STA121: Applied Regression Analysis

Linear Regression Analysis - Chapters 3 and 4 in Dielman

Artin Armagan

Department of Statistical Science

September 15, 2009

Outline

Purpose and Formulation

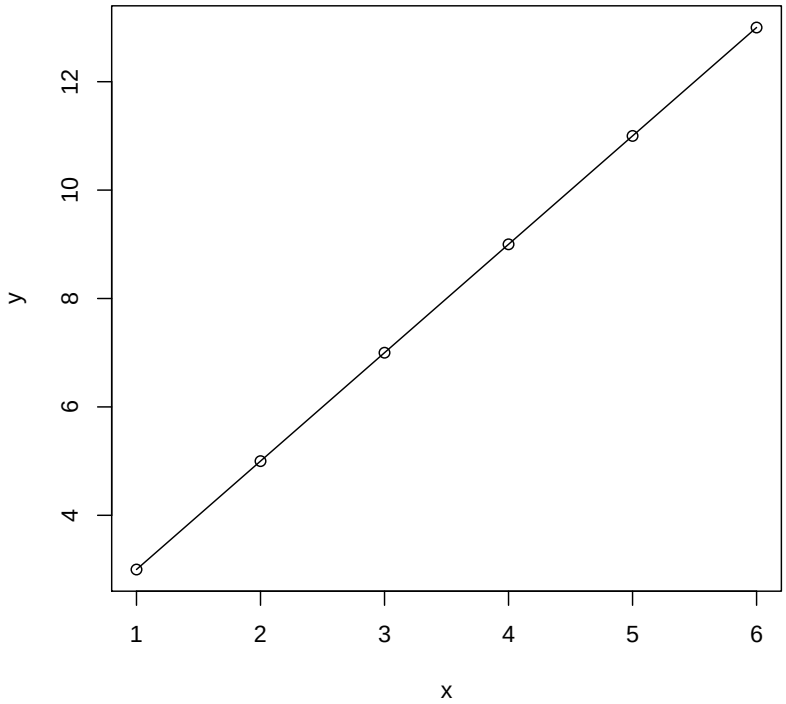
- Regression analysis is a statistical technique used to describe relationships among variables.
- In the simplest case where bivariate data are observed, the simple linear regression is used.
- The variable that we are trying to model is referred to as the *dependent* variable and often denoted by y.
- The variable that we are trying to explain y with is referred to as the *independent* or *explanatory* variable and often denoted by x.
- If a linear relationship between y and x is believed to exist, this relationship is expressed through an equation for a line:

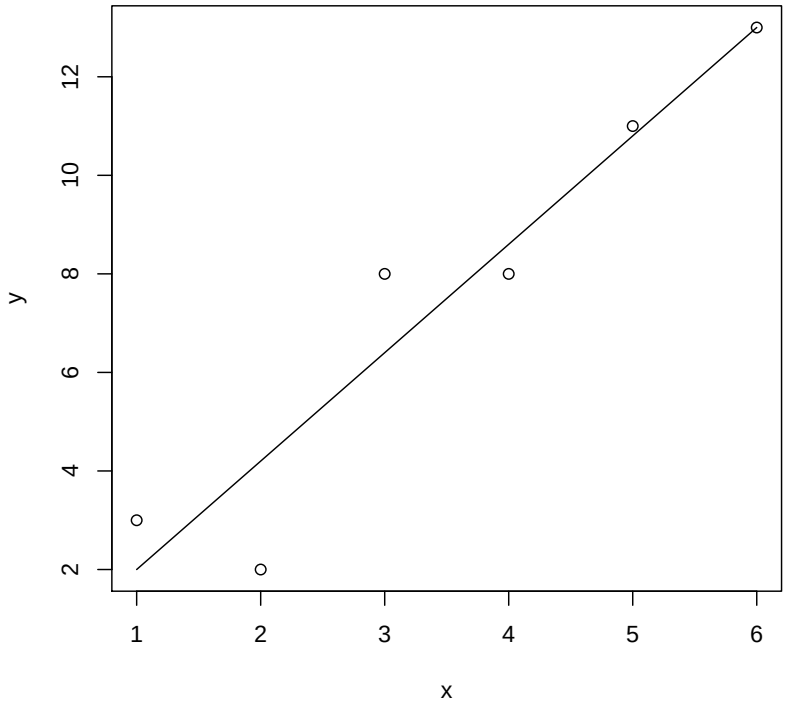
$$y = b_0 + b_1 x$$

Purpose and Formulation

- Above equation gives an *exact* or a *deterministic* relationship meaning there exists no randomness.
- In this case recall that having only two pairs of observations (x, y) would suffice to construct a line.
- However many things we observe have a random component to it which we try to understand through various probability distributions.

Example





$$y = 1 + 2x$$

$$\hat{y} = -0.2 + 2.2x$$

Least Squares Criterion to Fit a Line

- We need to specify a method to find the "best" fitting line to the observed data.
- When we pass a line through the the observations, there will be differences between the actual observed values and the values predicted by the fitted line. This difference at each x value is called a *residual* and represents the "error".
- It is only sensible to try to minimize the total error we make while fitting the line.
- The least squares criterion minimizes the sum of squared errors to fit a line, i.e. $\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

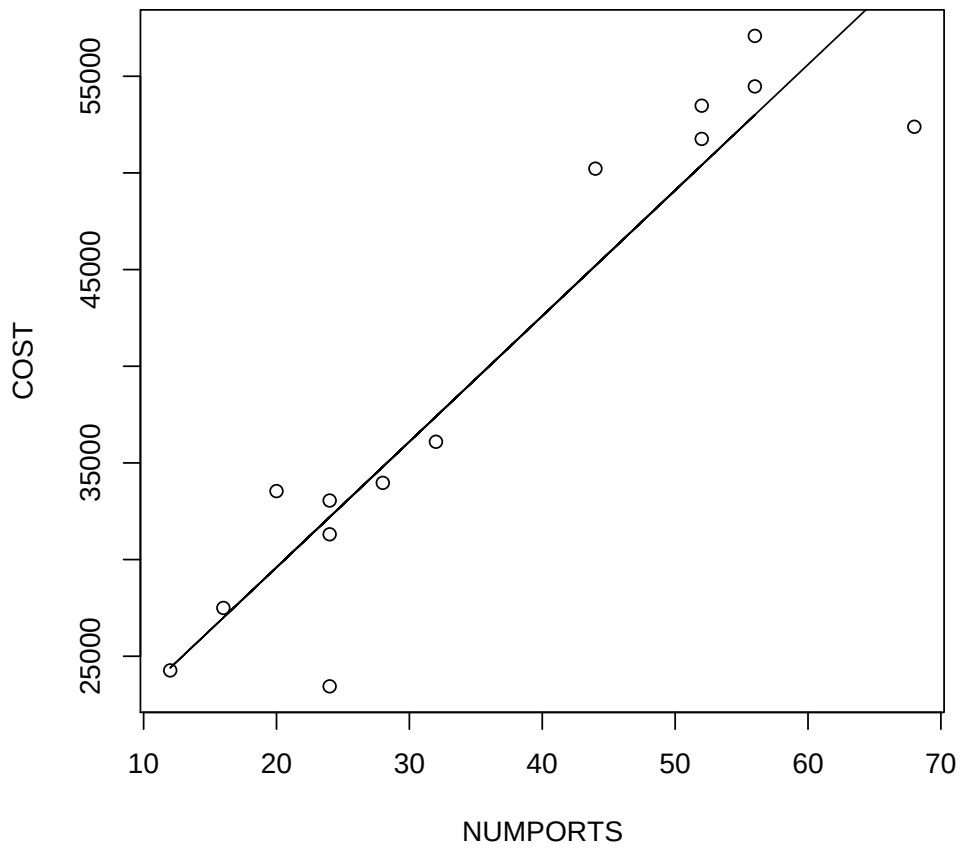
Least Squares Criterion to Fit a Line

This is a simple minimization problem and results in the following expressions for b_0 and b_1:

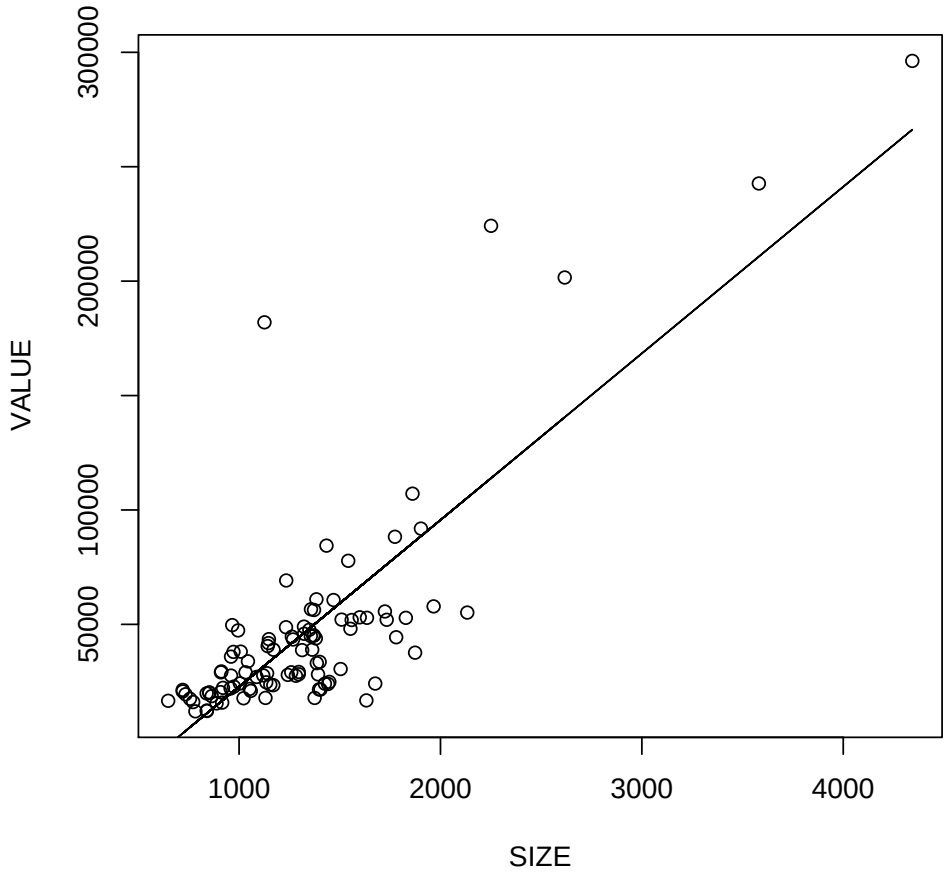
$$\begin{aligned} b_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ b_0 &= \bar{y} - b_1\bar{x} \end{aligned}$$

These are simply obtained by differentiating $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ ($\hat{y}_i = b_0 + b_1 x_i$) with respect to b_0 and b_1 and setting them equal to zero at the solution which leaves us with two equations and two unknowns.

Pricing Communication Nodes



Estimating Residential Real Estate Values



Assumptions

- It is assumed that there exists a linear deterministic relationship between x and the mean of y, $\mu_{y|x}$:

  $$\mu_{y|x} = \beta_0 + \beta_1 x$$

  Since the actual observations deviate from this line, we need to add a noise term giving

  $$y_i = \beta_0 + \beta_1 x_i + e_i.$$

- The expected value of this error term is zero: $E(e_i) = 0$.
- The variance of each e_i is equal to σ_e^2. This assumptions suggests a constant variance along the regression line.
- The e_i are normally distributed.
- The e_i are independent.

Inferences about β_0 and β_1

- The point estimates of β_0 and β_1 are justified by the least squares criterion such that b_0 and b_1 minimize the sum of squared errors for the observed sample.
- It should be also noted that, under the assumptions made earlier, the *maximum likelihood estimator* for β_0 and β_1 is identical to the least squares estimator.
- Recall that a statistic is a function of a sample (which is a realization of a random variable), thus is a random variable itself. b_0 and b_1 have sampling distributions.

Sampling Distribution of b_0

- $E(b_0) = \beta_0$
- $Var(b_0) = \sigma_e^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x}^2)^2} \right)$
- The sampling distribution of b_0 is normal.

Sampling Distribution of b_1

- $E(b_1) = \beta_1$
- $Var(b_0) = \frac{\sigma_e^2}{\sum_{i=1}^{n}(x_i - \bar{x}^2)^2}$
- The sampling distribution of b_1 is normal.

Properties of b_0 and b_1

- b_0 and b_1 are unbiased estimators for β_0 and β_1
- b_0 and b_1 are consistent estimators for β_0 and β_1
- b_0 and b_1 are minimum variance unbiased estimators for β_0 and β_1. That said, they have smaller sampling errors than any other unbiased estimator for β_0 and β_1.

Estimating σ_e^2

- The sampling distributions of b_0 and b_1 are normal when σ_e^2 is known.
- In realistic cases we won't know σ_2^2.
- An unbiased estimate of σ_e^2 is given by

$$s_e^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2} = MSE$$

where $\hat{y}_i = b_0 + b_1 x_i$.

- Substituting s_e for σ_e earlier, s_{b_0} and s_{b_1} can be obtained.

Constructing Confidence Intervals for β_0 and β_1

- Now that σ_e is not known, the sampling distributions of b_0 and b_1 are t, i.e. $\frac{b_0-\beta_0}{s_{b_0}} \sim t_{n-2}$ and $\frac{b_1-\beta_1}{s_{b_1}} \sim t_{n-2}$.
- $(1-\alpha)100\%$ confidence intervals then can be constructed as

$$(b_0 - t_{\alpha/2,n-2}s_{b_0} \quad , \quad b_0 + t_{\alpha/2,n-2}s_{b_0})$$
$$(b_1 - t_{\alpha/2,n-2}s_{b_1} \quad , \quad b_1 + t_{\alpha/2,n-2}s_{b_1}).$$

Hypothesis tests about β_0 and β_1

- Conducting a hypothesis test is no more involved than constructing a confidence interval. We make use of the same pivotal quantity, $\frac{b_1-\beta_1}{s_{b_1}}$ which is t distributed.
- Since we often include the intercept in our model anyway, a hypothesis test on β_0 may be redundant. Our main goal is to see whether there exists a linear relationship between the two variables which is implied by the slope, β_1.
- We first state the null and alternative hypotheses:

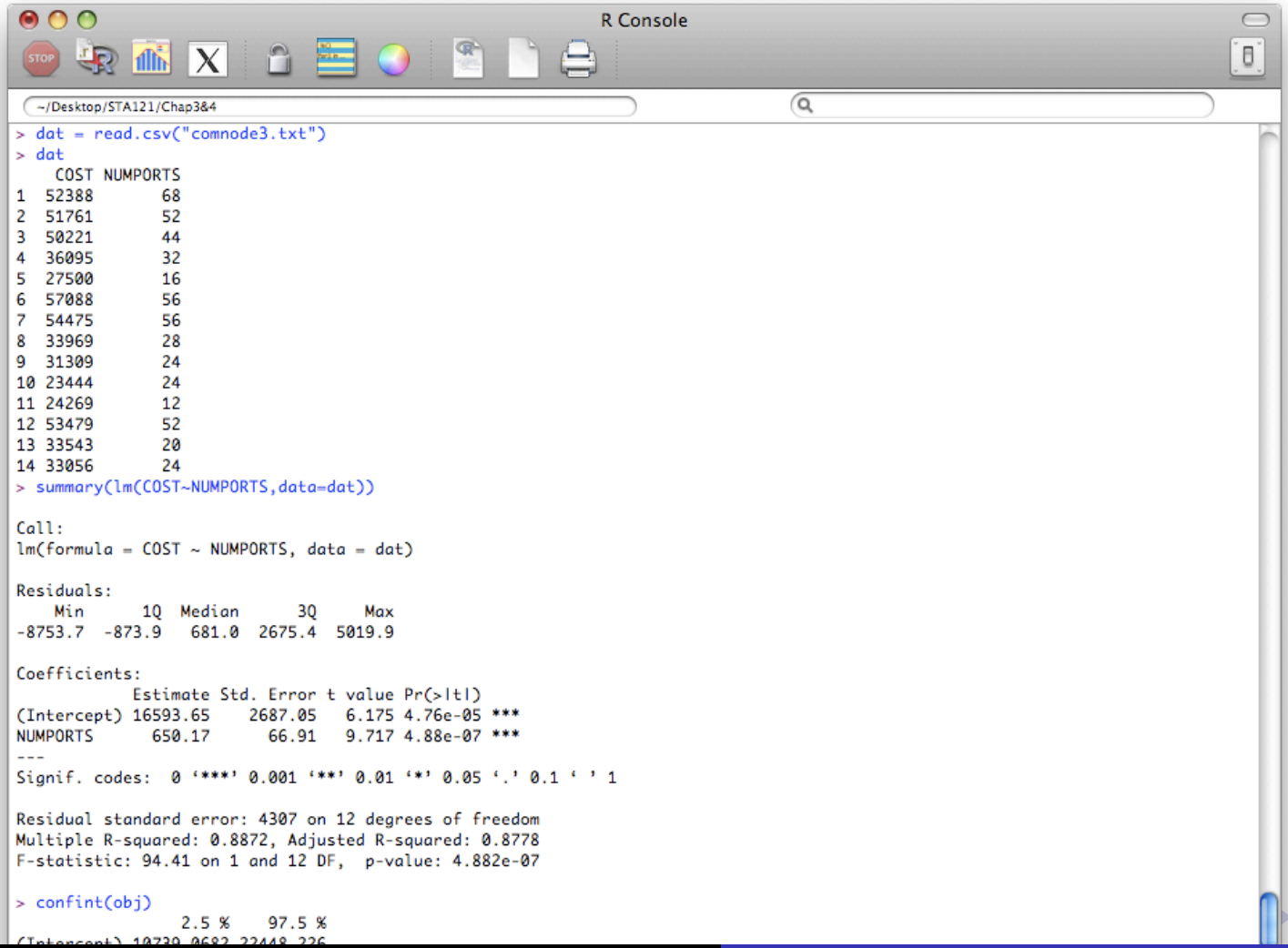
$$\begin{aligned} H_0 &: \beta_1 = (\geq, \leq)\beta_1^* \\ H_a &: \beta_1 \neq (<, >)\beta_1^* \end{aligned}$$

Hypothesis tests about β_0 and β_1

- To test this hypothesis, a t statistic is used, $t = \frac{b_1 - \beta_1^*}{s_{b_1}}$.
- A significance level, α, is specified to decide whether or not reject the null hypothesis.
- Possible alternative hypotheses and corresponding decision rules are

Alternative	**Decision Rule**
$H_a : \beta_1 \neq \beta_1^*$	Reject H_0 if $\lvert t \rvert > t_{\alpha/2,n-2}$
$H_a : \beta_1 < \beta_1^*$	Reject H_0 if $t < -t_{\alpha,n-2}$
$H_a : \beta_1 > \beta_1^*$	Reject H_0 if $t > t_{\alpha,n-2}$

Pricing Communication Nodes

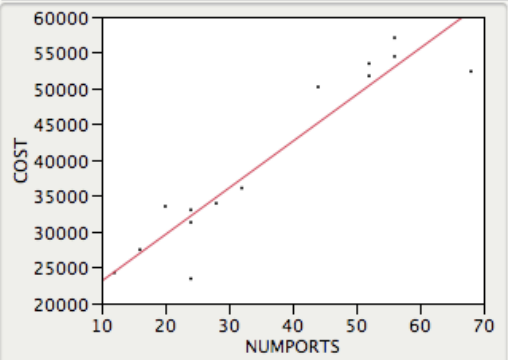



```
> dat = read.csv("comnode3.txt")
> dat
    COST NUMPORTS
1  52388       68
2  51761       52
3  50221       44
4  36095       32
5  27500       16
6  57088       56
7  54475       56
8  33969       28
9  31309       24
10 23444       24
11 24269       12
12 53479       52
13 33543       20
14 33056       24
> summary(lm(COST~NUMPORTS,data=dat))

Call:
lm(formula = COST ~ NUMPORTS, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-8753.7  -873.9   681.0  2675.4  5019.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16593.65    2687.05   6.175 4.76e-05 ***
NUMPORTS      650.17      66.91   9.717 4.88e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4307 on 12 degrees of freedom
Multiple R-squared: 0.8872, Adjusted R-squared: 0.8778
F-statistic: 94.41 on 1 and 12 DF,  p-value: 4.882e-07

> confint(obj)
                2.5 %    97.5 %
(Intercept) 10739.0682 22448.226
```

Pricing Communication Nodes

comnode3: Fit Y by X of COST by NUMPORTS

Bivariate Fit of COST By NUMPORTS



Linear Fit

Linear Fit

COST = 16593.647 + 650.16917*NUMPORTS

Summary of Fit

RSquare	0.887229
RSquare Adj	0.877832
Root Mean Square Error	4306.914
Mean of Response	40185.5
Observations (or Sum Wgts)	14

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1751268376	1.7513e+9	94.4105
Error	12	222594146	18549512	Prob > F
C. Total	13	1973862522		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	16593.647	2687.05	6.18	<.0001*	10739.068	22448.226

Lab Assignment #1

- Work on Exercises 6 and 7 in Chapter 3. You are encouraged to use `R` but may use JMP if you feel more comfortable.
- Due date is September 9.

ANOVA

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1751268376	1.7513e+9	94.4105
Error	12	222594146	18549512	Prob > F
C. Total	13	1973862522		<.0001*

The Coeficient of Determination

- In an exact or deterministic relationship, SSR=SST and SSE=0. This would imply that a straight line could be drawn through each observed value.
- Since this is not the case in real life, we need a a measure of how well the regression line fits the data.
- The coefficient of determination gives the proportion of total variation explained in the response by the regression line and is denoted by R^2.

$$R^2 = \frac{SSR}{SST}$$

The Correlation Coefficient

- For simple linear regression the correlation coefficient is $r = \pm\sqrt{R^2}$.
- This does not apply to multiple linear regression.
- If the sign of r is positive, then the relationship between the variables is direct, otherwise is inverse.
- r ranges between -1 and 1.
- A correlation of 0 merely implies no linear relationship.

The F Statistic

- An additional measure of how well the regression line fits the data is provided by the F statistic, which tests whether the equation $\hat{y} = b_0 + b_1 x$ provides a better fit to the data than the equation $\hat{y} = \bar{y}$.

$$F = \frac{MSR}{MSE}$$

  where $MSR = SSR/1$ and $MSE = SSE/(n-2)$.
- The degrees of freedom corresponding to SSR and SSE add up to the total degrees of freedom, $n - 1$.

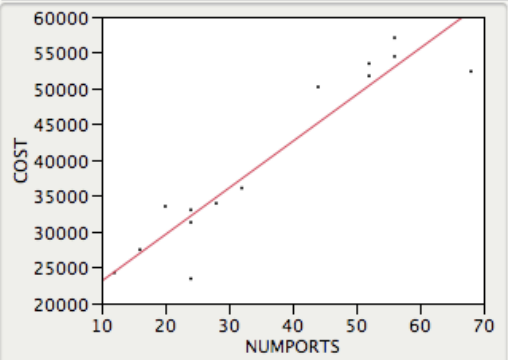
The F Statistic

- To formalize the use of F statistic, consider the hypotheses $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$.
- We reject H_0 if $F > F_{\alpha,1,n-2}$.
- For simple linear regression, $F = \frac{MSR}{MSE} = t^2$.
- Since both a t-test and an F-test will yield the same conclusions, it doesn't matter which one we use.

Pricing Communication Nodes

comnode3: Fit Y by X of COST by NUMPORTS

Bivariate Fit of COST By NUMPORTS



Linear Fit

Linear Fit

COST = 16593.647 + 650.16917*NUMPORTS

Summary of Fit

RSquare	0.887229
RSquare Adj	0.877832
Root Mean Square Error	4306.914
Mean of Response	40185.5
Observations (or Sum Wgts)	14

Lack Of Fit

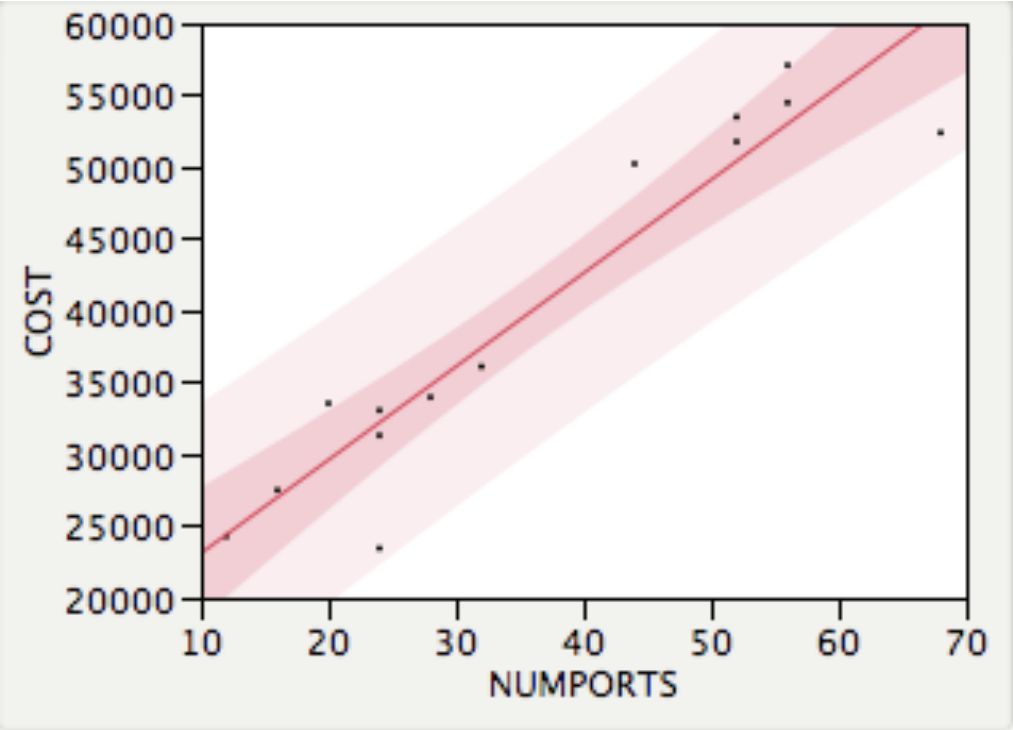
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1751268376	1.7513e+9	94.4105
Error	12	222594146	18549512	Prob > F
C. Total	13	1973862522		<.0001*

Parameter Estimates

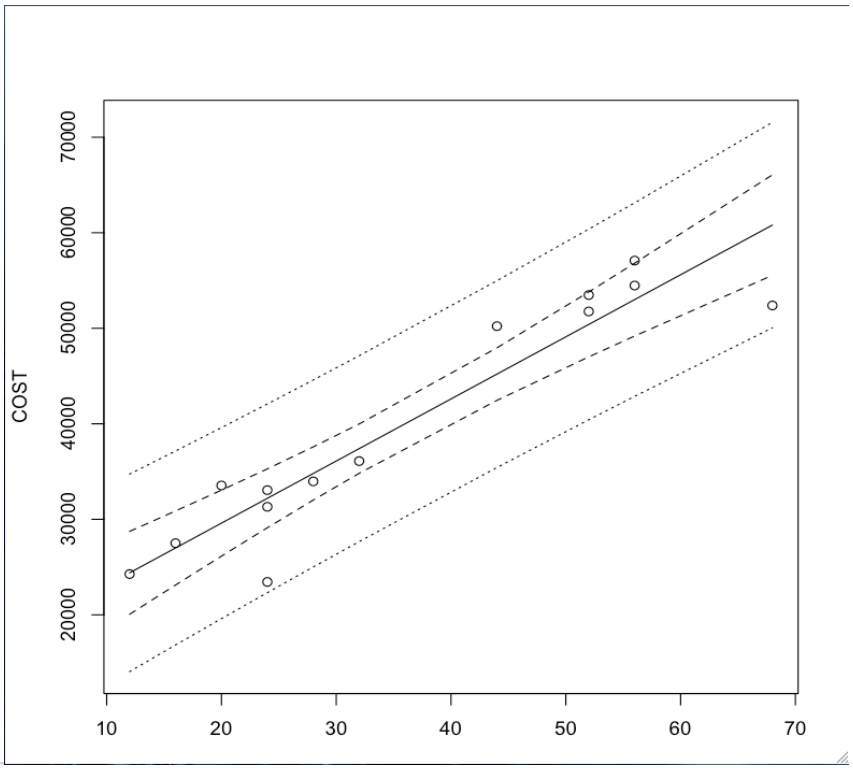
Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	16593.647	2687.05	6.18	<.0001*	10739.068	22448.226

Pricing Communication Nodes



Pricing Communication Nodes

```
>
> dat=read.csv("comnode3.txt")
> lm.comnode=lm(COST~NUMPORTS,data=dat)
> pc=predict(lm.comnode,int="c")
> pp=predict(lm.comnode,int="p")
Warning message:
In predict.lm(lm.comnode, int = "p") :
  Predictions on current data refer to _future_ responses

> plot(dat[,2],pc[,1],type="l",ylim=c(min(pp),max(pp)),ylab="COST",xlab="NUMPORTS")
> lines(sort(dat[,2]),sort(pc[,2]),lty="dashed")
> lines(sort(dat[,2]),sort(pc[,3]),lty="dashed")
> lines(sort(dat[,2]),sort(pp[,2]),lty="dotted")
> lines(sort(dat[,2]),sort(pp[,3]),lty="dotted")
> points(dat[,2],dat[,1])
>
```



What Makes a Prediction Interval Wider?

- The difference arises from the difference between the variation in the mean of y and the variation in one individual y value.
- $\mathbb{V}(\bar{y}_f) = \sigma_e^2 \left(\frac{1}{n} + \frac{(x_f - \bar{x})^2}{(n-1)s_x^2} \right)$
- $\mathbb{V}(y_f) = \sigma_e^2 \left(1 + \frac{1}{n} + \frac{(x_f - \bar{x})^2}{(n-1)s_x^2} \right)$
- Replace σ_e^2 by s_e^2 when the error variance is not known and is to be estimated.

Assessing the Quality of Fit

- The mean square deviation is used commonly.

$$MSD = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n_h}$$

where n_h is the size of the hold-out sample.

Lab Assignment #2

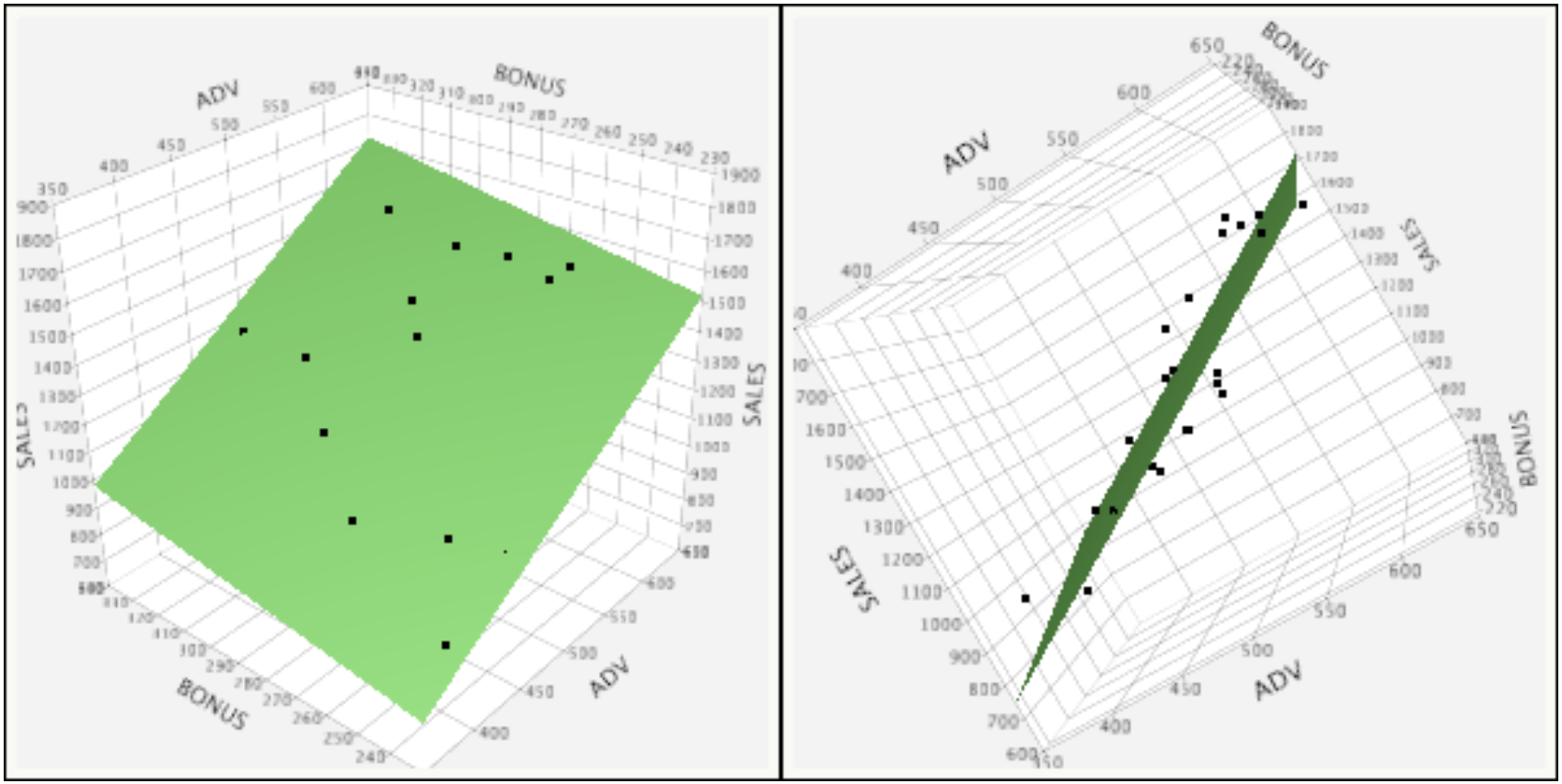
- Work on Exercises 8, 9, 10, 11, 12 and 13 in Chapter 3. You are encouraged to use `R` but may use JMP if you feel more comfortable.
- Due date is September 18.

Formulation

- If a linear relationship between y and a set of xs is believed to exist, this relationship is expressed through an equation for a plane:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + ... + b_p x_p$$

Meddicorp Sales



Assumptions

- Assumptions are the same with simple linear regression model. Thus the population regression equation is written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_p x_{ip} + e_i$$

Inferences About the Population Regression Coefficients

Summary of Fit

RSquare	0.8592
RSquare Adj	0.83104
Root Mean Square Error	93.76972
Mean of Response	1269.02
Observations (or Sum Wgts)	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	1073118.5	268280	30.5114
Error	20	175855.2	8793	Prob > F
C. Total	24	1248973.7		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	-593.5375	259.1959	-2.29	0.0330*	-1134.211	-52.86438
ADV	2.513138	0.314275	8.00	<.0001*	1.8575708	3.1687052
BONUS	1.905948	0.742386	2.57	0.0184*	0.3573588	3.4545372
MKTSHR	2.651007	4.635655	0.57	0.5738	-7.018801	12.320815
COMPET	-0.120731	0.371815	-0.32	0.7488	-0.896324	0.654861

```
> lm.med=lm(SALES~ADV+BONUS+MKTSHR+COMPET,data=dat)
> summary(lm.med)

Call:
lm(formula = SALES ~ ADV + BONUS + MKTSHR + COMPET, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-186.98  -73.97   16.95   55.62  125.52

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -593.5375   259.1959  -2.290   0.0330 *
ADV            2.5131     0.3143   7.997 1.17e-07 ***
BONUS          1.9059     0.7424   2.567   0.0184 *
MKTSHR         2.6510     4.6357   0.572   0.5738
COMPET        -0.1207     0.3718  -0.325   0.7488
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 93.77 on 20 degrees of freedom
Multiple R-squared: 0.8592, Adjusted R-squared: 0.831
F-statistic: 30.51 on 4 and 20 DF,  p-value: 2.937e-08
```

R^2 and R^2_{adj}

- An R^2 value is computed as in the case of simple linear regression.
- Although it has a nice interpretation, it also has a drawback in the case of multiple linear regression.
- Intuitively, R^2 will never decrease as we add more independent variable into the model disregarding the fact that the variables being thrown in to the model may be explaining an insignificant portion of the variation in y. And as far as we can tell, the closer R^2 is to 1, the better.
- This means, we have to somehow account for how many variables we include in our model. In other words, we need to somehow "penalize" for the number of variables included in the model.
- Always remember that, the simpler the model we come up with, the better.

R^2 and R^2_{adj}

- The adjusted R^2 does not suffer from this limitation and it accounts for the number of variables included in the model

$$R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

- Note that, in this case, if a variable is causing an insignificant amount of decrease in SSE, the denominator in the above equation may actually be increasing, leading to a smaller R^2_{adj} value.

F Statistic

- And F statistic is computed in a similar fashion to that of simple linear regression case.
- A different use of the F statistic will come into play with the comparison of nested models in the multiple linear regression case.
- This helps us compare a larger model to a reduced model which comprises a subset of variables included in the full model.
- This statistic is computed for each variable in the model if one uses `anova()` in `R` or looks at "Effect tests" in JMP.

F Statistic

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
ADV	1	1	562259.56	63.9457	<.0001*
BONUS	1	1	57954.64	6.5912	0.0184*
MKTSHR	1	1	2875.57	0.3270	0.5738
COMPET	1	1	927.07	0.1054	0.7488

```
> anova(lm.med)
Analysis of Variance Table

Response: SALES
           Df  Sum Sq Mean Sq  F value    Pr(>F)
ADV         1 1012408 1012408 115.1411 9.546e-10 ***
BONUS       1   55389   55389   6.2994   0.02079 *
MKTSHR      1    4394    4394   0.4997   0.48777
COMPET      1     927     927   0.1054   0.74877
Residuals 20  175855    8793
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F Statistic

- $H_0 : \beta_{l+1} = ... = \beta_p = 0$
  H_a : At least one of $\beta_{l+1}, ..., \beta_p$ is not equal to zero.
- This implies, under the reduced model we have

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_p x_{il} + e_i.$$

- If we want to compare this reduced model to the full model and find out if the reduction was reasonable, we have to compute an F statistic:

$$F = \frac{(SSE_R - SSE_F)/(p - l)}{SSE_F/(n - p - 1)}$$

Meddicorp

- $H_0 : \beta_3 = ... = \beta_4 = 0$
  H_a : At least one of β_3, β_4 is not equal to zero.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	1073118.5	268280	30.5114
Error	20	175855.2	8793	Prob > F
C. Total	24	1248973.7		<.0001*

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	1067797.3	533899	64.8306
Error	22	181176.4	8235	Prob > F
C. Total	24	1248973.7		<.0001*

- $F = \frac{(181176-175855)/2}{175855/20} = 0.303$
- 3.49 is the 5% F critical value with 2 numerator and 20 denominator degrees of freedom. Thus we accept H_0.

Using Conditional Sums of Squares

```
> anova(lm.med)
Analysis of Variance Table

Response: SALES
          Df  Sum Sq Mean Sq  F value    Pr(>F)
ADV        1 1012408 1012408 115.1411 9.546e-10 ***
BONUS      1   55389   55389   6.2994   0.02079 *
MKTSHR     1    4394    4394   0.4997   0.48777
COMPET     1     927     927   0.1054   0.74877
Residuals 20  175855    8793
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Lab Assignment #3

- Work on Exercises 1, 2, 3 and 4 in Chapter 4. You are encouraged to use `R` but may use JMP if you feel more comfortable.
- Due date is September 25.

Multicollinearity

- When explanatory variables are correlated with one another, the problem of multicollinearity is said to exist.
- The presence of a high degree multicollinearity among the explanatory variables result in the following problem:
  - The standard deviations of the regression coefficients are disproportionately large leading to small t-score although the corresponding variable may be an important one.
  - The regression coefficient estimates are highly unstable. Due to high standard errors, reliable estimation is not possible.

Detecting multicollinearity

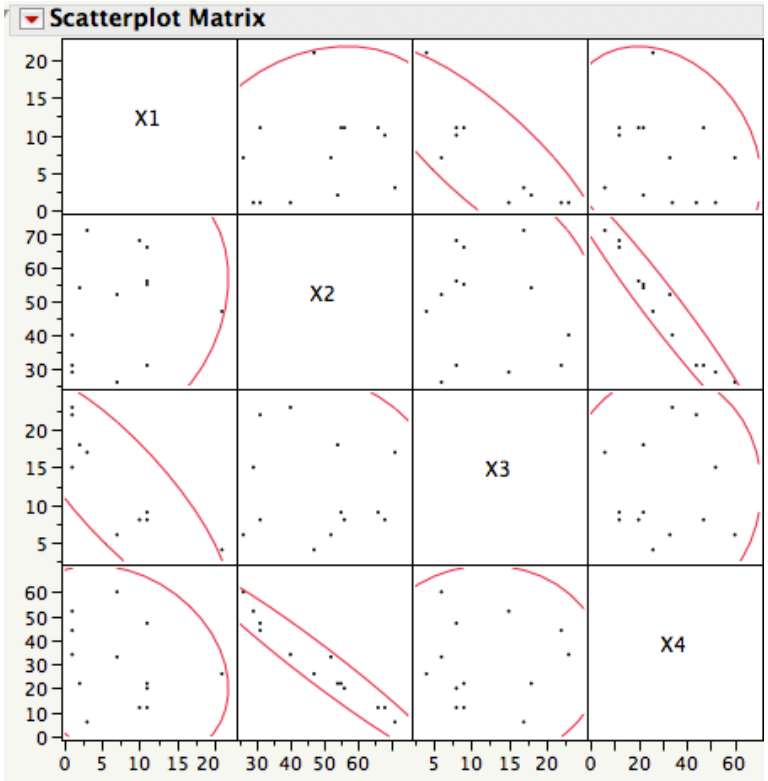
- Pairwise correlations.
- Large F, small t.
- Variance inflation factor (VIF).

Example - Hald's data

Montgomery and Peck (1982) illustrated variable selection techniques on the Hald cement data and gave several references to other analysis. The response variable y is the heat evolved in a cement mix. The four explanatory variables are ingredients of the mix, i.e., x1: tricalcium aluminate, x2: tricalcium silicate, x3: tetracalcium alumino ferrite, x4: dicalcium silicate. An important feature of these data is that the variables x1 and x3 are highly correlated (corr(x1,x3)=-0.824), as well as the variables x2 and x4 (with corr(x2,x4)=-0.975). Thus we should expect any subset of (x1,x2,x3,x4) that includes one variable from highly correlated pair to do as any subset that also includes the other member.

Example - Hald's data

	[,1]	[,2]	[,3]	[,4]
[1,]	1.0000000	0.2285795	-0.82413376	-0.24544511
[2,]	0.2285795	1.0000000	-0.13924238	-0.97295500
[3,]	-0.8241338	-0.1392424	1.00000000	0.02953700
[4,]	-0.2454451	-0.9729550	0.02953700	1.00000000



Example - Hald's data

```
Call:
lm(formula = y.hald ~ x.hald)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1750 -1.6709  0.2508  1.3783  3.9254

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.4054    70.0710   0.891   0.3991
x.hald1       1.5511     0.7448   2.083   0.0708 .
x.hald2       0.5102     0.7238   0.705   0.5009
x.hald3       0.1019     0.7547   0.135   0.8959
x.hald4      -0.1441     0.7091  -0.203   0.8441
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.446 on 8 degrees of freedom
Multiple R-squared: 0.9824, Adjusted R-squared: 0.9736
F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07
```

Summary of Fit

RSquare	0.982376
RSquare Adj	0.973563
Root Mean Square Error	2.446008
Mean of Response	95.42308
Observations (or Sum Wgts)	13

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	2667.8994	666.975	111.4792
Error	8	47.8636	5.983	Prob > F
C. Total	12	2715.7631		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	62.405369	70.07096	0.89	0.3991
X1	1.5511026	0.74477	2.08	0.0708
X2	0.5101676	0.723788	0.70	0.5009
X3	0.1019094	0.754709	0.14	0.8959
X4	-0.144061	0.709052	-0.20	0.8441

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
X1	1	1	25.950911	4.3375	0.0708
X2	1	1	2.972478	0.4968	0.5009
X3	1	1	0.109090	0.0182	0.8959
X4	1	1	0.246975	0.0413	0.8441

Example - Hald's data

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	62.405369	70.07096	0.89	0.3991	.
X1	1.5511026	0.74477	2.08	0.0708	38.496211
X2	0.5101676	0.723788	0.70	0.5009	254.42317
X3	0.1019094	0.754709	0.14	0.8959	46.868386
X4	-0.144061	0.709052	-0.20	0.8441	282.51286

Example - Hald's data

Summary of Fit

RSquare	0.978678
RSquare Adj	0.974414
Root Mean Square Error	2.406335
Mean of Response	95.42308
Observations (or Sum Wgts)	13

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	2657.8586	1328.93	229.5037
Error	10	57.9045	5.79	Prob > F
C. Total	12	2715.7631		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	52.577349	2.286174	23.00	<.0001*	.
X1	1.4683057	0.121301	12.10	<.0001*	1.055129
X2	0.6622505	0.045855	14.44	<.0001*	1.055129

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
X1	1	1	848.4319	146.5227	<.0001*
X2	1	1	1207.7823	208.5818	<.0001*