Linear Regression with SAS

Linear regression is a method for modeling the relationship between two variables: one independent (x) and one dependent (y). The history and mathematical models behind regression analysis can be found at (http://en.wikipedia.org/wiki/Linear_regression). Scientists are typically interested in getting the equation of the line that describes the best least-squares fit between two datasets. They may also be interested in the coefficient of determination (R^2) which describes the proportion of variability in a data that is accounted for by the linear model.

Let's look at the Example:

Table 10.1 Mileage and Groove Depth of a Car Tire

Mileage (in 1000 miles)	Groove Depth (in mils)
0	394.33
4	329.50
8	291.00
12	255.17
16	229.33
20	204.83
24	179.00
28	163.83
32	150.33

```
x<-Mileage<-c(0,4,8,12,16,20,24,28,32)
y<-Groove_Depth<-c(394.33,329.50,291.00,255.17,229.33,204.83
,179.00,163.83,150.33)
p1<-plot(x,y)
title(main="Scatter Plot", xlab="Mileage(in 1000 miles)",
ylab="Groove Depth(in miles)")
mod<-lm(y~x)
abline(mod)
anova(mod)
par(mfrow=c(2,2))
plot(mod)
summary(mod)
cor.test(x,y)
```

SAS code

```
Data cars;
Input x y;
Datalines;
0   394.33
4   329.50
8   291.00
12   255.17
16   229.33
20   204.83
24   179.00
28   163.83
32   150.33 ;
Run;

proc reg data=cars;
 model y=x;
run;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	**50887**	50887	140.71	<.0001
Error	7	2531.52943	361.64706		
Corrected Total	8	53419			

Root MSE	19.01702	**R-Square**	0.9526
Dependent Mean	244.14667	**Adj R-Sq**	0.9458
Coeff Var	7.78918		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	**1**	360.63667	11.68855	30.85	<.0001
x	**1**	-7.28062	0.61377	-11.86	<.0001

1. A marketing manager conducted a study to determine whether there is a linear relationship between money spent on advertising and company sales. The data are listed in the following table.

Advertising expenses (1000s of \$), x	2.4	1.6	2.0	2.6	1.4	1.6	2.0	2.2
Company sales (1000s of \$), y	225	184	220	240	180	184	186	215

Some summary statistics are as follows: $\sum x = 15.8$, $\sum y = 1634$, $\sum xy = 3289.8$, and $\sum x^2 = 32.44$.

(a) What is the correlation coefficient between these two variables?

(b) Write down the least squares regression equation.

(c) What is the coefficient of determination of your regression?

(d) At $\alpha = 0.01$, is there a significant linear relationship between these two variables?

(e) Suppose a company plans to spend \$1,800 on advertisement, what is the expected sales?

Question 1:

```
DATA Company;
INPUT x y;
DATALINES;
2.4 225
1.6 184
2.0 220
2.6 240
1.4 180
1.6 184
2.0 186
2.2 215
RUN;
PROC CORR DATA = Company OUTP = corr;
    VAR x y;
RUN;
PROC REG DATA = Company ALPHA = 0.01;
    MODEL y = x;
RUN;
```

Selected SAS Output:

Pearson Correlation Coefficients, N = 8
Prob > |r| under H0: Rho=0

	x	y
x	1.00000	0.91291
		0.0015
y	0.91291	1.00000
	0.0015	

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3178.15587	3178.15587	30.01	0.0015
Error	6	635.34413	105.89069		
Corrected Total	7	3813.50000			

Root MSE	10.29032	**R-Square**	0.8334
Dependent Mean	204.25000	**Adj R-Sq**	0.8056
Coeff Var	5.03810		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	**1**	104.06073	18.64622	5.58	0.0014
x	**1**	50.72874	9.25967	5.48	0.0015

Interpretation:

1) The correlation coefficient is 0.91291.
2) The LS regression equation is $\hat{y} = 50.73x + 104.06$.
3) The coefficient of determination is 83.34%.
4) We perform t-test to test linear relationship between the two variables. Null Hypothesis and Alternative Hypothesis for the test:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

The p-value of the t-test is 0.0015. Since $0.0015 < \alpha = 0.01$, we reject the null hypothesis in favor of the alternative and

conclude that there is a significant linear relationship between the two variables. The coefficient of determination is 83.34% also confirms that there is a significant linear relationship between the two variables.

5) Since $y(1.8) = 50.73\,(1.8) + 104.06 = 195.374$, which gives us that the expected sales is 195.374.