Multiple Linear Regression &

General Linear Model in *SAS*

Multiple linear regression is used to model the relationship between one numeric outcome or response or dependent variable (Y), and several (multiple) explanatory or independent or predictor or regressor variables (X). When some predictors are categorical variables, we call the subsequent regression model as the General Linear Model.

```
Data <-read.csv("http://www.math.uah.edu/stat/data/Galton.csv", header = T)
y<-data$Height
x1<-data$Father
x2<-data$Mother
x3<-as.numeric(data$Gender)-1
```

Multiple regression using the **lm()** function

```
mod<-lm(y ~ x1+x2+x3)
summary(mod)

par(mfrow=c(2,2))
plot(mod)

x3<- data$Gender
mod1<-glm(y~x1+x2+factor(x3))
summary(mod1)

x3<-relevel(factor(x3),ref="M")

mod1<-glm(y~x1+x2+factor(x3))
summary(mod1)

library(xlsx)

data1<-read.xlsx("d:/ams394/galton/heat.xlsx", 1)

library(leaps)

attach(data1)

leaps1<-regsubsets(Y~X1+X2+X3+X4,data=data1,nbest=10)

Summary(leaps1)

step(lm(Y~X1+X2+X3+X4), data=data1)

summary(step(lm(Y~X1+X2+X3+X4), data=data1))
```

```
/* simple linear regression */
proc reg;
model y = x;
/* multiple regression */
proc reg;
model y = x1 x2 x3;
```

Here are some print options for the model phrase:

```
model y = x / noint;  /* regression with no intercept */
model y = x / ss1;    /* print type I sums of squares */
model y = x / p;      /* print predicted values and residuals */
model y = x / r;      /* option p plus residual diagnostics */
model y = x / clm;    /* option p plus 95% CI for estimated mean */
model y = x / cli;    /* option p plus 95% CI for predicted value */
model y = x / r cli clm; /* options can be combined */
```

The CLM option adds confidence limits for the mean predicted values. The CLI option adds confidence limits for the individual predicted values.

It is possible to let SAS do the predicting of new observations and/or estimating of mean responses. The way to do this is to enter the x values (or x1,x2,x3 for multiple regression) you are interested in during the data input step, but put a period (.) for the unknown y value.

```
data new;
input x y;
datalines;
1 0
2 3
3 .
4 3
5 6
;
run;
proc reg;
model y = x / p cli clm;
```

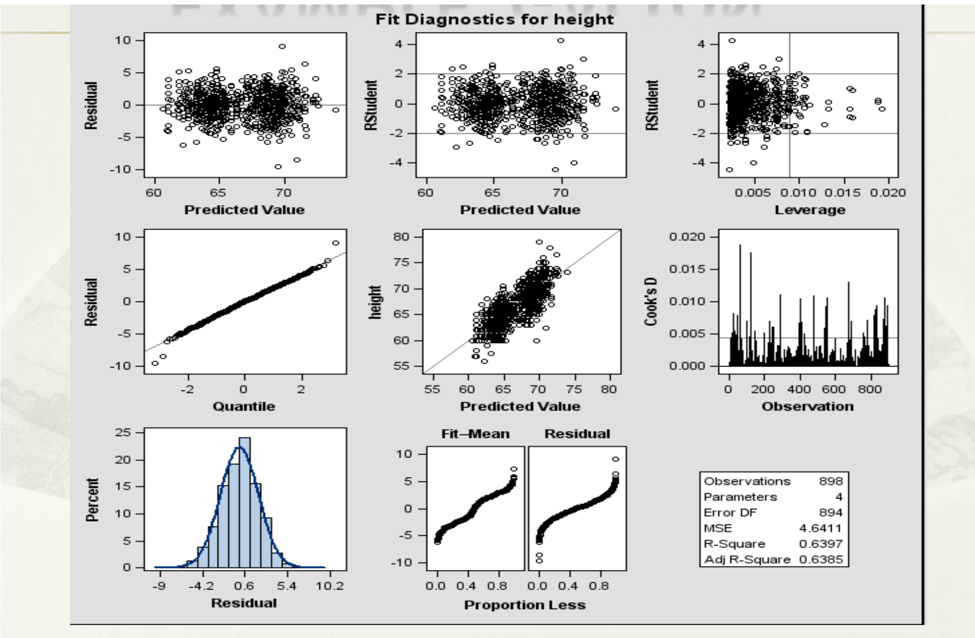
```
Data Galton;
Input Family $ Father Mother Gender $ Height Kids;
Datalines;
1       78.5    67      M       73.2    4
1       78.5    67      F       69.2    4
1       78.5    67      F       69      4
1       78.5    67      F       69      4
2       75.5    66.5    M       73.5    4
2       75.5    66.5    M       72.5    4
2       75.5    66.5    F       65.5    4
    2       75.5    66.5    F       65.5    4
…
;
Run;

data revise;
set Galton;
if Gender = 'F' then sex = 1.0;
else sex = 0.0;
run;
proc reg data=revise;
title "proc reg; Dependence of Child Heights on Parental Heights";
model height = father mother sex / vif collin;
run;
quit;
```

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	7365.90034	2455.30011	529.03	<.0001
Error	894	4149.16204	4.64112		
Corrected Total	897	11515			

Root MSE	2.15433	R-Square	0.6397
Dependent Mean	66.76069	Adj R-Sq	0.6385
Coeff Var	3.22694		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	20.57071	2.74067	7.51	<.0001	0
Father	1	0.40598	0.02921	13.90	<.0001	1.00607
Mother	1	0.32150	0.03128	10.28	<.0001	1.00660
sex	1	-5.22595	0.14401	-36.29	<.0001	1.00188



Proc GLM
Alternatively, one can use proc GLM procedure that can incorporate the categorical variable(sex) directly via the class statement

```
proc glm data=Galton;
Class gender;
model height = father mother gender;
        run;
        quit;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Father	1	896.716584	896.716584	193.21	<.0001
Mother	1	490.217369	490.217369	105.62	<.0001
Gender	1	6111.965365	6111.965365	1316.92	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	20.57071133	B	2.74066703	7.51	<.0001
Father	0.40597803		0.02920696	13.90	<.0001
Mother	0.32149514		0.03128178	10.28	<.0001
Gender F	-5.22595131	B	0.14400791	-36.29	<.0001
Gender M	0.00000000	B	.	.	.

Example:
The following table shows data on the heat evolved in calories during the handending of cement on a per gram basis(y) along with the percentages of four ingredients: tricalcium aluminate(x1), tricalcium silicate(x2),tetracalcium alumino ferrite(x3), and dicalcium silicate(x4)

No.	X1	X2	X3	X4	Y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	1159
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

```
data example115;
input x1 x2 x3 x4 y;
datalines;
 7 26  6 60  78.5
 1 29 15 52  74.3
11 56  8 20 104.3
11 31  8 47  87.6
 7 52  6 33  95.9
11 55  9 22 109.2
 3 71 17  6 102.7
 1 31 22 44  72.5
 2 54 18 22  93.1
21 47  4 26 115.9
 1 40 23 34  83.8
11 66  9 12 113.3
10 68  8 12 109.4
;
run;
proc reg data=example115;
   model y = x1 x2 x3 x4 /selection=stepwise;
run;
```

Selected SAS output

The REG Procedure
Model: MODEL1
Dependent Variable: y

Stepwise Selection: Step 4

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	52.57735	2.28617	3062.60416	528.91	<.0001
x1	1.46831	0.12130	848.43186	146.52	<.0001
x2	0.66225	0.04585	1207.78227	208.58	<.0001

Bounds on condition number: 1.0551, 4.2205

* All variables left in the model are significant at the 0.1500 level.
* No other variable met the 0.1500 significance level for entry into the model.
* Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x4		1	0.6745	0.6745	138.731	22.80	0.0006
2	x1		2	0.2979	0.9725	5.4959	108.22	<.0001
3	x2		3	0.0099	0.9823	3.0182	5.03	0.0517
4		x4	2	0.0037	0.9787	2.6782	1.86	0.2054

Best subsets regression & SAS

```
proc reg data=example115;
   model y = x1 x2 x3 x4 /selection=ADJRSQ;
run;
```

For the selection option, SAS has implemented 9 methods in total. For best subset method, we have the following options:

- *R^2 Selection (RSQUARE)*
- *Adjusted R^2 Selection (ADJRSQ)*
- *Mallows' C_p Selection (CP)*