Applied Regression

Multiple Linear Regression Model (MLR Model)

Adequacy

Module 4 Lecture - 4-2

So far we have learned how to estimate the regression coefficients and get the estimated regression equation. Once again our MLR model is

$$Y = \beta_0 + \beta_1.X_1 + \beta_2.X_2 + \ldots + \beta_k.X_k + e$$

We have n data points which creates n equations as

$$Y_i = \beta_0 + \beta_1.X_{i1} + \beta_2.X_{i2} + \ldots + \beta_k.X_{ik} + e_i \quad \text{for } i = 1, 2...., n.$$

As a result, MLR model (in matrix form) is written as

$$Y_{n\times 1} = X_{n\times(k+1)} \times \beta_{(k+1)\times 1} \; + \; e_{n\times 1}$$

Using the least square theory, we get the estimated value of the β-vector as

$$\hat{\beta} = (X'X)^{-1}X'Y$$

To provide any inference regarding the random behavior of estimated coefficients or the models we need an assumption on the random part of the

model which are the errors. Like in simple linear regression we have the same assumption $e_i \sim N(0, \sigma^2)$ and they are all independent. As a result it implies,

$$Y_i \sim N(\beta_0 + \beta_1.X_{i1} + \beta_2.X_{i2} + \ldots + \beta_k.X_{ik}, \sigma^2)$$

and they are all independent.

Next we have estimated σ which plays a crucial role in the analysis.

We developed an estimator of σ^2 from the residual sum of squares

$$\text{Sum of Squares of Error} = SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e}$$

Substituting $\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\beta}$ and after few simplification we get
$SSE = \mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - \hat{\beta}\mathbf{X}'\mathbf{y}$

Like in SLR that the error (residual) sum of squares has (n - k - 1) degrees of freedom associated with it since (k+1) parameters are estimated in the regression model. Hence the mean square error is

$$MSE = \frac{SSE}{n-k-1} \quad \text{and} \quad \hat{\sigma^2} = MSE \quad \Rightarrow \quad \hat{\sigma} = \sqrt{MSE}$$

Hypothesis Testing in MLR model

Once we have estimated the parameters in the model, we face two immediate questions:

1. How good is this model? - It is the macro view of the model, all the variables as a whole.

2. How important are the variables individually? - It is the micro view of the model and it tries to justify the importance and presence of the variables individually.

To built a good model it is important to first look at overall model and then look at the individual variables. It is possible that overall model looks good (i.e significant) but each variable looks insignificant. We will first look at the macro view and then get into the micro view.

Overall Significance Test of The MLR Model

The test for significance of the MLR model is a test to determine if there is a linear relationship between the response y and any of the regressor variables $x_1, x_2, .., x_k$. This procedure is often thought of as an overall or global test of model adequacy. This test tries to determine if there is any X_i which contributes significantly to the variation of Y. The test procedure is a generalization of the analysis of variance used in simple linear regression. The hypothesis are written as:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$
$$H_1 : \beta_j \neq 0 \quad \text{at least for one j}$$

The test statistic is based on the partitions of SST (SSR & SSE).

$$SST = SSR + SSE \quad \text{where} \quad SSR = \hat{\beta} X' y - n.\bar{y}^2$$
$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n.\bar{y}^2 = y'y - n.\bar{y}^2$$
$$SSE = \sum_{i=1}^{n} (y_i - \hat{y})^2 = y'y - \hat{\beta} X' y$$

The statistic used for the above test is $F - stat = \frac{SSR/k}{SSE/(n-k-1)}$

Note that SSR and SSE are both random variables and it can be proven that they are independently distributed. Using that fact, it can be derived that $F - stat \sim F(k, n-k-1)$ under H_0 and the test rejects H_0 for higher values of the F-stat.

So here is the summary to test the overall significance:

Hypothesis : $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \;\; vs$

H_1 : at least one of the $\beta_j' s$ is non zero

Test-Statistic: $F - stat = \frac{SSR/k}{SSE/n-k-1} = \frac{R^2/k}{(1-R^2)/n-k-1}$

Rejection Region: Reject $H_0 \quad \text{if} \quad F - stat > F_{\alpha,k,n-k-1}$

p-value: $P(F > F - stat) \quad$ where $F \sim F(k, n-k-1)$

Remark: As H_0 is a model with no independent variables at all, it is usually a weak test in practice. Even if one variable is important, it comes out significant. So it is more like must happen else variables are all together not significant.

ANOVA TABLE FOR MLR MODEL

Source	SS	df	MS	F-Stat	p-Value
Regression	SSR	(k+1-1)	$MSR = \frac{SSR}{(k)}$	$F - stat = \frac{MSR}{MSE}$	$P(F > F - stat)$
Error	SSE	(n-k-1)	$MSE = \frac{SSE}{(n-k-1)}$		
Total	SST	(n-1)			

Test for Individual Independent Variables - The Micro View

Once we have determined that at least one of the regressors is important, a logical question becomes which one(s). The basic question here is whether any particular X_i can justify its presence in the model. It means that if we fit a model with all but one particular X_i first and then we fit a second model including the X_i along with others, do we improve the model significantly (in statistical terms). It is a well known fact that adding a variable to a regression model always increase the sum of square of regression (i.e SSR) and decrease the sum of squares of errors (i.e SSE). We must decide whether the increase in the regression sum of squares is sufficient to warrant using the additional regressor in the model. The addition of a regressor may also increase the variance of the fitted value so we must be careful to include only regressors that are of real value in explaining the response. Furthermore, adding an unimportant regressor may increase the residual mean square, which may decrease the usefulness of the model.

The hypotheses for significance testing of any individual variable (say X_j) is done in terms of β_j and the hypothesis are framed as

$H_0 : \beta_j = 0 \quad vs \quad H_1 : \beta_j \neq 0$

Note the fact that H_0 represent the model with all but the variable X_j and H_1 represents the model with all X's including X_j.

Now Variance of $\hat{\beta}_j$ is $Var(\hat{\beta}_j) = S^2_{\hat{\beta}_j}$ (need to pick it up from output)

To test β_j against 0, the test statistic used is

$t - stat = \frac{\hat{\beta}_j}{S_{\hat{\beta}_j}}$ which follows t-distribution with (n-k-1) degrees of freedom under H_0.

Hence this is really a test of contribution by X_j given that the other variables are already in the model. Though the following test is a two sided test but one sided tests can be used where necessary.

Hence the summary of the test is as follows:

Hypothesis : $H_0 : \beta_j = 0 \quad vs \quad H_1 : \beta_j \neq 0$

Test-Statistic: $t - stat = \frac{\hat{\beta}_j}{S_{\hat{\beta}_j}}$

Rejection Region: Reject $H_0 \quad$ if $\quad |t - stat| > t_{\alpha/2,n-k-1}$

p-value: $\quad = 2.P(t > |t - stat|) \quad$ where $t \sim t_{n-k-1}$

Any software will produce the testing output for individual β_j's as follows:

Parameter Estimates

Parameter	Estimate	St. Deviation	t-stat	p-Value	Conf. Interval
β_0	$\hat{\beta}_0$	$S_{\hat{\beta}_0}$	$t_0 = \frac{\hat{\beta}_0}{S_{\hat{\beta}_0}}$	$2P(t > \lvert t_0 \rvert)$	$\hat{\beta}_0 \pm t_{\alpha/2,n-k-1}.S_{\hat{\beta}_0}$
β_1	$\hat{\beta}_1$	$S_{\hat{\beta}_1}$	$t_1 = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}$	$2P(t > \lvert t_1 \rvert)$	$\hat{\beta}_1 \pm t_{\alpha/2,n-k-1}.S_{\hat{\beta}_1}$
β_2	$\hat{\beta}_2$	$S_{\hat{\beta}_2}$	$t_2 = \frac{\hat{\beta}_2}{S_{\hat{\beta}_2}}$	$2P(t > \lvert t_2 \rvert)$	$\hat{\beta}_2 \pm t_{\alpha/2,n-k-1}.S_{\hat{\beta}_2}$
...	...	....	.....	.....	.....
β_k	$\hat{\beta}_k$	$S_{\hat{\beta}_k}$	$t_k = \frac{\hat{\beta}_k}{S_{\hat{\beta}_k}}$	$2P(t > \lvert t_k \rvert)$	$\hat{\beta}_k \pm t_{\alpha/2,n-k-1}.S_{\hat{\beta}_k}$

General Testing Table for β_j

Hypothesis	$H_o : \beta_j \geq \beta_{jo}$ $H_1 : \beta_j < \beta_{jo}$	$H_o : \beta_j \leq \beta_{jo}$ $H_1 : \beta_j > \beta_{jo}$	$H_o : \beta_j = \beta_{jo}$ $H_1 : \beta_j \neq \beta_{jo}$
Test Statistic	$t - stat = \frac{\hat{\beta}_j - \beta_{jo}}{S_{\hat{\beta}_j}}$	$t - stat = \frac{\hat{\beta}_j - \beta_{jo}}{S_{\hat{\beta}_j}}$	$t - stat = \frac{\hat{\beta}_j - \beta_{jo}}{S_{\hat{\beta}_j}}$
Rejection	$t - stat < -t_{\alpha,n-k-1}$	$t - stat > t_{\alpha,n-k-1}$	$\lvert t - stat \rvert > t_{\alpha/2,n-k-1}$
p-value	$Prob(t < t - stat)$	$Prob(t > t - stat)$	$2.Prob(t > \lvert t - stat \rvert)$

Using the same statistic the confidence interval for β_j is calculated.

$100(1 - \alpha)$ % confidence interval interval is

$$\hat{\beta}_j \pm t_{\alpha/2,n-k-1}.S_{\hat{\beta}_j}$$

Example 4.1 (Revisit) - Consider the National Football League data in Table B.1. (You can download it from the site) Below is the output from R.

ANOVA TABLE (Macro View)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x2	1	76.19	76.19	26.17	0.0000
x8	1	166.83	166.83	57.31	0.0000
x7	1	14.07	14.07	4.83	0.0378
Residuals	24	69.87	2.91		

Parameter Estimates Table (Micro View)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.8084	7.9009	-0.23	0.8209
x2	0.0036	0.0007	5.18	0.0000
x8	-0.0048	0.0013	-3.77	0.0009
x7	0.1940	0.0882	2.20	0.0378

b. Construct the analysis-of-variance table and test for significance of regression.

Note: The output from two different softwares are little different regarding the presentation of ANOVA table. SAS output (next slide) shows the overall SSR, SSE and SST where as R-output shows the intermediate values of those when you add one variable at a time. But both of them have the same error line. So if the 3 lines in R added together then you get the regression line of SAS output which is the goal here. So we can add the Sum-of-Square and degrees of freedom columns from R to get to the Model line of SAS output and then we can calculate the other entries of the SAS output. But you cannot add the other columns (e.g Mean-Square, F-stat or p-value)

$H_0 : \beta_2 = \beta_7 = \beta_8 = 0 \quad vs \quad H_1$: at least one of three is non-zero

So SSR $= 76.19 + 166.83 + 14.07 = 257.09$ (same from the SAS output).

$F - stat = \frac{SSR/k}{SSE/(n-k-1)} = \frac{257.09/3}{69.87/24} = 29.44$

As F-stat $= 29.44 > \quad F_{\alpha,k,nk-1} = F_{0.05,3,24} = 3.01$ we reject the null hypothesis. So data provides sufficient evidence to support the fact the the model is significant at 5% level.

c. Calculate t statistics for testing the hypotheses

$H_0 : \beta_2 = 0, H_0 : \beta_7 = 0, \ and \ H_0 : \beta_8 = 0.$

What conclusions can you draw about the roles the variables $x_2, x_7, \text{ and }, x_8$ play in the model?

Solution:

For testing β_2 we set $H_0 : \beta_2 = 0 \ vs \ H_0 : \beta_2 \neq 0$

Now, $t - stat = \frac{\hat{\beta}_2}{S_{\hat{\beta}_2}} = \frac{0.00360}{0.000695} = 5.18 \Rightarrow |t - stat| > t_{\alpha/2,n-k-1} = 2.064$

For testing β_7 we set $H_0 : \beta_7 = 0 \ vs \ H_0 : \beta_7 \neq 0$

Now, $t - stat = \frac{\hat{\beta}_7}{S_{\hat{\beta}_7}} = \frac{0.19396}{0.08823} = 2.20 \Rightarrow |t - stat| > t_{\alpha/2,n-k-1} = 2.064$

For testing β_8 we set $H_0 : \beta_8 = 0 \ vs \ H_0 : \beta_8 \neq 0$

Now, $t - stat = \frac{\hat{\beta}_8}{S_{\hat{\beta}_8}} = \frac{-0.00482}{0.00128} = -3.77 \Rightarrow |t - stat| > t_{\alpha/2,n-k-1} = 2.064$

Hence we reject all three null hypothesis. So the data provides sufficient evidence to support the fact that each independent variable provides enough contribution on top of other variables to stay in the model at 5% level.

Output from SAS - Problem 4.1

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: y y

Number of Observations Read	28
Number of Observations Used	28

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	257.09428	85.69809	29.44	<.0001
Error	24	69.87000	2.91125		
Corrected Total	27	326.96429			

Root MSE	1.70624	**R-Square**	0.7863
Dependent Mean	6.96429	**Adj R-Sq**	0.7596
Coeff Var	24.49984		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	-1.80837	7.90086	-0.23	0.8209
x2	x2	1	0.00360	0.00069500	5.18	<.0001
x7	x7	1	0.19396	0.08823	2.20	0.0378
x8	x8	1	-0.00482	0.00128	-3.77	0.0009

Coefficient of Determination - R^2

Like in Simple linear regression, model adequacy is measured by R^2 as

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

The above R^2 represent the square of the correlation coefficient between Y and $\hat{Y}$.

Mathematically it can be shown that R^2 increases with every addition of independent variables even if that variable has nothing to do with the dependent variable. Due to this unwanted property, it is better to pay attention to adjusted R^2 as it can decrease with the addition of a bad variable. $adj - R^2$ only increases only if MSE goes down with the addition of variable(s), because it adjust R^2 using degrees of freedom. $adj - R^2$ is always less than R^2 and it can be even negative.

Adjusted Coefficient of Determination - $adj - R^2$

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

Problem 4.1 (Re-visit) - Consider the National Football League data in Table B.1. (You can download it from the site)

d. Calculate R^2 and for this model.

Solution: Note that from the SAS output we can get SSR=257.09428, and SSE = 69.87000.

So $R^2 = \frac{SSR}{SST} = \frac{257.09}{257.09+69.87} = 0.7863$

It means that 78.63% of the variations in Y can be explained by this model.

Now we can also calculate $adj - R^2$ as

$adj - R^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - \frac{69.87/24}{(257.09+69.87)/27} = 0.7596$

Testing Partial Models

Consider the regression model with k regressors

$$Y_{n\times 1} = X_{n\times(k+1)} \times \beta_{(k+1)\times 1} \; + \; e_{n\times 1}$$

We would like to determine if some subset of $r < k$ regressors contributes enough as a group to stay in the regression model or they can be dropped as a whole group. Let the vector of regression coefficients be partitioned as follows:

where $\boldsymbol{\gamma_1}$ is the first (k - r) x 1 β's and $\boldsymbol{\gamma_2}$ is the rest r x 1 β's. Writing in terms of row vector, $\beta' = (\boldsymbol{\gamma_1'}, \boldsymbol{\gamma_2'})$ and We wish to test the hypotheses that addition of r many independent variables as a group improve the model significantly.

So our hypothesis is $H_0 : \gamma_2 = \underline{0} \quad vs \quad H_1 : \gamma_2 \neq \underline{0}$

In terms of model we are testing (Partition of $\mathbf{X} = \mathbf{X_1^\star} \,\big|\, \mathbf{X_2^\star}$)

$$H_0 : Y = \mathbf{X_1^\star}\gamma_1 = \beta_0 + \beta_1.X_1 + \ldots + \beta_{k-r}.X_{k-r} + e$$
$$H_1 : Y = \mathbf{X_1^\star}\boldsymbol{\gamma_1} + \mathbf{X_2^\star}\gamma_2 = X.\beta$$
$$= \beta_0 + \beta_1.X_1 + \ldots + \beta_{k-r}.X_{k-r} \quad + \beta_{k-r+1}.X_{k-r+1}e + \ldots + \beta_k.X_k + e$$

Let's say the MLR model in H_0 is model-1 and model in H_1 is model-2. Then it is known that $SSR_2 > SSR_1$ and $SSE_1 > SSE_2$ because model-1 is a subset of model-2. As SST are same for both the model, difference between the SSR and SSE are same as well. And that difference account for the improvement of addition of γ_2 parameters or in other words addition of r many variables. It is also known that the difference (i.e $SSR_2 - SSR_1 = SSE_1 - SSE_2$) is independent of SSE_2. That's why statistic for testing the hypothesis is built as

$$F - stat = \frac{(SSR_2 - SSR_1)/r}{SSE_2/(n-p)} = \frac{(SSE_1 - SSE_2)/r}{SSE_2/(n-p)}$$

where $F - stat \sim F(r, n-p)$ under H_0

Note that $\frac{SSE_2}{(n-p)}$ is the MSE for the original MLR model.

Summary of the Partial Model Testing:

Hypothesis : $H_0 : \gamma_2 = 0 \;\; vs \;\; H_1 : \gamma_2 \neq 0$

Test-Statistic: $F - stat = \frac{(SSR_2 - SSR_1)/r}{SSE_2/(n-p)} = \frac{(SSE_1 - SSE_2)/r}{SSE_2/(n-p)}$

Rejection Region: Reject H_0 if $F - stat > F_{\alpha,r,n-p}$

p-value: $P(F > F - stat) \quad$ where $F \sim F(r, n-p)$

Remark: Note that if we set r=1 then it is same as testing individual regressors as before. Though we used t-test to perform the hypothesis but it is equivalent to the F-test (known as partial F-test) with r=1. Technically $F - stat = (t - stat)^2$ and $F_{\alpha,1,n-p} = (t_{\alpha/2,n-p})^2$.

e. Using the partial F test, determine the contribution of x_7 to the model.

How is this partial F statistic related to the t test for β_7 calculated in part c above?

To use the partial F-test we will rerun the model without X_7 and find the SSE with only X_2 & X_8 in the model.

Then the new SSE_2 is 83.94 and new $SSR_2 = 243.03$ (rerun yourself and check).

Hypothesis for the partial F-test is

$H_0 : \beta_7 = 0 \; vs \; H_1 : \beta_7 \neq 0$ is

Then $F - stat = \frac{(SSR - SSR_2)/1}{SSE/(n-k-1)} = \frac{(257.09 - 243.03}{69.87/24} = 4.84$

As $F - stat = 4.84 \; > \; F_{0.05,1,24} = 4.26$ we reject the null hypothesis and say that X_7 is significant.

Note that partial $F - stat = 4.84 = (2.20)^2 = (t - stat)^2$

The t-stat is in the parameter estimation table of the original model.

Testing the General Linear Hypothesis

Many different hypotheses about regression coefficients can be tested using a unified approach. All the previous hypothesis are a special case of this procedure. In this general set up test statistic for testing the hypothesis is usually calculated as the difference between between the two residual sums of squares under two different models. Here is the basic outline of the procedure,

Consider our usual MLR model with (k+1) many β's with n observations. Suppose that the null hypothesis of interest can be expressed as $H_0 : T\beta = c$, where T is an m x (k+1) matrix of constants, such that only r of the m equations in $T\beta = 0$ are independent.

The full model is $y = X\,\beta + e$, with the sum of squares of error is SSE.

To obtain the reduced model, the r independent equations in $T\beta = c$ are used to replace r of the regression coefficients in the full model in terms of the remaining (k + 1 - r) regression coefficients. Hence it is possible to write the new reduced model in terms of new parameters and transformed X matrix. Let's call the sum of squares of this reduced model as SSE_{RM}. It can be easily shown that $SSE_{RM} > SSE$ and the difference $SS_H = SSE_{RM} - SSE$ provides the amount of increase due to constraints.

Hence the the testing procedure is as follows:

Hypothesis: $H_0 : T\beta = c \; vs \; H_1 : T\beta \neq c$

Test Statistic: $$F - stat = \frac{(SSE_{RM} - SSE)/r}{SSE/(n-k-1)}$$

$$= \frac{(T\hat{\beta}' - c)'[T(X'X)^{-1}T']^{(-1)}(T\hat{\beta}' - c)/r}{SSE/(n-k-1)}$$

Reject H_0 if $F - stat > F_{\alpha,r,n-k-1}$.

Remark: Observe that any kind of linear hypothesis can be tested using this theory.

For example, to test $\beta_2 = 1$ in Problem 4.1 we can set T = (0 1 0 0) and c = 1.

Similarly to test $\beta_7 = \beta_8$ we can set T = (0 0 1 -1) and c = 0.

Confidence and prediction interval for any given $X = X_0$.

Given any given value of x $= x_0 = (x_{01}, x_{02}, \ldots, x_{0k})$ we can find the $100(1-\alpha)\%$ confidence interval for the mean of y given x (i.e $E(y|x_0)$) as

$$\left(x_0'\hat{\beta} - t_{\alpha/2,n-k-1}\,\hat{\sigma}\,\sqrt{x_0'(X'X)^{(-1)}x_0}\;,\;\; x_0'\hat{\beta} + t_{\alpha/2,n-k-1}\,\hat{\sigma}\,\sqrt{x_0'(X'X)^{(-1)}x_0}\right)$$

Similarly $100(1-\alpha)\%$ prediction interval for one future value (i.e $\hat{y_0}$) is

$$\left(x_0'\hat{\beta} - t_{\alpha/2,n-k-1}\,\hat{\sigma}\,\sqrt{1 + x_0'(X'X)^{(-1)}x_0}\;,\;\; x_0'\hat{\beta} + t_{\alpha/2,n-k-1}\,\hat{\sigma}\,\sqrt{1 + x_0'(X'X)^{(-1)}x_0}\right)$$

Problem: 3.13 An engineer studied the effect of four variables on a dimensionless factor used to describe pressure drops in a screen-plate bubble column. Table B.9 summarizes the experimental results.

a. Fit a multiple linear regression model relating this dimensionless number to these regressors.

b. Test for significance of regression. What conclusions can you draw?

c. Use t tests to assess the contribution of each regressor to the model. Discuss your findings.

d. Calculate R^2 and for this model. Compare these values to the R^2 and for the multiple linear regression model relating the dimensionless number to x_2 and x_3. Discuss your results.

e. Find a 99% CI for the regression coefficient for x_2 for both models in part d. Discuss any differences.

Output from SAS - Problem 3.13

Problem 3.13

The REG Procedure
Model: MODEL2
Dependent Variable: y y

Number of Observations Read	62
Number of Observations Used	62

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	4	3209.71775	802.42944	31.92	<.0001
Error	57	1432.78822	25.13664		
Corrected Total	61	4642.50597			

Root MSE	5.01364	**R-Square**	0.6914
Dependent Mean	23.50806	**Adj R-Sq**	0.6697
Coeff Var	21.32734		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	5.89453	4.32508	1.36	0.1783
x1	x1	1	-0.47790	0.34002	-1.41	0.1653
x2	x2	1	0.18271	0.01718	10.63	<.0001
x3	x3	1	35.40284	11.09960	3.19	0.0023
x4	x4	1	5.84391	2.90978	2.01	0.0494

Test nametest Results for Dependent Variable y				
Source	**DF**	**Mean Square**	**F Value**	**Pr > F**
Numerator	1	49.65663	1.98	0.1653
Denominator	57	25.13664		

a. Fit a multiple linear regression model relating this dimensionless number to these regressors.

Using the output, we can see the estimated regression hyperplane is

$$\hat{y} = 5.89453 - 0.47790 \times x_1 \; + \; 0.18271 \times x_2 \; + \; 35.40284 \times x_3 \; + \; 5.84391 \times x_4$$

b. Test for significance of regression. What conclusions can you draw?

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \quad vs \quad H_1 :$ at least one of them is not zero

F-stat $= 31.92$ and p-value < 0.0001. Hence H_0 is rejected at any reasonable $\alpha > 0.0001$.

Conclusion: Overall this regression model is significant at $\alpha < 0.001$.

c. Use t tests to assess the contribution of each regressor to the model. Discuss your findings.

For each of the β_j, we set the hypothesis $H_0 : \beta_j = 0 \quad vs \quad H_1 : \beta_j \neq 0$.

By looking at the p-values we can see that β_1 is not significant at 5% level. Also β_4 is barely significant at 5% level but not for any $\alpha < 0.0494$.

But β_2 and β_3 are both significant for smaller values of α.

Output from SAS - Problem 3.13

Problem 3.13

The REG Procedure
Model: MODEL2
Dependent Variable: y y

Number of Observations Read	62
Number of Observations Used	62

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	3091.41724	1545.70862	58.80	<.0001
Error	59	1551.08873	26.28964		
Corrected Total	61	4642.50597			

Root MSE	5.12734	**R-Square**	0.6659
Dependent Mean	23.50806	**Adj R-Sq**	0.6546
Coeff Var	21.81099		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	7.18971	4.03031	1.78	0.0796
x2	x2	1	0.18456	0.01755	10.52	<.0001
x3	x3	1	35.11616	11.33308	3.10	0.0030

For part (d) and (e) we need to run the second model and outputs are shown above.

d. Calculate R^2 and for this model. Compare these values to the R^2 and for the multiple linear regression model relating the dimensionless number to x_2 and x_3. Discuss your results.

$R^2 = 69.14\%$ and $adj - R^2 = 66.97\%$ for this model but for the model only with x_2 and x_3 has $R^2 = 66.49\%$ and $adj - R^2 = 65.46\%$.

Though it appears that the reduced model has fairly close R^2 but it cannot be confirmed until we test the reduced model against the full model. (Try to do that on your own with SAS or R).

e. Find a 99% CI for the regression coefficient for x_2 for both models in part d. Discuss any differences.

99% CI for β_2 in the full model is

$$\hat{\beta}_2 \pm t_{\alpha/2,n-k-1} S_{\hat{\beta}_2} = 0.1817 \pm 2.665 \times 0.01718 = (0.1359, 0.2275)$$

where as for the reduced model,

$$\hat{\beta}_2 \pm t_{\alpha/2,n-k-1} S_{\hat{\beta}_2} = 0.18456 \pm 2.662 \times 0.01755 = (0.1378, 0.2313).$$

as we that they are close.

(Here we have used p-value and output but you need to prepare to do the test as before using t-table if p-value if not available).

Summary - What you should learn

1. Testing of hypothesis in MLR model.
2. How to test importance of each variables.
3. Concept and how to test partial models.
4. R^2 and $adj - R^2$ concepts of MLR model.