STOR 455
STATISTICAL METHODS I

Jan Hannig

Total Sum of Squares

- 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
- 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

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

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 *<u>linear</u>* relationship between two variables

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

R^2 and r^2

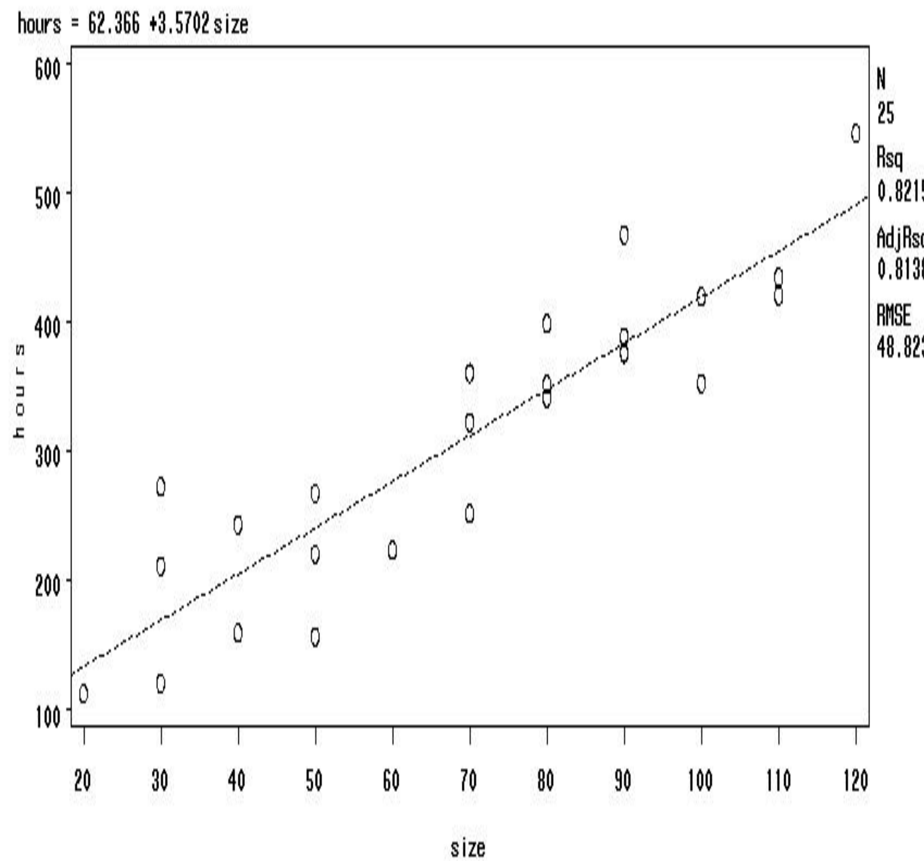
- We use R^2 when the number of explanatory variables is arbitrary (simple and multiple regression)
- r^2=R^2 only for simple regression
- R^2 is often multiplied by 100 and thereby expressed as a percent

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

Error in variables (Section 3.10)

- Sometimes the values of Y and X are observed with error (measured by some instrument – e.g., X is temperature)
  - $Y_o=Y+\xi$, $X_o=X+\zeta$
  - errors assumed independent normal
- The measurement error in Y can be folded into the regression error
  - all estimators and CIs remain valid with exception of σ_Y and $\rho_{Y|X}$
- The measurement error in X is more serious but still can be folded into the regression error for simple linear regression
  - Most estimators and CIs remain unaffected. No valid CIs exists for Y, $\mu_Y(x)$, σ_Y and $\rho_{Y|X}$
- A quantity cannot be estimated if no direct measurements are made or assumed.

Regression through the origin (Section 3.11)

- Assume that $\beta_0=0$ (the regression line goes through [0,0]).
- The formulas are given in the book (page 210)
  - Notice that d.f.=n-1 (why?)
- In SAS use
  - model y =x / noint;
- DO NOT USE THIS WITHOUT A GOOD REASON
  - E.g., if the regression curve is not strictly linear but must go through the origin you should include intercept.

Why REGRESSION?

- Remember in SLR
  - b_1=SXY/SSX=r $(SSY/SSX)^{1/2}$= r S_Y/S_X
  - b_0= $\overline{Y}$ - b_1 $\overline{X}$
  - The equation

$$\hat{Y} = b_0 + b_1 X = \bar{Y} + r\frac{S_y}{S_X}(X - \bar{X})$$

- Remark
  - If predictor is at the mean of X, response is predicted at the mean of Y
  - If the predictor is 1 sd above the mean of X then the response is predicted r sd above the mean of Y
  - Regression toward the mean

Studios Example

- Dwaine Studios currently operates in 21 medium size cities, specialized in portraits of children.
- They want to expand to other cities.
- Want to know: relationship between sales, young population, and disposable income

Generalize SLR to MLR:

- Y is sales in thousands of dollars
- X_1 is the target population
- X_2 is per capita disposable income
- More than one predictor variables correlated with response variable
- Need to generalize Simple Linear Regression to Multiple Linear Regression

Multiple Regression (Chapter 4)

- More than one predictor.
- Relationship still linear
- Questions
  - How to fit
  - How to spot departures from assumptions
  - How to select important predictors

Observables in MLR (Section 4.2)

- For each population item we observe p+1 variables Y, $X_1, \ldots, X_p$
  - Y is the response (only one response value per item)
  - $X_1, \ldots, X_p$ are the predictors (multiple predictors per item)
- Two possible assumptions
  - $(Y, X_1, \ldots, X_p)$ are jointly normal
  - $X_1, \ldots, X_p$ are fixed and every subpopulation $Y | X_1, \ldots, X_p$ is normal

Graphical and numerical summaries for individual and pairs of observables

- Histogram (proc univariate)
- Scatter plot (SAS Macro scatter.sas)
- Mean, s.d., min, max (proc means)
- Correlation (proc corr)

Do It in SAS

```
*Data shown on page 237 of the OPTIONAL
  textbook - file CH06FI05.txt;
data studios;
  input x1 x2 y;
  x1x2=x1*x2;
  label x1='targtpop'
        x2='dispoinc';
cards;
  68.5  16.7  174.4
  45.2  16.8  164.4
  91.3  18.2  244.2
  ...

  52.3  16.0  166.5
  ;

run;
```

Do it in SAS

```
* Descriptive statistics;
proc means data=studios;
run;
* Check correlation;
proc corr data = studios;
run;
proc univariate data = studios
  noprint;
var y x1 x2;
histogram y x1 x2;
run;
```

Do it in SAS

```
* Making scatter
  plot using
  macro;
%include "T:\...
  \Macro
  \scatter.sas";
%scatter(data =
  studios, var = y
  x1 x2);
```