STOR 455
STATISTICAL METHODS I

Jan Hannig

Homework Comments

- The test/CI for standard deviation does not use T but chi-square.

Simple Linear Regression Model (Section 3.3)

- $Y_i = \beta_0 + \beta_1 X_i + \xi_i$
  - Y_i is the value of the response variable for the i^{th} case
  - X_i is the value of the explanatory variable for the i^{th} case
  - ξ_i are independent normally distributed random errors with mean 0 and variance σ^2
- Parameters
  - β_0 the intercept
  - β_1 the slope
  - $\sigma^2 = var(\xi_i)$ the variance of the error term

Least Squares Solution

$$b_1 = \frac{SXY}{SXX}$$

$$b_0 = \overline{Y} - b_1 \overline{X}$$

$$SXY = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$SSX = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- $(\overline{X}, \overline{Y})$ on fitted regression equation.

Estimation of σ^2

$$s^2 = \frac{\sum\left(\mathrm{Y_i} - \hat{\mathrm{Y}}_\mathrm{i}\right)^2}{n-2} = \frac{\sum \hat{e}_\mathrm{i}^{\,2}}{n-2} = \frac{SSE}{df_E} = MSE$$

$$s = \sqrt{s^2} = \text{Root } MSE$$

Alternative expression for SSE

$$SSE = SSY - \frac{(SXY)^2}{SSX}$$

SMSA Data

- Info about 440 most populous counties between 1990 and 1992
- From US Bureau of the Census
- Variables: id, area, total population, percent of city population, percent of population 65 or older, number of physicians, number of hospital bed, percent high school graduates, total labor, total income, total crime, region (NE, NC, S, W)

Do it in SAS

```
/* SMSA data set, see Appendix C.2 */
data smsa;
  infile 'T:\...\APPENC02.TXT';
  input id county$ state$ area tp pyon
  pold phy bed cri highsch bach pov
  unemp pcinc tinc reg;
proc print data=smsa;
run;
```

Do it in SAS

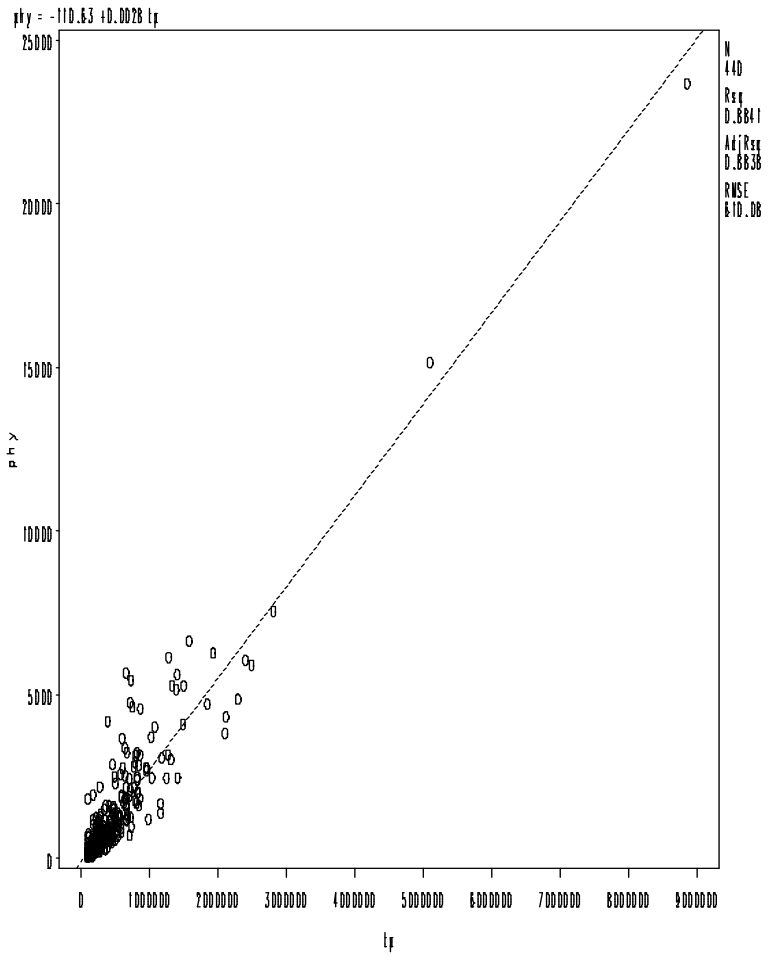
```
                                                                              h
    c                                 i
    o     s                            g         u  p
    u     t    a      p  p              h  b      n   c    t
O   n      a    r       y   o   p   b    c   s   a   p   e   i    i r
b  it      t    e   t    o   l   h    e    r   c   c   o   m   n    n e
s  dy       e    a   p    n   d   y    d    i   h   h   v   p   c    c  g

 1  1 Los_Ange CA  4060 8863164 32.1  9.7 23677 27700 688936 70.0 22.3 11.6  8.0 20786 184230 4
 2  2 Cook     IL   946 5105067 29.2 12.4 15153 21550 436936 73.4 22.8 11.1  7.2 21729 110928 2
 3  3 Harris   TX  1729 2818199 31.3  7.1  7553 12449 253526 74.9 25.4 12.5  5.7 19517  55003 3
 4  4 San_Dieg CA  4205 2498016 33.5 10.9  5905  6179 173821 81.9 25.3  8.1  6.1 19588  48931 4
 5  5 Orange   CA   790 2410556 32.6  9.2  6062  6369 144524 81.2 27.8  5.2  4.8 24400  58818 4
 6  6 Kings    NY    71 2300664 28.3 12.4  4861  8942 680966 63.7 16.6 19.5  9.5 16803  38658 1
 7  7 Maricopa AZ  9204 2122101 29.2 12.5  4320  6104 177593 81.5 22.1  8.8  4.9 18042  38287 4
 8  8 Wayne    MI   614 2111687 27.4 12.5  3823  9490 193978 70.0 13.7 16.9 10.0 17461  36872 2
```

SMSA data

- Relation between # of physicians and population, land area, income
- Relation between crime and population
- Is there any regional difference?

Do it in SAS

```
/* Physician vs.
   total population
   */
proc reg
   data=smsa;
   model phy=tp;
   plot (phy)*(tp);
run;
```



Do it in SAS

The REG Procedure
Model: MODEL1
Dependent Variable: phy

Number of Observations Read	440
Number of Observations Used	440

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1243181164	1243181164	3340.06	<.0001
Error	438	163025135	372204		
Corrected Total	439	1406206299			

Root MSE	610.08483	R-Square	0.8841
Dependent Mean	987.99773	Adj R-Sq	0.8838
Coeff Var	61.74962		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-110.63478	34.74602	-3.18	0.0016
tp	1	0.00280	0.00004837	57.79	<.0001

Do it in SAS

```
/* Physician vs. land area, output residual */
proc reg data=smsa noprint;
  model phy = area;
  output out=temp p=yhat r=residual;
run;

data temp1;
  set temp;
  rsq = residual**2;
run;
proc print data = temp1;
  var phy area yhat residual rsq;
run;
```

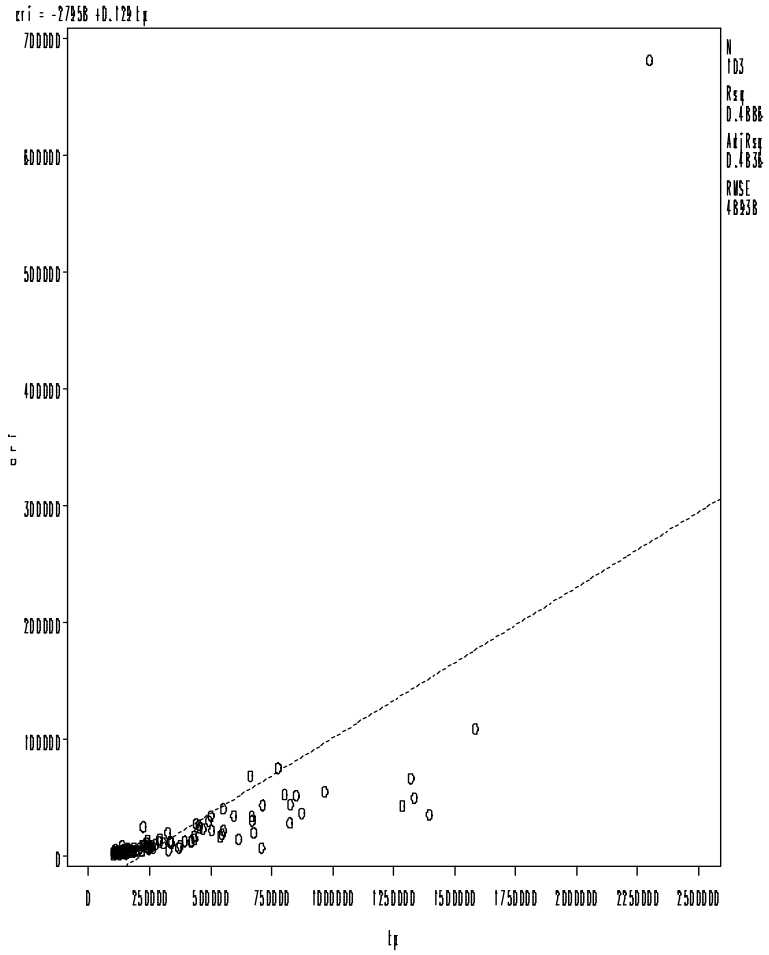
Do it in SAS

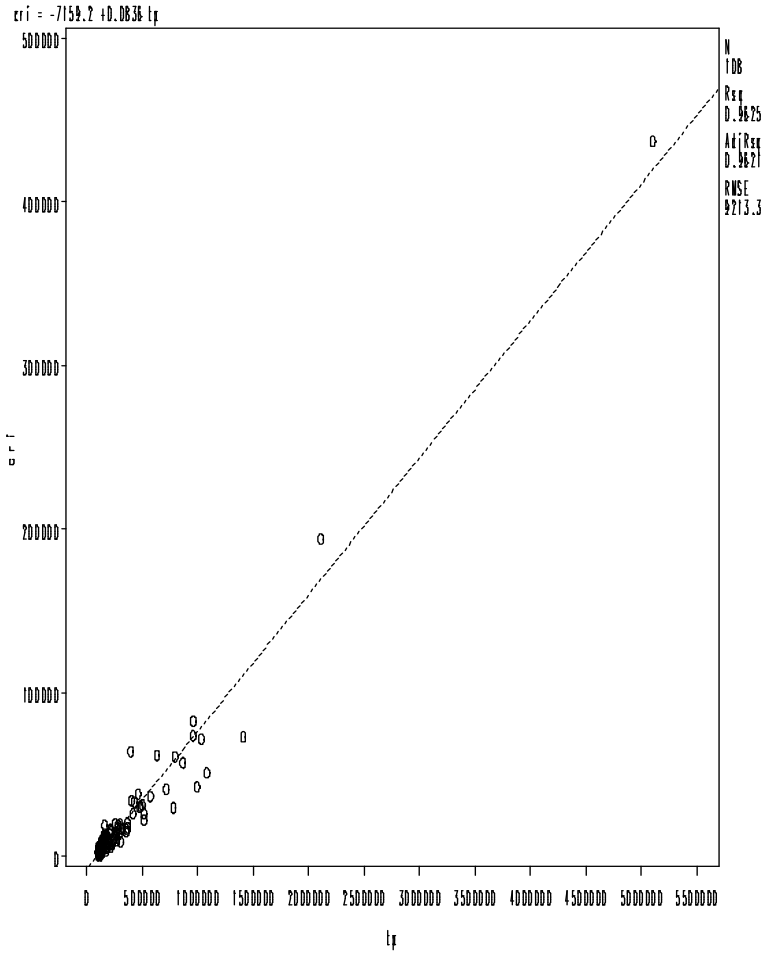
Obs	phy	area	yhat	residual	rsq
1	23677	4060	1260.14	22416.86	502515600.83
2	15153	946	979.40	14173.60	200891054.23
3	7553	1729	1049.99	6503.01	42289169.48
4	5905	4205	1273.21	4631.79	21453452.57
5	6062	790	965.33	5096.67	25976028.81
6	4861	71	900.51	3960.49	15685482.99
7	4320	9204	1723.90	2596.10	6739733.03
8	3823	614	949.46	2873.54	8257207.88
9	6274	1945	1069.46	5204.54	27087223.49
10	4718	880	973.45	3744.55	14021687.68
11	6641	135	906.28	5734.72	32887016.81

Do it in SAS

```
/* Crime vs. total population, NE region */
proc reg data=smsa;
    where reg = 1;
     model cri= tp;
     plot ( cri)*( tp)
;
run;


/* Crime vs. total population, NC region */
proc reg data=smsa;
    where reg = 2;
     model cri= tp;
     plot ( cri)*( tp)
;
run;
```

Do it in SAS


crr = -27958 +0.129 tp
N
103
Rsq
0.4886
Adj Rsq
0.4836
RMSE
48938
crr
tp


crr = -7154.2 +0.0836 tp
N
108
Rsq
0.9625
Adj Rsq
0.9621
RMSE
9213.3
crr
tp

Do it in SAS

Model: MODEL1
Dependent Variable: cri

Number of Observations Read	103
Number of Observations Used	103

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.311376E11	2.311376E11	96.51	<.0001
Error	101	2.418884E11	2394934612		
Corrected Total	102	4.73026E11			

Root MSE	48938	R-Square	0.4886
Dependent Mean	23086	Adj R-Sq	0.4836
Coeff Var	211.98132		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-27958	7088.63230	-3.94	0.0001
tp	1	0.12895	0.01313	9.82	<.0001

Do it in SAS

Model: MODEL1
Dependent Variable: cri

Number of Observations Read	108
Number of Observations Used	108

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.309087E11	2.309087E11	2720.28	<.0001
Error	106	8997722792	84884177		
Corrected Total	107	2.399064E11			

Root MSE	9213.26095	R-Square	0.9625
Dependent Mean	21781	Adj R-Sq	0.9621
Coeff Var	42.30003		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-7159.19264	1045.87028	-6.85	<.0001
tp	1	0.08360	0.00160	52.16	<.0001

Change of Units

- Fahrenheit = 32 +(9/5).Celsius
- What happens to the regression line if we change units?
  - $Y^*=c+d.Y \quad X^*=a+b.X$
  - The regression function
    $\mu_Y(x) = \beta_0 + \beta_1 x, \ \mu_{Y^*}(x^*) = \beta^*_0 + \beta^*_1 x^*$
  - $\beta^*_1=(d/b).\beta_1, \beta^*_0=c+(d/b).(b.\beta_0-a.\beta_1), \sigma^*=|d|\sigma$
  - Same for estimator
- Read Conversation 3.4

Diagnostics (Section 3.5)

- Diagnostics: look at the data to check the assumptions of our model
- Focus on graphic methods, formal tests also discussed
  - Graph the variable themselves
  - Scatter plot, Residual plots
  - Rankit – normal quantile plot

Look at the data

- Before trying to describe the relationship between a response variable (Y) and an explanatory variable (X), we should look at the distributions of these variables
- We should always look at X
- If Y depends on X, looking at Y alone may not be very informative

Diagnostics for X

- Distribution of X
  - outliers, influential cases
  - range and concentration
  - variance, skewness
- Time dependence and dependence on other variables

Graphical tools for X

- Boxplot
- dot plot, stem-leaf plot
- Q-Q plot
- Time sequence plot

Do it in SAS

```
data score;
input x;
cards;
21.5
16
…
20.5
16
;

proc univariate plot data=score;
run;
```

- **proc univariate** plot:

  Summary statistics, stem-leaf, box, and QQ plots.

Do it in SAS

The UNIVARIATE Procedure
Variable: x

Moments

N	33	Sum Weights	33
Mean	19.0606061	Sum Observations	629
Std Deviation	3.12439388	Variance	9.76183712
Skewness	0.179907	Kurtosis	-1.3149512
Uncorrected SS	12301.5	Corrected SS	312.378788
Coeff Variation	16.3918916	Std Error Mean	0.54388716

Basic Statistical Measures

Location		Variability	
Mean	19.06061	Std Deviation	3.12439
Median	18.50000	Variance	9.76184
Mode	15.50000	Range	10.00000
		Interquartile Range	5.50000

Do it in SAS

Tests for Location: Mu0=0

Test	-Statistic-	-----p Value------
Student's t	t 35.04515	Pr > |t| <.0001
Sign	M 16.5	Pr >= |M| <.0001
Signed Rank	S 280.5	Pr >= |S| <.0001

Quantile	Estimate
100% Max	24.5
99%	24.5
95%	24.0
90%	23.5
75% Q3	21.5
50% Median	18.5
25% Q1	16.0
10%	15.5
5%	15.0
1%	14.5
0% Min	14.5

Do it in SAS

The UNIVARIATE Procedure
Variable: x
Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
14.5	28	23.5	3
15.0	13	23.5	19
15.0	5	23.5	27
15.5	32	24.0	6
15.5	30	24.5	15

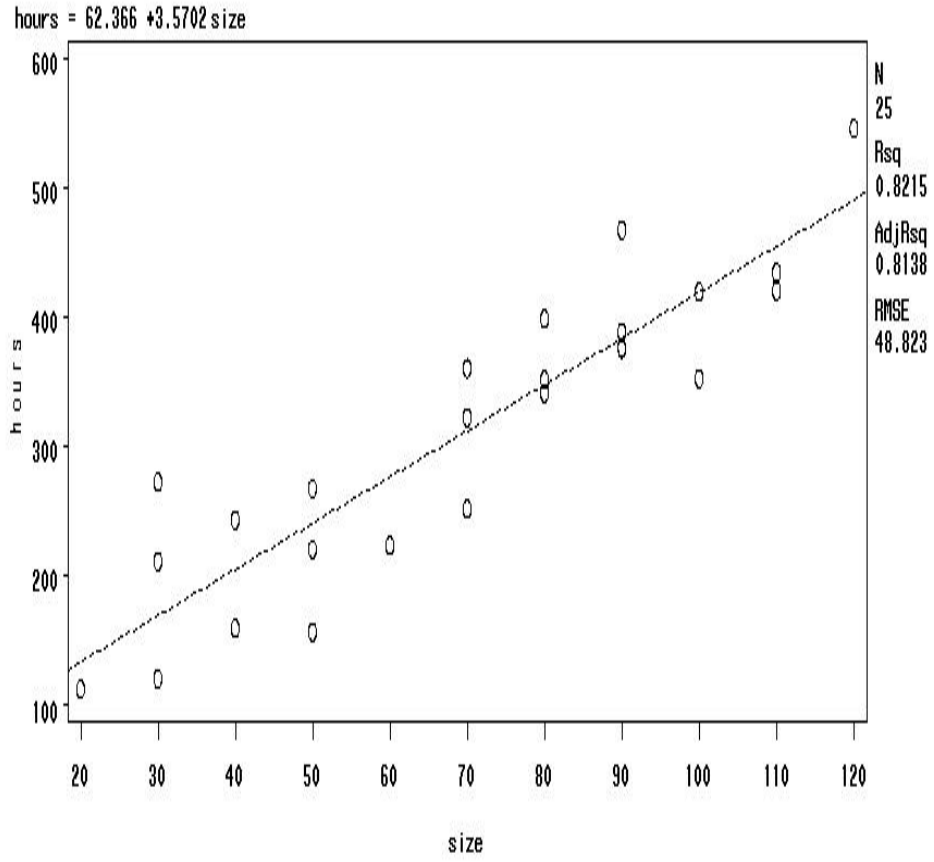
Stem	Leaf	#	Boxplot
24	05	2	
23	555	3	
22	005	3	
21	0555	4	+-----+
20	05	2	| |
19	0	1	| + |
18	05555	5	*-----*
17	555	3	| |
16	00	2	+-----+
15	0055555	7	|
14	5	1	|

Toluca Example

- Toluca Company try to find out the relationship between lot size and labor hours needed to produce the lot

Do it in SAS

```
data lot;
   infile 'T:\...\CH01TA01.TXT';
   input size hours;
run;
proc print data=lot;
proc reg data=lot;
      model hours=size;
      plot hours*size;
run;
```



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	252378	252378	105.88	<.0001
Error	23	54825	2383.71562		
Corrected Total	24	307203			

Root MSE	48.82331	R-Square	???
Dependent Mean	312.28000	Adj R-Sq	0.8138
Coeff Var	15.63447		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	62.36586	26.17743	2.38	0.0259
size	1	3.57020	0.34697	10.29	<.0001

Diagnostics for residuals

- Model: $Y_i = \beta_0 + \beta_1 X_i + \xi_i$
- Predicted values: $\hat{Y}_i = b_0 + b_1 X_i$
- Residuals: $e_i = Y_i - \hat{Y}_i$
  - The book recommends to standardize the residuals.

$$h_{i,i} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SSX} \qquad r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{i,i}}}$$

Residuals

- The e_i should be similar to the ξ_i
  - The model assumes ξ_i iid N(0, σ)
- Similarly the r_i should be similar to the ξ_i/σ

Residuals

- Is the relationship linear?
- Is the variance a constant?
- Are there outliers?
- Are the errors normal?
- Are the errors dependent?

Is the Relationship Linear?

- Scatter plot of Y vs X
- Scatter plot of e vs X
- Plot of e vs X emphasize deviations from linear pattern

Do it in SAS

```
/* Residual plots */
symbol1 v=dot h=.8 c=blue;

/* simulated data, is it linear?
   */
Data resid;
  do x=1 to 30;
   y=x*x-10*x+30+25*normal
    (0);
   output;
  end;
proc print data=resid;
run;
```

Obs	x	y
1	1	20.268
2	2	20.292
3	3	-11.644
4	4	33.722
5	5	-46.357
6	6	-6.092
7	7	-13.902
8	8	13.392
9	9	26.473
10	10	39.391
11	11	16.491
12	12	56.712
13	13	18.588
14	14	57.087
…		

Do it in SAS

```
proc reg data=resid
    noprint;
model y=x;
plot y*x r.*x student.*x;
run;
```