STOR 455
STATISTICAL METHODS I

Jan Hannig

Exam 1

- Results are on blackboard
- I should have printouts and hardcopies to return on Tuesday.
- Grading Scale
  - 90+ A
  - 80+ B
  - 70+ C
  - 60+ D

Inference for β$_1$

$$b_1 \sim N(\beta_1, \sigma^2(b_1))$$

$$\text{where } \sigma^2(b_1) = \sigma^2 \Big/ \sum (X_i - \overline{X})^2$$

$$t = (b_1 - \beta_1) / s(b_1)$$

$$\text{where } s(b_1) = \sqrt{s^2 \Big/ \sum (X_i - \overline{X})^2}$$

$$t \sim \mathrm{t}(n-2)$$

Confidence Interval for β_1

- $b_1 \pm t^* s(b_1)$
- where $t^* = t(1-\alpha/2;n-2)$, the upper $(1-\alpha/2)$ 100 percentile of the t distribution with n-2 degrees of freedom
- $1-\alpha$ is the confidence level

Significance tests for β$_1$

$$H_0 : \beta_1 = 0 \ \text{ vs } \ H_1 : \beta_1 \neq 0$$

$$t = (b_1 - 0) / s(b_1)$$

$$\text{Reject } \ H_0 \text{ if } |t| \geq t^*, t^* = t(1 - \alpha/2, n-2)$$

$$p - value = \mathrm{P}(|\mathrm{T}| > |t|), \text{ where } \ \mathrm{T} \sim t(n-2)$$

The book discourages tests in favor of CIs

Inference for β_0

$$b_0 \sim N(\beta_0, \sigma^2(b_0))$$

$$\text{where}\ \sigma^2(b_0) = \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum(X_i - \overline{X})^2}\right]$$

$$t = (b_0 - \beta_0)/s(b_0)$$

$$\text{for}\ s(b_0)\ \text{replace}\ \sigma^2\ \text{by}\ s^2$$

$$t \sim \mathrm{t}(n-2)$$

Confidence Interval for β_0

- $b_0 \pm t^* s(b_0)$
- where $t^* = t(1-\alpha/2;n-2)$, the upper $(1-\alpha/2)$ 100 percentile of the t distribution with n-2 degrees of freedom
- $1-\alpha$ is the confidence level

Significance tests for β$_0$

$$H_0 : \beta_0 = 0 \text{ vs } H_a : \beta_0 \neq 0$$

$$t = (b_0 - 0)/s(b_0)$$

$$\text{Reject } H_0 \text{ if } |t| \geq t^*, t^* = t(1 - \alpha/2, n - 2)$$

$$P = \text{Prob } (|\text{z}| > |t|), \text{where } z \sim t(n - 2)$$

Point Estimation of μ_{Yh}

- $\mu_{Yh} = \beta_0 + \beta_1 X_h$, the mean value of Y for the subpopulation with $X = X_h$
- Point estimate of μ_{Yh}: $\hat{Y}_h = b_0 + b_1 X_h$
- Unbiased: $E(\hat{Y}_h) = \mu_{Yh}$

Inference for $E(Y_h)$

- Estimate $\sigma^2(\hat{Y}_h)$ by

$$s^2(\hat{Y}_h) = s^2\left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum (X_i - \overline{X})^2}\right]$$

- $$t = \frac{\hat{Y}_h - E(Y_h)}{s(\hat{Y}_h)} \sim t(n-2)$$

Inference for $Y_{h(new)}$

- Estimate prediction variance by:

$$s^2(\text{pred}) = s^2\left[1 + \frac{1}{n} + \frac{\left(X_h - \bar{X}\right)^2}{\sum\left(X_i - \bar{X}\right)^2}\right]$$

- t-distribution:

$t = (Y_{h(new)} - \hat{Y}_h)/s(\text{pred}) \sim t(n-2)$

Confidence band for regression line

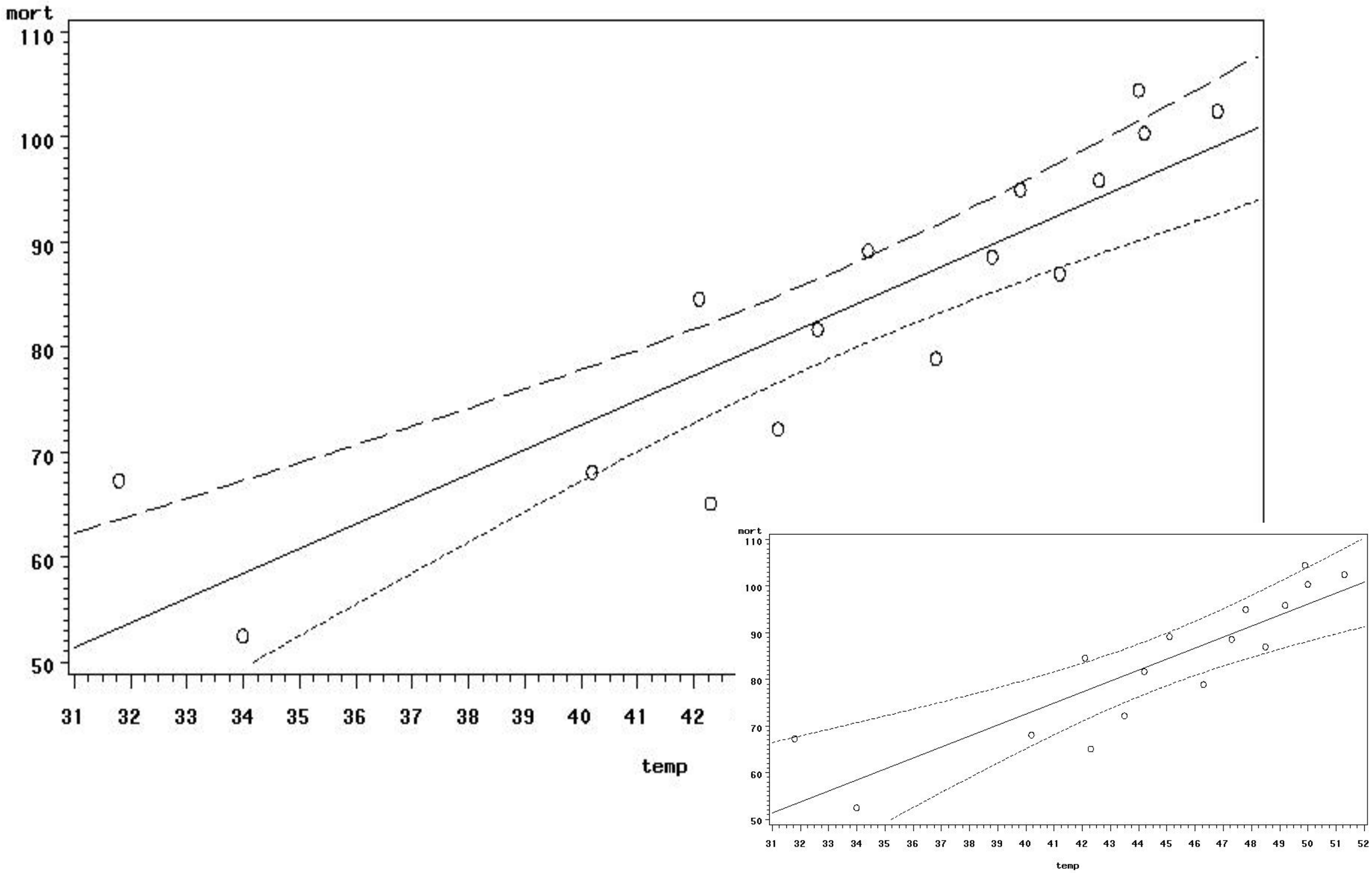
- Working-Hotelling CB: $\hat{Y}_h \pm Ws(\hat{Y}_h)$
- where $W^2=2F(1-\alpha; 2, n-2)$
- This gives intervals for *all* X_h
- CI narrower when X_h close to $\overline{X}$

Do it in SAS

```
/*plot confidence band */
symbol1 v=circle i=rlclm99;
proc gplot data=breastcancer;
   plot mort*temp;
run;

symbol1 v=circle i=rlclm95;
proc gplot data=breastcancer;
   plot mort*temp;
run;
```

95% and 99% Confidence band



Example: Breast Cancer

- What's the relationship between mean annual temperature and the mortality rate for a type of breast cancer in women? The subjects from regions of Great Britain, Norway, and Sweden.
- Mortality: Mortality index for neoplasms of the female breast
- Temperature: Mean annual temperature (in degrees F)
- The Data (http://www.ncsec.org/cadre2/team6_2/modelII.pdf)

Mort	Temp
102.5	51.3
104.5	49.9
100.4	50.0
95.9	49.2

……

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2599.53358	2599.53358	45.67	<.0001
Error	14	796.90580	56.92184		
Corrected Total	15	3396.43938			

Root MSE	7.54466	R-Square	0.7654
Dependent Mean	83.34375	Adj R-Sq	0.7486
Coeff Var	9.05246		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	99% Confidence Limits	
Intercept	1	-21.79469	15.67190	-1.39	0.1860	-68.44747	24.85809
temp	1	2.35769	0.34888	6.76	<.0001	1.31913	3.39626

ANalysis Of VAriance (ANOVA) [Section 3.8]

- Total (corrected) sum of squares in Y is

$$SSTO=\Sigma(Y_i - \overline{Y})^2$$

- Partition of SSTO into:
  - – Regression sum of square SSR
  - – Error sum of square SSE
- SSTO=SSR+SSE

Total Sum of Squares

- If ignoring X_h, predict $E(Y_h)$ use $\overline{Y}$
- SSTO is the sum of squared deviations from this predictor, SSTO= $\Sigma(Y_i - \overline{Y})^2$
- SAS uses **Corrected Total** for SSTO
- Uncorrected total: ΣY_i^2
- "Corrected" means subtract the mean before squaring

Total Sum of Squares

- df_{Total} = n-1
- MST = SSTO/df_{Total} (sample variance)
- MST measures the variability of Y if there are no explanatory variables

Regression Sum of Squares

- $SSR = \Sigma(\hat{Y}_i - \overline{Y})^2$
- $df_R = 1$ (number of explanatory variable)
- SAS call it model sum of square (SSM)
- $MSR = SSR/df_R$

Error Sum of Squares

- $SSE = \Sigma(Y_i - \hat{Y}_i)^2$
- $df_E = df_{Total} - df_R = n-2$
- $MSE = SSE/df_E$
- MSE is an estimate of the variance of residual e_i
- $MSE = s^2$

ANOVA Table

Source	df	SS	MS
Regression	1	$\Sigma(\hat{Y}_i - \overline{Y})^2$	SSR/df_R
Error	n-2	$\Sigma(Y_i - \hat{Y}_i)^2$	SSE/df_E
Total	n-1	$\Sigma(Y_i - \overline{Y})^2$	$SSTO/df_T$

Expected Mean Squares

- MSR, MSE are random variables
- $E(MSR) = \sigma^2 + \beta_1^2 \Sigma(X_i - \overline{X})^2$
- $E(MSE) = \sigma^2$
- When $H_0 : \beta_1 = 0$ is true

  $E(MSR) = E(MSE)$

F test

- F=MSR/MSE ~ $F(df_R, df_E) = F(1, n-2)$
- When H_0: $\beta_1=0$ is false, MSR tends to be larger than MSE
- We reject H_0 when F is large

  $F \geq F(1-\alpha, df_R, df_E) = F(.95, 1, n-2)$
- In practice we use P values

Breast cancer example

Number of Observations Read 16
Number of Observations Used 16
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	???	???	???	???
Error	14	???	56.9		
Corrected Total	15	3396			

- What's the F-value? What's the distribution of the F statistics? P-value?

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2599.53358	2599.53358	45.67	<.0001
Error	14	796.90580	56.92184		
Corrected Total	15	3396.43938			

Root MSE	7.54466	R-Square	0.7654
Dependent Mean	83.34375	Adj R-Sq	0.7486
Coeff Var	9.05246		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	99% Confidence Limits	
Intercept	1	-21.79469	15.67190	-1.39	0.1860	-68.44747	24.85809
temp	1	2.35769	0.34888	6.76	<.0001	1.31913	3.39626

F test and t test

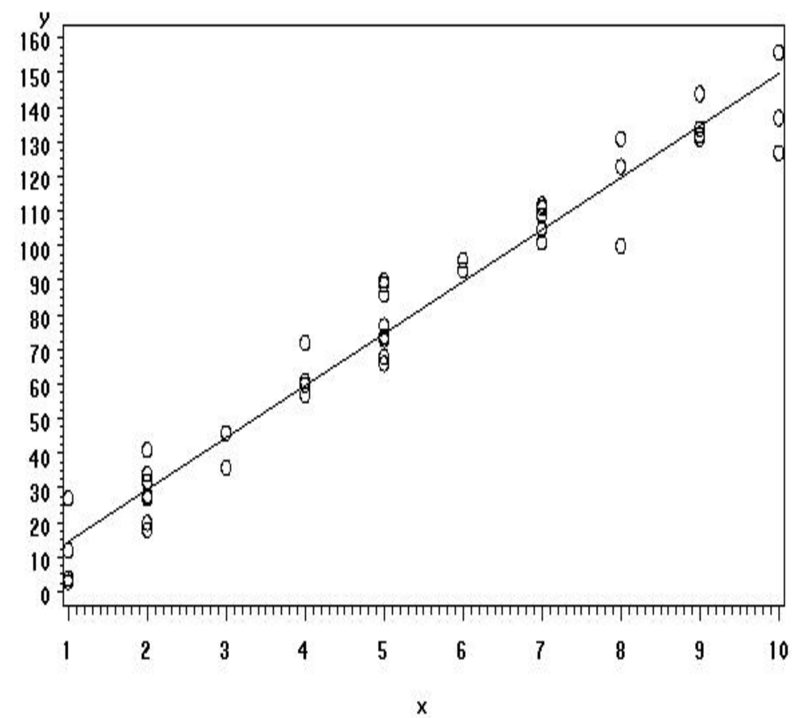
- When H_0: β_1=0 is false, F has a <u>*noncentral*</u> F distribution
- This can be used to calculate power
- Recall $t = b_1/s(b_1)$ tests $H_0 : \beta_1=0$
- It can be shown that $t^2 = F$
- Two approaches give same P-value

Example: Copier maintenance

- Routine preventive maintenance service
- X: number of machines serviced
- Y: number of minutes spent
- How much you should charge for one more machine that needs service?

Do it in SAS

```
/* Copier maintenance data */
data copier;
    infile 'CH01PR20.txt';
    input y x;
symbol1 v=circle i=rl;
proc gplot data=copier;
     plot y*x;
run;
proc reg data=copier;
    model y=x;
run;
```



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	76960	76960	???	???
Error	43	3416.37702	79.45063		
Corrected Total	44	80377			

Root MSE	8.91351	R-Square	0.9575
Dependent Mean	76.26667	Adj R-Sq	0.9565
Coeff Var	11.68729		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.58016	2.80394	-0.21	0.8371
x	1	15.03525	0.48309	31.12	<.0001

R^2 (Section 3.9)

- R^2 = SSR/SSTO = 1 – SSE/SSTO
- $100 * R^2$ = percentage of variation in the response variable explained by the explanatory variable

Pearson Correlation

- r: the usual correlation coefficient
- A number between –1 and +1
- Measures the strength of the *linear* relationship between two variables

$$r = \frac{\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum\left(X_i - \overline{X}\right)^2 \sum\left(Y_i - \overline{Y}\right)^2}}$$

Pearson Correlation

- Notice that

$$r = b_1 \sqrt{\frac{\sum (X_i - \overline{X})^2}{\sum (Y_i - \overline{Y})^2}}$$

$$= b_1 (SXX / SYY)^{1/2}$$

- Test H_0: β_1=0 similar to H_0: r = 0

R^2 and r^2

$$r^2 = \mathbf{b}_1^2 \left(\frac{\sum \left(\mathbf{X}_i - \overline{\mathbf{X}} \right)^2}{\sum \left(\mathbf{Y}_i - \overline{\mathbf{Y}} \right)^2} \right)$$

$$= \mathbf{SSR} / \mathbf{SSTO}$$

- Ratio of explained and total variation

R^2 and r^2

- We use R^2 when the number of explanatory variables is arbitrary (simple and multiple regression)
- $r^2 = R^2$ only for simple regression
- R^2 is often multiplied by 100 and thereby expressed as a percent

NBA Salary Example

- Data collected by Steven Couper from the web.
- Variable: salary, ppg, cppg.

Obs	salary	ppg	cppg
1	5.4	9.9	18.2
2	7.5	18.8	15.4
3	12.1	17.1	16.8
4	1.2	2.8	2.8
…			
204	15.1	24.6	22.3

Do it in SAS

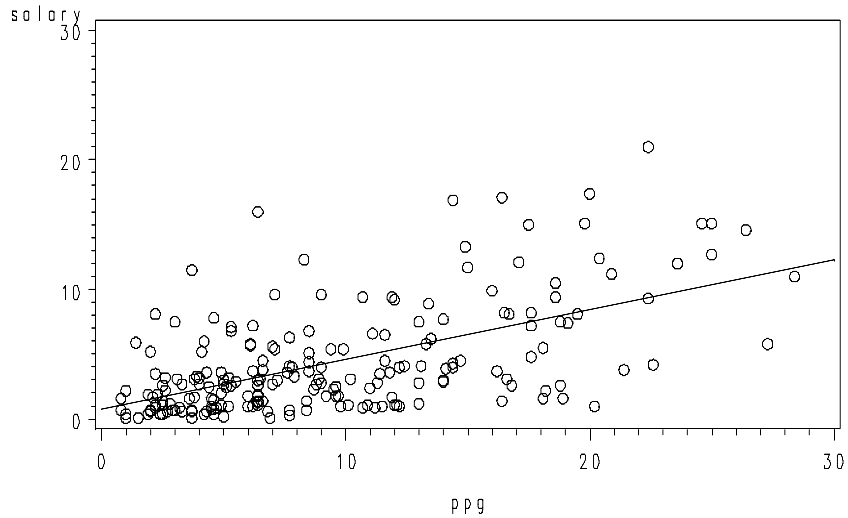
- Import: File->Import Data, then follow instruction.
- Alternative:

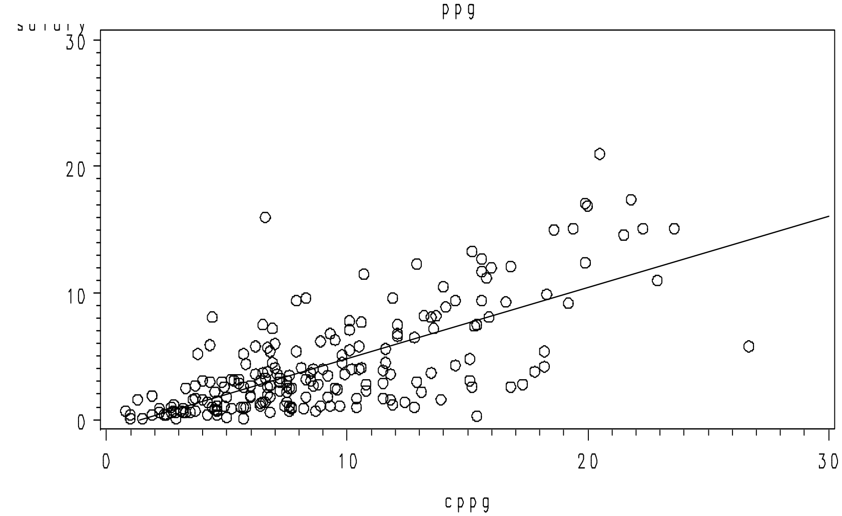
```
PROC IMPORT OUT= WORK.NBAPPG
        DATAFILE= "T:\....steve.nbappg.xls"
        DBMS=EXCEL REPLACE;
    SHEET="Sheet1$";
    GETNAMES=YES;
    MIXED=NO;
    SCANTEXT=YES;
    USEDATE=YES;
    SCANTIME=YES;
RUN;
```

Do it in SAS

```
symbol1 v=circle i=rl;
proc gplot data=nbappg;
    plot salary*ppg
    salary*cppg;
run;

proc reg data=nbappg;
    model salary=ppg;
     model salary=cppg;
run;
```





Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1223.97682	1223.97682	109.25	<.0001
Error	204	2285.57697	11.20381		
Corrected Total	205	3509.55379			

Root MSE	3.34721	R-Square	0.3488
Dependent Mean	4.34757	Adj R-Sq	0.3456
Coeff Var	76.99029		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.71525	0.41852	1.71	0.0890
ppg	ppg	1	0.38688	0.03701	10.45	<.0001

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1764.82673	1764.82673	206.35	<.0001
Error	204	1744.72705	8.55258		
Corrected Total	205	3509.55379			

Root MSE	2.92448	R-Square	0.5029
Dependent Mean	4.34757	Adj R-Sq	0.5004
Coeff Var	67.26696		

Parameter Estimates

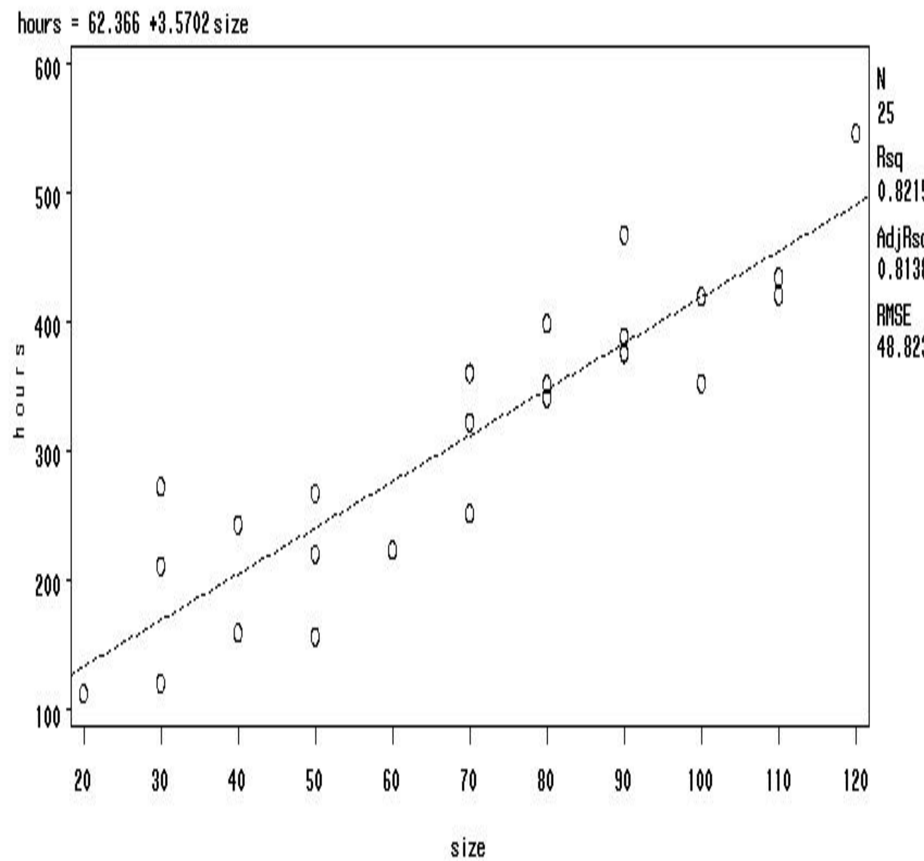
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-0.77038	0.41043	-1.88	0.0619
cppg	cppg	1	0.56208	0.03913	14.36	<.0001

Toluca Example

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

Do it in SAS

```
data lot;
  infile '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