CH 12 - Multiple Linear Regression

[Note: Simple linear regression is a special case of multiple linear regression. When number of independent variables is reduced to one (k=1), multiple linear regression become simple linear regression.]

Population Regression Model:

$$\mu_Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k$$

$x_k x_1, x_2, \ldots, x_k$ = Independent variables (also called predictors or regressors)
μ_Y = mean value of Y for a given set of values of $x_1, x_2, \ldots, x_k$

For the *i*-th observation, the observed value of *Y*

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

ε_i=error in *i*-th observation

We assume that errors are normally distributed with mean 0 and standard deviation σ^2, so the error term vanishes when you take the mean value of Y.

Sample Regression Model:

The sample regression equation:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

For the *i*-th observation, the predicted/estimated value of *Y*

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + e_i$$

Where $b_0, b_1, \ldots\ldots, b_k$ are estimates of regression coefficients $\beta_0, \beta_1, \ldots\ldots, \beta_k$ respectively.

Residual:

For the *i*-th observation $e_i = y_i - \hat{y_i}$

Relationship of b_k with β_k :

Now we make the following general assumption to relate b_k with β_k. We assumed that the errors are normally distributed with mean 0 and standard deviation σ^2. Then it can be shown that

$$\text{the } mean\ of\ b_k = \beta_k.$$

$$\text{the } variance\ of\ b_k = c_{kk}\sigma^2$$

The c_{kk} values are the diagonal elements of the matrix $(\boldsymbol{X'X})^{-1}$ and will be **given to you.**

The error variance σ^2 is estimated by the residual mean square (MSE). The residual mean square is the sum of squares of the residuals divided by the associated degrees of freedom (n-k-1).

$$\text{Estimate of } \sigma^2 = s^2 = MSE = \frac{SSE}{n-k-1},\ SSE = \sum_{i=1}^{n} e_i{}^2 = \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

Problem # 1 (Based on Exercise 12.17)

Consider the following set of data:

y	x_1	x_2
0.231	740	1.10
0.107	740	0.62
0.053	740	0.31
0.129	805	1.10
0.069	805	0.62
0.030	805	0.31
1.005	980	1.10
0.559	980	0.62
0.321	980	0.31
2.948	1235	1.10
1.633	1235	0.62
0.934	1235	0.31

The following fitted/sample regression equation (surface) is obtained using least-squares method.

$$\hat{y} = -3.3727 + 0.00362\, x_1 + 0.9476\, x_2$$

Estimate $error\ variance, \sigma^2$.

Also, find the standard error of the estimate.

Solution:

Estimate of $\sigma^2 = s^2 = MSE = \frac{SSE}{n-k-1} = \frac{\sum_{i=1}^{n} e_i^2}{12-2-2} = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{9} = \frac{1.489}{9} = 0.165$

The standard error of estimate is square root of 0.165 =0.406.

[Calculation detail:

Note that $e_1 = y_i - \hat{y}_i = 0.231 — 3.3727 + 0.00362 * 740 + 0.9476 * 1.1 = -0.11746, etc.$

If you are curious, detailed calculations are given below

i	y	x1	x2	predicted y (y_hat)	e=y-y_hat	e^2
1	0.231	740	1.1	0.34846	-0.11746	0.013797
2	0.107	740	0.62	-0.106388	0.213388	0.045534
3	0.053	740	0.31	-0.400144	0.453144	0.205339
4	0.129	805	1.1	0.58376	-0.45476	0.206807
5	0.069	805	0.62	0.128912	-0.059912	0.003589
6	0.03	805	0.31	-0.164844	0.194844	0.037964
7	1.005	980	1.1	1.21726	-0.21226	0.045054
8	0.559	980	0.62	0.762412	-0.203412	0.041376
9	0.321	980	0.31	0.468656	-0.147656	0.021802
10	2.948	1235	1.1	2.14036	0.80764	0.652282
11	1.633	1235	0.62	1.685512	-0.052512	0.002758
12	0.934	1235	0.31	1.391756	-0.457756	0.209541
	0.66825				SSE=	**1.485845**
					n	12
					k	2
					n-k-1	9
					MSE=s^2	0.165094

Problem # 2 (Based on Exercise 12.20)

Consider the data set given in Problem # 1. The following fitted regression equation is obtained using least-squares method.

$$\hat{y} = -3.3727 + 0.00362x_1 + 0.9476x_2$$

Estimate $the\ variance\ of\ b_1\ and\ b_2$

Given: c_{ii} , the diagonal elements of $(X'X)^{-1}$ are computed to be

c_{00} = 2.450262 c_{11} = 2.26898E-6, and c_{22}=0.788975

Solution:

(i) Estimate of $\sigma^2 = s^2 = MSE = \frac{SSE}{n-k-1} = 0.16509$ (From problem # 1)

The *variance of* $b_k = c_{kk}\sigma^2$

Therefore, the estimate of the variance of $b_1 = c_{11}s^2$=2.26898E-6*0.16509=3.74713E-7

Similarly, the estimate of the variance of $b_2 = c_{22}s^2$=0.788975*0.16509=0.13024

[**Note:** The values of c11 and c22 values will be given to you during exam. If you are curious to see the detail calculation:

$\boldsymbol{X} =$

1	740	1.1
1	740	0.62
1	740	0.31
1	805	1.1
1	805	0.62
1	805	0.31
1	980	1.1
1	980	0.62
1	980	0.31
1	1235	1.1
1	1235	0.62
1	1235	0.31

$(\boldsymbol{X'X})^{-1}$

$=$

2.450262	-0.0021337	-0.53387
-0.00213	2.26989E-06	-3E-18
-0.53387	-3.5927E-18	0.788975

]

Inferences in Multiple Linear Regression (Inference on Individual Coefficients):

[Note: We will test hypotheses (i) on overall regression model using f-statistic and (ii) on individual coefficients using t-statistic.]

In some regression situations, individual coefficients are of importance to the experimenter and s/he need to perform hypothesis test on individual coefficients. We can use a t-statistic to do that.

We assume that the $b_j (j = 0,1,2, \ldots, k)$ are normally distributed with mean β_j and standard deviation $c_{jj}\sigma^2$. The constants c_{jj} are the diagonal elements of the $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ matrix and will be given. We also estimated σ^2 with s^2, the variance of the residuals. (Recall problems 1 and 2). Therefore, we can use the following t-statistic with $df = \nu = n - k - 1$ degrees of freedom to test hypotheses as usual.

$$t = \frac{b_j - mean\ of\ b_j}{standard\ deviation\ of\ b_j}$$

$$\Longrightarrow t = \frac{b_j - \beta_j}{S_{bj}} = \frac{b_j - \beta_j}{s\sqrt{c_{jj}}}$$

Problem # 3 (Based on Exercise 12.23)

For the data and fitted model given in problem # 1 and 2, test the hypothesis that $\beta_1 = 0$ at the 0.5 level of significance against the alternative that $\beta_1 \neq 0$. Use c_{ii} values from problem 2. Based on the result whether it is justified to keep the independent variable x1 in the fitted model.

Solution:

[Note: We apply techniques learnt in Ch 10. The level of significance, alpha, is given, and therefore, recall that we can reach conclusion in two different ways - (i) either compute P-value from the observed t-statisitc and compare with alpha and reach conclusion (reject Ho when P-value is less than or equal to alpha) or (ii) Find the critical region, then make decision based on the observed t-statistic – whether it belongs to the reject region or in the do not reject region. Go ahead and do it.]

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

P-value method

Observed value of t-statistic: $t = \frac{b_1 - \beta_{10}}{s\sqrt{c_{11}}} = \frac{0.0036 - 0}{0.4067 * \sqrt{2.26898\text{E}-6}} = \frac{0.0036}{0.000612} = 5.91$

We used the s value computed in problem # 1.

$$P - value = 2 * (t > 5.91) \; with \; df = n - k - 1 = 12 - 2 - 1 = 9$$
$$P - value < 0.0005 \; (page \; 738, table \; A.4)$$

Therefore, P-value is much less than given level of significance and thus we reject the null hypothesis. The conclusion drawn is that the variable x1 is significant, that is, it explains a significant amount of variation in y. It is justified to keep this variable in the model.

Critical region method

$$\alpha = 0.05, df = n - k - 1 = 12 - 2 - 1 = 9$$

Critical region: $t > t_{0.025,9} \; and \; t < t_{0.025,9}$

$$\Longrightarrow t > 2.262 \; and \; t < -2.262$$

Observed value of t-statistic: $t = \frac{b_1 - \beta_{10}}{s\sqrt{c_{11}}} = \frac{0.0036 - 0}{0.4067 * \sqrt{2.26898\text{E}-6}} = \frac{0.0036}{0.000612} = 5.91$

We find that the observed value of t belongs to the reject region and therefore we reject the null hypothesis, and conclude that the variable x1 explain significant amount of variation in y. It is justified to keep this variable in the model.

Regression sum of squares:

The regression sum of squares (SSR) is a useful quantity in determining the adequacy of a fitted regression model. (Later, we will construct two important quantities based on SSR –(i) coefficient of determination and (ii) a f-statistic).

The regression sum of squares measures the improvement by predicting y using the fitted model over the mean value of y. Imagine the vertical distances of predicted y values ($\hat{y}$) from a horizontal line $y = \bar{y}$. SSR is a measure of the variability explained by the model.

SSR = regression sum of squares = $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$, with df = (k+1)-1=k

We can relate SSR with the previously studied sum of residual sum of squares (SSE)

$$SSE = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad with\ df = n - (k+1)$$

by the sum-of-squares identity. Adding SSR and SSE we obtain,

$$SSR + SSE = SST$$

$$where, SST = total\ sum\ of\ squares = \sum_{i=1}^{n}(y_i - \bar{y})^2\ with\ df = n - 1$$

Coefficient of Determination:

One criterion that is commonly used to illustrate the adequacy of a fitted regression model is the coefficient of determination, R^2.

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{SSE}{SST}$$

R^2 is a measure of the proportion of total variation in the response Y explained by the fitted model.

$R^2 * 100\%$ is percent variation in the response Y explained by the fitted model.

The Adjusted Coefficient of Determination:

$$R^2_{adj} = 1 - \frac{\frac{SSE}{n-k-1}}{\frac{SST}{n-1}} = 1 - \frac{s^2}{\frac{SST}{n-1}}$$

R^2_{adj} provides an adjustment for degrees of freedom. Addition of new independent variable to the model increases SSR increases, but does not affect SST (because it does not involve predicted y-values). Consequently, R^2 increases (SSE decreases) by the addition of any new variable, adding credibility to the model. However, sometimes addition of new variable adds no significant information to the model. R^2_{adj} takes into consideration both the additional information brought by new independent variable and the changed degrees of freedom. Note that addition of new variable increases k, thus decreases n-k-1.

Problem # 4

Compute and interpret R^2 and R^2_{adj} for the data and fitted model given in problem # 1.

Solution:

SSR = $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ = 6.91,
MSR=SSR/k=3.456
$SSE = \sum_{i=1}^{n} e_i{}^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ =1.486,
MSE=SSE/(n-k-1) =0.165
$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ =8.386

$$R^2 = \frac{SSR}{SST} = 0.824 = 82.4\%$$

$$R^2_{adj} = 1 - \frac{MSE}{\frac{SST}{n-1}} = 1 - \frac{0.165}{\left(\frac{8.386}{11}\right)} = 1 - \frac{0.165}{0.762} = 0.783 = 78.3\%$$

This results indicates that a relatively high proportion of the variation of the dependent variable y is accounted for by the independent variables in this model. A comparison of adjusted R^2 (0.824) with the adjusted R^2 (0.783) shows that the adjusted R^2 reduces the overall proportion of variation of y accounted. The gap between R^2 and adjusted R^2 tend to increase as non-significant independent variables are included to the model. As n increases, the difference between the two become less.

[Detailed calculation:

i	y	x1	x2	predicted y (y_hat)	e=y-y_hat	e^2		y_hat - y_avg	Square of y_hat - y_avg	y -y_avg	(y - y_avg)^2
1	0.231	740	1.1	0.34846	-0.11746	0.013797		-0.31979	0.102266	-0.43725	0.191188
2	0.107	740	0.62	-0.106388	0.213388	0.045534		-0.77464	0.600064	-0.56125	0.315002
3	0.053	740	0.31	-0.400144	0.453144	0.205339		-1.06839	1.141466	-0.61525	0.378533
4	0.129	805	1.1	0.58376	-0.45476	0.206807		-0.08449	0.007139	-0.53925	0.290791
5	0.069	805	0.62	0.128912	-0.059912	0.003589		-0.53934	0.290885	-0.59925	0.359101
6	0.03	805	0.31	-0.164844	0.194844	0.037964		-0.83309	0.694046	-0.63825	0.407363
7	1.005	980	1.1	1.21726	-0.21226	0.045054		0.54901	0.301412	0.33675	0.113401
8	0.559	980	0.62	0.762412	-0.203412	0.041376		0.094162	0.008866	-0.10925	0.011936
9	0.321	980	0.31	0.468656	-0.147656	0.021802		-0.19959	0.039838	-0.34725	0.120583
10	2.948	1235	1.1	2.14036	0.80764	0.652282		1.47211	2.167108	2.27975	5.19726
11	1.633	1235	0.62	1.685512	-0.052512	0.002758		1.017262	1.034822	0.96475	0.930743
12	0.934	1235	0.31	1.391756	-0.457756	0.209541		0.723506	0.523461	0.26575	0.070623
	0.66825				SSE=	**1.485845**		SSR	6.911372	SST	8.38652
	y_avg				n	12		MSR	3.455686		
	0.66825				k	2		R2	0.824105		
					n-k-1	9		R2_adj	0.783458		
					MSE=s^2	0.165094		f	20.93165		

]

Inferences in Multiple Linear Regression (Inference on Overall Model):

The regression sum of squares (RSS) can be used to give some indication concerning whether or not the model is an adequate explanation of the true situation. We can test the null hypothesis that the regression is not significant by forming the following ratio f (where f follows F-distribution)

$$f = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)} = \frac{SSR/k}{s^2}$$

and rejecting the hypothesis at the alpha-level of significance when

$$f > f_\alpha(k, n-k-1).$$

The above f-statistic is also used to test the following hypotheses

$$H_0: \beta_1 = \beta_2 = \cdots \ldots = \beta_k$$
$$H_1: At\ least\ one\ of\ the\ coefficients\ is\ nonzero$$

The null hypothesis is rejected at the alpha-level of significance when

$$f > f_\alpha(k, n-k-1).$$

A rejection of the null hypothesis will indicate that at least one of the variables add significant predictability for y. Failure to reject the null hypothesis will indicate that the regression model has no significant predictability for the dependent variable.

Problem # 5

Test whether the model given in problem # 1 is significant at 0.05 level of significance.

Solution:

Let us conduct the f-test. The observed value of $f = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)} = \frac{3.456}{0.165} = 20.9$

Considering significant level of 0.05, we can determine from table A-6 page 741 that the critical value on the f-distribution (with 2 and 9 degrees of freedom) is almost 4.26.

Because, f = 20.9 > $f_{0.25}(2,9) = 4.26$, we can reject H_0 and conclude that the regression is significant (not by chance).