; var Estimate;
 histogram Estimate / kernel ;
 output out=a4 mean=bmean std=bsterr pctlpre=perc_ pctlpts=2.5,5,95,97.5;
proc print data=a4; run;
```

Results

```
Obs   bmean    bsterr    perc_2_5   perc_5    perc_95   perc_97_5
1   18.0332  0.076763    17.9033   17.9173   18.1728    18.2072
```

- Bootstrap mean (18.0332) close to the regression estimate (18.0283) so there is little bias
- Distribution somewhat skewed suggesting t interval not appropriate
- Percentile CI : (17.9033, 18.2072)
- Reflection CI : (17.8494, 18.1533)
- Both CIs reflect asymmetry in the sampling distribution but go about addressing this asymmetry in different ways

Background Reading

- KNNL Section 11.2-11.6
- knnl435.sas
- KNNL Chapter 16