WEEK 14

12. Multiple Regression and Model Building

In the above discussion, we have studied the simple linear regression (means that there is only one explanatory or independent variable). In the regression analysis we may have more than one explanatory variables.

Let $(Y, X_1, X_2, ..., X_p)$ be the random variables with joint probability (or probability density) function $f_{Y,X_1,X_2,...,X_p}(y, x_1, x_2, ..., x_p)$. In a similar way, we can find the conditional probability (or probability density) function of Y given $X_1 = x_1, X_2 = x_2, ..., X_p = x_p$ as

$$f_{Y|X_1=x_1,X_2=x_2,...,X_p=x_p}(y \mid x_1, x_2, ..., x_p) = \frac{f_{Y,X_1,X_2,...,X_p}(y, x_1, x_2, ..., x_p)}{f_{X_1,X_2,...,X_p}(x_1, x_2, ..., x_p)}$$

where $f_{X_1,X_2,...,X_p}(x_1, x_2, ..., x_p) > 0$. And in a similar way, we can calculate the conditional expectation of Y given $X_1 = x_1, X_2 = x_2, ..., X_p = x_p$ as

$$E(Y \mid X_1 = x_1, X_2 = x_2, ..., X_p = x_p) = h(x_1, x_2, ..., x_p).$$

This conditional expectation is known as multiple regression of Y on the explanatory variables $X_1, X_2, ..., X_p$. As it is obviously seen, this conditional expectation is a function of $x_1, x_2, ..., x_p$, namely, $h(x_1, x_2, ..., x_p)$. If $h(x_1, x_2, ..., x_p)$ is a linear function of x's then it is a multiple linear regression of Y on the variables $x_1, x_2, ..., x_p$ namely, if

$$h(x_1, x_2, ..., x_p) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_p x_p$$

then it is multiple linear regression of Y on the variables $x_1, x_2, ..., x_p$, otherwise it is a non-linear regression. For the case $p = 1$ the linear regression is named as "simple" linear regression. In this class we are going to investigate the "linear case" and for simplicity we will have two or three explanatory variables in the discussion. Consider the following multiple regression of Y on the explanatory variables x_1, x_2 and x_3

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + e_i \ , \ i = 1, 2, ..., n \tag{1}$$

The assumptions of a multiple regression model is the same as the simple linear regression,

- the error terms (e_i) are independent and identically distributed random variables (for statistical inferences we can add the normality)

- the explanatory variables (x_1, x_2 and x_3) are fixed, in the sense that the are not random.

The model can be written in a matrix notation as $\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{e}$ where

$$\underset{\sim}{y} = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_n \end{bmatrix}, \; X = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & x_{3,1} \\ 1 & x_{1,2} & x_{2,2} & x_{3,2} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 1 & x_{1,n} & x_{2,n} & x_{3,n} \end{bmatrix}, \; \underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \text{ and } \underset{\sim}{e} = \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix}.$$

Therefore the OLS estimator of the parameter vector is $\hat{\underset{\sim}{\beta}} = (X'X)^{-1}X'\underset{\sim}{y}$. Using this OLS estimator we write the fitted regression equation as $\hat{\underset{\sim}{y}} = X\hat{\underset{\sim}{\beta}}$ and thus the residual vector will be written as $\hat{\underset{\sim}{e}} = \underset{\sim}{y} - \hat{\underset{\sim}{y}}$.

It is important to note that if at least one of the column of the X matrix is linearly related to any other columns of X, the matrix $X'X$ is singular and therefore $(X'X)^{-1}$ matrix is undefined. In this case there is a multicolinearity problem in the data. In this class we are not going to discuss the multicolinearity problem.

<u>Notes</u>: If a k variate random vector $\underset{\sim}{X} = (X_1, X_2, ..., X_k)'$ has mean $E(\underset{\sim}{X}) = \underset{\sim}{\mu}$ and the variance covariance matrix Σ then

- $E(a + \underset{\sim}{b}'\underset{\sim}{X}) = a + \underset{\sim}{b}'E(\underset{\sim}{X}) = a + \underset{\sim}{b}'\underset{\sim}{\mu}$ and $Var(a + \underset{\sim}{b}'\underset{\sim}{X}) = \underset{\sim}{b}'Var(\underset{\sim}{X})\underset{\sim}{b} = \underset{\sim}{b}'\Sigma\underset{\sim}{b}$
- when $\underset{\sim}{X}$ is a multivariate normally distributed random vector with the mean vector $\underset{\sim}{\mu}$ and variance covariance matrix Σ ($\underset{\sim}{X} \sim N(\underset{\sim}{\mu}, \Sigma)$) then $a + \underset{\sim}{b}'\underset{\sim}{X} \sim N(a + \underset{\sim}{b}'\underset{\sim}{\mu}, \underset{\sim}{b}'\Sigma\underset{\sim}{b})$

.

Under the normality of the error term ($\underset{\sim}{e} \sim N(\underset{\sim}{0}, \sigma^2 I_n)$) the dependent vector is also normally distributed ($\underset{\sim}{y} \sim N(X\underset{\sim}{\beta}, \sigma^2 I_n)$) and Since the OLS estimator $\hat{\underset{\sim}{\beta}}$ is a linear combination of $\underset{\sim}{y}$, $\hat{\underset{\sim}{\beta}}$ is also normally distributed random vactor with the mean vector and the variance covariance matrix $\underset{\sim}{\beta}$ and $\sigma^2(X'X)^{-1}$ respectively because

$$E(\hat{\underset{\sim}{\beta}}) = E((X'X)^{-1}X'\underset{\sim}{y}) = (X'X)^{-1}X'E(\underset{\sim}{y}) = (X'X)^{-1}X'X\underset{\sim}{\beta} = \underset{\sim}{\beta}$$

and

$$\begin{aligned} Var(\hat{\underset{\sim}{\beta}}) &= Var((X'X)^{-1}X'\underset{\sim}{y}) = (X'X)^{-1}X'Var(\underset{\sim}{y})X(X'X)^{-1} \\ &= (X'X)^{-1}X'\sigma^2 I_n X(X'X)^{-1} = \sigma^2(X'X)^{-1}. \end{aligned}$$

Therefore "similar to the simple linear regression", we can do any statistical inferences (hypothesis testing, confidence intervals etc.) for the regression parameters.

Example: Assume that there is a linear relationship between the test scores (Y) and the IQ level (X_1), study hour (X_2) for preparation to the examination. That is, we assume that the following data set is appropriate for a multiple linear regression model. In the model the third variable X_3 is just the product of X_1 and X_2 ($X_3 = X_1 X_2$) which is known as the interaction term. The data contains variables Y (test score), X_1 (IQ level) and X_2 (working hour) and we want to look at the contributions of the explanatory variables on the test scores. The model we want to investigate is

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{1,i} x_{2,i} + e_i \ , \ i = 1,2,\dots,8 .$$

X_1	X_2	Y	X_1X_2	X_1Y	X_2Y	X_1^2	X_2^2
105	10	75	1050	7875	750	11025	100
110	12	79	1320	8690	948	12100	144
120	6	68	720	8160	408	14400	36
116	13	85	1508	9860	1105	13456	169
122	16	91	1952	11102	1456	14884	256
130	8	79	1040	10270	632	16900	64
114	20	98	2280	11172	1960	12996	400
102	15	76	1530	7752	1140	10404	225

In the matrix notation a regression model can be written as

$$\underset{\sim}{y} = X \underset{\sim}{\beta} + \underset{\sim}{e}$$

with the following matrices:

$$\underset{\sim}{y} = \begin{bmatrix} 75 \\ 79 \\ 68 \\ 85 \\ 91 \\ 79 \\ 98 \\ 76 \end{bmatrix} , \quad X = \begin{bmatrix} 1 & 105 & 10 & 1050 \\ 1 & 110 & 12 & 1320 \\ 1 & 120 & 6 & 720 \\ 1 & 116 & 13 & 1508 \\ 1 & 122 & 16 & 1952 \\ 1 & 130 & 8 & 1040 \\ 1 & 114 & 20 & 2280 \\ 1 & 102 & 15 & 1530 \end{bmatrix}, \quad \underset{\sim}{\beta} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} , \quad \underset{\sim}{e} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \end{bmatrix}$$

Using these matrices we calculate

$$X'X = \begin{bmatrix} 8 & 919 & 100 & 11400 \\ 919 & 106165 & 11400 & 1306102 \\ 100 & 11400 & 1394 & 158366 \\ 11400 & 1306102 & 158366 & 18068568 \end{bmatrix} , \quad X'\underset{\sim}{y} = \begin{bmatrix} 651 \\ 74881 \\ 8399 \\ 959682 \end{bmatrix}$$

and the OLS estimates of the parameters, the fitte values and residuals are

$$\underset{\sim}{\hat{\beta}} = (X'X)^{-1}X'\underset{\sim}{y} = \begin{bmatrix} 72.206 \\ -0.131 \\ -4.111 \\ 0.053 \end{bmatrix}, \qquad \underset{\sim}{\hat{y}} = X\underset{\sim}{\hat{\beta}} = \begin{bmatrix} 73.05 \\ 78.50 \\ 70.01 \\ 83.58 \\ 94.02 \\ 77.46 \\ 96.03 \\ 78.36 \end{bmatrix} \quad \text{and} \quad \underset{\sim}{\hat{e}} = \underset{\sim}{y} - \underset{\sim}{\hat{y}} = \begin{bmatrix} 1.95 \\ 0.50 \\ -2.01 \\ 1.42 \\ -3.02 \\ 1.54 \\ 1.97 \\ -2.36 \end{bmatrix}$$

Note that (considering the rounding error) we have the following results:

$$\sum_{i=1}^{8} y_i = \sum_{i=1}^{8} \hat{y}_i = 651.0, \qquad \sum_{i=1}^{8} \hat{e}_i = 0 \;, \quad \sum_{i=1}^{8} x_{1,i}\,\hat{e}_i = 0$$

$$\sum_{i=1}^{8} x_{2,i}\,\hat{e}_i = 0 \;, \qquad \sum_{i=1}^{8} x_{3,i}\,\hat{e}_i = 0, \qquad \sum_{i=1}^{8} \hat{y}_i\,\hat{e}_i = 0.$$

The fitted regrtession equation given below

$$\hat{Y}_i = 72.2 - 0.131x_{1,i} - 4.111x_{2,i} + 0.053x_{1,i}\,x_{2,i}, \quad i = 1,2,3,...,n$$

and the predicted values with residuals are given in the following table.

Y	75	79	68	85	91	79	98	76
$\hat{Y}$	73.046 5	78.497 6	70.010	83.576 8	94.0198	77.458 7	96.032 0	78.3585
$\hat{e}$	1.9534 6	0.5024 1	-2.010	1.4232 5	- 3.01983	1.5412 8	1.9679 7	-2.35854

Using these predicted values and residuals we calculate the following suma of squares as

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y}_n) = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}_n^2 = 641.875,$$

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_n) = \sum_{i=1}^{n} \hat{Y}_i^2 - n\bar{Y}_n^2 = 610.81033,$$

$$SSE = SST - SSR = 641.875 - 610.81033 = 31.06467.$$

Now we can construct the ANOVA table as

SoV	$d.f$	SS	MS	F
Regresyon	3	610.81033	610.81033	26.22
Artıklar	4	31.06467	7.76617	
Toplam	7	641.875.		

Notice that the value of R^2 is $R^2 = SSR/SST = 0.9516 \cong 0.95$. This means that almost 95% of all variability in Y is explained by the explanatory variables (X 's). Moreover, since

$$F_h = 26.22 > F^{0.05}(3,4) = 6.59$$

the model seems to be significant at 5% level (we reject the null hypotheisi $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ at 5% level). On the other hand, the standard variance of the interaction term is $s(\hat{\beta}_3) = 0.03858$ (because $s^2(\hat{\beta}_3) = MSE(X'X)^{-1}_{4,4}$ which implies that $s^2(\hat{\beta}_3) = MSE(0.0001916598) = 0.001488$ and thus the standard error is $s(\hat{\beta}_3) = 0.03858$). Let us try to test whether the interaction term is significant or not. In order to test whether the interaction term is significant or not, we need to test $H_0 : \beta_3 = 0$ against the alternative of $H_a : \beta_3 \neq 0$. If we reject this null hypothesis we can conclude that the interaction term (or the parameter β_3) is significant. The value of t statistics is calculated as

$$t_h = \hat{\alpha}_3 / s(\hat{\alpha}_3) = 0.053071 / 0.03858 = 1.376 .$$

and since $|t_h| = 1.376 < t_4(0.025) = 2.7667$ we fail to reject the null hypothesis. This means that the interaction term is insignificant at 5% level. In a similar way, we can test the other parameters ($H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$) and we notice that all three parameters are insignificant. the results are summarized in the following table.

Parameter	estimate	Stand. error	T: $H_0 : \beta_i = 0$	5% critical value	result
β_1	-0.131170	0.45529954	-0.288	2.7667	Accept H_0
β_2	-4.111072	4.52430095	-0.909	2.7667	Accept H_0
β_3	0.053071	0.03858059	1.376	2.7667	Accept H_0

According to the table values, there seems to be a contradiction because when we want to test $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ at 5% level we rejected the null (means that all three parameters are not zero) however if we want to test these parameters seperately we failed to reject the null hypotheses. Moreover, we can see (from the table) that the value of R^2 is quite large. Actually this is not a contradiction because rejecting (or failing to reject) the null of $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ does not imply to reject the null of $H_0 : \beta_1 = 0$ (or others). Similarly, rejecting (or failing to reject) $H_0 : \beta_i = 0$ for all i does not imply to reject (or fail to reject) $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$.

The SAS codes and output of the analysis are given in the following table. according to table, even we reject the null of $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ at 5% level ($F_h = 26.22 > F^{0.05}(3,4)$) we fail to reject the null of $H_0 : \beta_i = 0$ for all $i = 1,2,3$ (even the intercept term) and if we notişce that the percentage of the variability explaned by the model (the value of R^2) is quite large, more that 95%.

```
data a; input x1 x2 y; x3=x1*x2; cards;
105 10 75
110 12 79
120  6 68
116 13 85
122 16 91
130  8 79
114 20 98
102 15 76
;
proc reg; model y=x1 x2 x3;
output out=out residual=ehat predicted=yhat; proc print data=out; run;
```

****

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	610.81033	203.60344	26.22	0.0043
Error	4	31.06467	7.76617		
Corrected Total	7	641.87500			

****

Root MSE 2.78678 R-Square **0.9516**

Dependent Mean	81.37500	Adj R-Sq	0.9153
Coeff Var	3.42462		

****

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	72.20608	54.07278	1.34	0.2527
x1	1	-0.13117	0.45530	-0.29	0.7876
x2	1	-4.11107	4.52430	-0.91	0.4149
x3	1	0.05307	0.03858	1.38	0.2410

****

Obs	x1	x2	y	x3	yhat	ehat
1	105	10	75	1050	73.0465	1.95346
2	110	12	79	1320	78.4976	0.50241
3	120	6	68	720	70.0100	-2.01000
4	116	13	85	1508	83.5768	1.42325
5	122	16	91	1952	94.0198	-3.01983
6	130	8	79	1040	77.4587	1.54128
7	114	20	98	2280	96.0320	1.96797
8	102	15	76	1530	78.3585	-2.35854

Now we consider the regression model without an intercept term as

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + e_i \ , \ i = 1, 2, ..., 8 .$$

According to this model the parameters are now significant (intercept term is still insignicicant, the cprresponding p-value is large). The SAS codes and ouput for this model is given below. The value of R^2 decreased from 95% to 93%. That is there is a little loss from the percentage of the variability. This is always the case because if you add a new explanatory variable to the model, the value of R^2 increases (here the loss is very little). The main question here is to search the contributions of the explanatory variables to the dependent variables. That is, we want to calculate the partial coefficient of determination(R^2 is also known as the coefficient of total determination).

```
data a; input x1 x2 y;
x3=x1*x2;
cards;
105 10 75
110 12 79
120  6 68
116 13 85
122 16 91
130  8 79
114 20 98
102 15 76
;
proc reg; model y=x1 x2;
output out=out residual=ehat predicted=yhat;
proc print data=out;  run;
**************************************************************************
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	596.11512	298.05756	32.57	0.0014
Error	5	45.75988	9.15198		
Corrected Total	7	641.87500			

```
**************************************************************************
Root MSE            3.02522    R-Square     0.9287
```

Dependent Mean	81.37500	Adj R-Sq	0.9002
Coeff Var	3.71763		

****

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.73655	16.26280	0.05	0.9656
x1	1	0.47308	0.12998	3.64	0.0149
x2	1	2.10344	0.26418	7.96	0.0005

****

Obs	**x1**	**x2**	**y**	**x3**	**yhat**	**ehat**
1	105	10	75	1050	71.4447	3.55529
2	110	12	79	1320	78.0170	0.98300
3	120	6	68	720	70.1272	-2.12722
4	116	13	85	1508	82.9589	2.04106
5	122	16	91	1952	92.1077	-1.10775
6	130	8	79	1040	79.0649	-0.06493
7	114	20	98	2280	96.7368	1.26318
8	102	15	76	1530	80.5426	-4.54264

In order to calculate the partial determination we need to define ther partial sums of squares. We now consider the following four models:

Model I $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + e_i$, $i = 1,2,3,...,n$

Model II $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + e_i$, $i = 1,2,3,...,n$

Model III $Y_i = \beta_0 + \beta_1 x_{1,i} + e_i$, $i = 1,2,3,...,n$

Model IV $Y_i = \beta_0 + \beta_2 x_{2,i} + e_i$, $i = 1,2,3,...,n$

and according to these differfent models we calculate regression sum of squares and errror sum of squares.

- For model I, regression sum of squares and error sum of squares are denoted by $SSR(X_1, X_2, X_3)$ and $SSE(X_1, X_2, X_3)$
- For model II, regression sum of squares and error sum of squares are denoted by $SSR(X_1, X_2)$ and $SSE(X_1, X_2)$
- For model III, regression sum of squares and error sum of squares are denoted by $SSR(X_1)$ and $SSE(X_1)$
- For model II, regression sum of squares and error sum of squares are denoted by $SSR(X_2)$ and $SSE(X_2)$.

Generally, the partial sum of squares, for example the partial sum of squares related to the explanatory variable X_2 when there are three explanatory variables in the model (Model I) is defined as

$$SSR(X_2 \mid X_1, X_3) = SSE(X_1, X_3) - SSE(X_1, X_2, X_3) .$$

Using this partial sum of squares, the partial coefficient of determination (the percentage of the variability explained by the explanatory variable X_2 when there are two more explanatory variables X_1 and X_2) is defined as

$$r^2_{Y2.13} = \frac{SSR(X_2 \mid X_1, X_3)}{SSE(X_1, X_3)} .$$

Using these sums of squares, we can define the partial sums of squares. Here, we assume that model I is the full model. Actually there are two types of partial sum of squares. One is known as Type I SS or squential SS or additive sums of square.

<u>Type I SS</u>:

Consider the full model

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + e_i \text{ , } i = 1, 2, 3, ..., n$$

and calculate $SSR(X_1)$ and $SSE(X_1)$ from Model III

calculate $SSR(X_1, X_2)$ and $SSE(X_1, X_2)$ from Model II

calculate $SSR(X_1, X_2, X_3)$ and $SSE(X_1, X_2, X_3)$ from Model I

then the sequential SS are calculated as follows:

X_1 : $SSR(X_1)$

X_2: $SSR(X_2 \mid X_1) = SSE(X_1) - SSE(X_1, X_2)$

X_3: $SSR(X_3 \mid X_1, X_2) = SSE(X_1, X_2) - SSE(X_1, X_2, X_3)$

Note that

$$
\begin{aligned}
&SSR(X_1)+SSR(X_2 \mid X_1)+SSR(X_3 \mid X_1,X_2)\\
&\qquad = SSR(X_1)+[SSE(X_1)-SSE(X_1,X_2)]+[SSE(X_1,X_2)-SSE(X_1,X_2,X_3)]\\
&\qquad = SST-\cancel{SSE(X_1)}+\cancel{SSE(X_1)}-\cancel{SSE(X_1,X_2)}+\cancel{SSE(X_1,X_2)}-SSE(X_1,X_2,X_3)\\
&\qquad = SST-SSE(X_1,X_2,X_3)=SSR(X_1,X_2,X_3)
\end{aligned}
$$

and therefore the sequential SS's are additive.

<u>Type II SS</u>:

In order to calculate Type II sums of squares we consider the full model as

$$Y_i=\beta_0+\beta_1 x_{1,i}+\beta_2 x_{2,i}+\beta_3 x_{3,i}+e_i \ , \ i=1,2,3,\ldots,n$$

and calculate Type II SS's as

X_1 : $SSR(X_1 \mid X_2,X_3)=SSE(X_2,X_3)-SSE(X_1,X_2,X_3)$

X_2 : $SSR(X_2 \mid X_1,X_3)=SSE(X_1,X_3)-SSE(X_1,X_2,X_3)$

X_3 : $SSR(X_3 \mid X_1,X_2)=SSE(X_1,X_2)-SSE(X_1,X_2,X_3)$.

Notice that these SS's are not additive.

<u>Example</u>: Connsider the previous example and we run four models given above. The ANOVA tables for these models are given below. In order to calculate especially for Type II SS's, we also need $SSR(X_1,X_3)$, $SSE(X_1,X_3)$, $SSR(X_2,X_3)$ and $SSE(X_2,X_3)$. These values are also calculated by running two more regression equations and given below. In a summary, we have the following results:

$SSR(X_1,X_2,X_3)=610.81033$	,	$SSE(X_1,X_2,X_3)=31.06467$	,	$SST=641.875$
$SSR(X_1,X_2)=596.11512$	,	$SSE(X_1,X_2)=45.75988$		
$SSR(X_1)=15.9393$	,	$SSE(X_1)=625.9357$		
$SSR(X_2)=474.87674$	,	$SSE(X_2)=166.99826$		
$SSR(X_1,X_3)=604.39802$	,	$SSE(X_1,X_3)=37.47698$		
$SSR(X_2,X_3)=610.16574$	,	$SSE(X_2,X_3)=31.70926$		

Now, we can calculate the Type I and Type II sums of squares are calculated as follows.

<u>Type I SS's</u>:

X_1 : $SSR(X_1)=15.9393$

X_2 : $SSR(X_2 \mid X_1)=SSE(X_1)-SSE(X_1,X_2)=625.9357-45.75988=580.17582$

X_3 : $SSR(X_3 \mid X_1,X_2)=SSE(X_1,X_2)-SSE(X_1,X_2,X_3)=45.75988-31.06467=14.69521$

Type II SS’s:

X_1 : $SSR(X_1 \mid X_2, X_3) = SSE(X_2, X_3) - SSE(X_1, X_2, X_3) = 31.70926 - 31.06467 = 0.64459$

X_2 : $SSR(X_2 \mid X_1, X_3) = SSE(X_1, X_3) - SSE(X_1, X_2, X_3) = 37.47698 - 31.06467 = 6.41231$

X_3 : $SSR(X_3 \mid X_1, X_2) = SSE(X_1, X_2) - SSE(X_1, X_2, X_3) = 45.75988 - 31.06467 = 14.69521$

Therefore using the Type I SS’ss and Type II SS’s, the partial determinations of the explanatory variables X_1, X_2 and X_3 are calculated as follows.

If we have three explanatory variables in the regression model given by

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + e_i \text{ , } i = 1,2,3,...,n$$

the multiple coefficient of determination is defined as $R^2 = SSR(X_1, X_2, X_3) / SST$. Let us show this multiple coefficient of determination as

$$R^2 = \frac{SSR(X_1, X_2, X_3)}{SST} = R^2_{Y.123} .$$

Using the partial sums of squares, the partial coefficients of determinations care calculated according to ***Type I SS’s*** as

$$r^2_{Y.1} = \frac{SSR(X_1)}{SST} \quad , \quad r^2_{Y2.1} = \frac{SSR(X_2 \mid X_1)}{SSE(X_1)} \quad , \quad r^2_{Y3.12} = \frac{SSR(X_3 \mid X_1, X_2)}{SSE(X_1, X_2)}$$

and the values are

$$r^2_{Y.1} = \frac{SSR(X_1)}{SST} = \frac{15.939299}{641.875} \cong 0.0248$$

$$r^2_{Y2.1} = \frac{SSR(X_2 \mid X_1)}{SSE(X_1)} = \frac{580.175816}{625.9357} = 0.92689$$

$$r^2_{Y3.12} = \frac{SSR(X_3 \mid X_1, X_2)}{SSE(X_1, X_2)} = \frac{14.695210}{45.75988} \cong 0.321.$$

Note that more than 95% of all variability in Y is explained by the model. That is,

$$R^2 = SSR(X_1, X_2, X_3) / SST = R^2_{Y.123} = 0.9516$$

and among all these variability more than 92% of all variability is explaned by only X_2. In other words, the IQ level has no effect (only about 2.5%) on the test scores.

Finally, we consider the multiple linear regression model given above. Suppose we want to do some statistical inferences about the parameters. We consider the regression model as

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + e_i \text{ , } i = 1,2,3,...,8$$

The values of the OLS estimators and their standard errors are given in the following table.

$\hat{\beta}_0 = 0.73655$, $s(\hat{\beta}_0) = 16.2628$, $\hat{\beta}_1 = 0.47308$, $s(\hat{\beta}_1) = 0.12998$, $\hat{\beta}_2 = 2.10344$, $s(\hat{\beta}_1) = 0.26418$

Let us write 95% confidence interval for the model parameters β_0 , β_1 and β_2 . Note that from $P(t_5 > t_5(0.025)) = 0.025$ we find the critical value from the t table as $t_5(0.025) = 2.571$. The $(1-\alpha)100\%$ confidence intervals for the parameters can be calculated as $\hat{\beta}_i \pm s(\hat{\beta}_i) t_{n-3}(\alpha / 2)$ for $i = 0,1,2$. Here, p is the number of parameters in the model.

Therefore a 95% confidence interval for β_0 is

$$\hat{\beta}_0 \pm s(\hat{\beta}_0) t_5(0.025) \Leftrightarrow 0.73655 \pm (16.268)(2.571) \Leftrightarrow (-41.09, 42.57)$$

Note that it is a very wide confidence interval for the intercept term. This is meaningful because the intercept term is insignificant because we failed to reject the null hypothesis of $H_0 : \beta_0 = 0$.

A 95% confidence interval for β_1

$$\hat{\beta}_1 \pm s(\hat{\beta}_1) t_5(0.025) \Leftrightarrow 0.473 \pm (0.13)(2.57) \Leftrightarrow (0.139, 0.807)$$

and for β_2

$$\hat{\beta}_2 \pm s(\hat{\beta}_2) t_5(0.025) \Leftrightarrow 2.103 \pm (0.26)(2.57) \Leftrightarrow (1.435, 2.771) .$$

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	610.81033	203.60344	26.22	0.0043
Error	4	31.06467	7.76617		
Corrected Total	7	641.87500			

Root MSE	2.78678	R-Square	0.9516
Dependent Mean	81.37500	Adj R-Sq	0.9153
Coeff Var	3.42462		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	72.20608	54.07278	1.34	0.2527
x1	1	-0.13117	0.45530	-0.29	0.7876
x2	1	-4.11107	4.52430	-0.91	0.4149
x3	1	0.05307	0.03858	1.38	0.2410

****

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	596.11512	298.05756	32.57	0.0014
Error	5	45.75988	9.15198		
Corrected Total	7	641.87500			

Root MSE	3.02522	R-Square	0.9287
Dependent Mean	81.37500	Adj R-Sq	0.9002
Coeff Var	3.71763		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.73655	16.26280	0.05	0.9656
x1	1	0.47308	0.12998	3.64	0.0149
x2	1	2.10344	0.26418	7.96	0.0005

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	15.93930	15.93930	0.15	0.7094
Error	6	625.93570	104.32262		
Corrected Total	7	641.87500			

Root MSE	10.21384	R-Square	0.0248
Dependent Mean	81.37500	Adj R-Sq	-0.1377
Coeff Var	12.55158		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	62.57113	48.24164	1.30	0.2423
x1	1	0.16369	0.41877	0.39	0.7094

****

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	474.87674	474.87674	17.06	0.0061
Error	6	166.99826	27.83304		
Corrected Total	7	641.87500			

Root MSE	5.27570	R-Square	0.7398		
Dependent Mean	81.37500	Adj R-Sq	0.6965		
Coeff Var	6.48320				

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	58.67535	5.80344	10.11	<.0001
x2	1	1.81597	0.43964	4.13	0.0061

Note that the intercept term is insignificant either by considering the inretactin term or not. Therefore it is reasonable to consider a regression model without having an intercept term. Another point is to note that the interaction term is insignificant. That's why we consider a regression model with two explanatory variables (IQ level and study-hour) as

$$Y_i = \beta_1 x_{1,i} + \beta_2 x_{2,i} + e_i \text{ , } i = 1,2,3,...,8 .$$

The OLS estimators of the parameters and some statistical results with the ANOVA table is in the following table. According to this model (model without intercept) all the parameters are significant and the percentage of the variability increased from 95% to more that 99%. On the other hand, if the model includes an intercept term, we know that the residuals are orthogonal to the explanatory variables and the predicted values. Moreover the sum of the residuals is zero. However if the model does not inclue an intercept term this may not be true. remember that if the model includes an intercept term we always have

$$\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i \text{ ,} \qquad \sum_{i=1}^{n} \hat{e}_i = 0 \text{ ,} \qquad \sum_{i=1}^{n} \hat{e}_i \hat{y}_i = 0 \text{ ,} \qquad \sum_{i=1}^{n} \hat{e}_i x_{1,i} = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} \hat{e}_i x_{2,i} = 0$$

and when we consider the case without an intercept term we have the following sums

- $\sum_{i=1}^{n} y_i = 651$ and $\sum_{i=1}^{n} \hat{y}_i = 650.974512$ so that $\sum_{i=1}^{n} \hat{y}_i \neq \sum_{i=1}^{n} y_i$

- $\sum_{i=1}^{n} \hat{e}_i = 0.00319595$ so that $\sum_{i=1}^{n} \hat{e}_i \neq 0$
- $\sum_{i=1}^{n} \hat{e}_i \hat{y}_i = 0.00012761$, $\sum_{i=1}^{n} \hat{e}_i x_{1,i} = 0.00018375$ and $\sum_{i=1}^{n} \hat{e}_i x_{2,i} = 0.00001$

```
data a; input x1 x2 y;
cards;
105 10 75
110 12 79
120  6 68
116 13 85
122 16 91
130  8 79
114 20 98
102 15 76
;
proc reg; model y=x1 x2/noint ss1 ss2;
output out=out residual=ehat predicted=yhat; proc print data=out;  run;
*****************************************************************************************
```

NOTE: No intercept in model. R-Square is redefined.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	53571	26786	3510.67	<.0001
Error	6	45.77866	7.62978		
Uncorrected Total	8	53617			

```
*****************************************************************************************
```

Root MSE 2.76220 R-Square 0.9991

Dependent Mean 81.37500 Adj R-Sq 0.9989

Coeff Var 3.39441

****

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
x1	1	0.47885	0.02429	19.72	<.0001	52816	2966.34258
x2	1	2.10915	0.21193	9.95	<.0001	755.65914	755.65914

****

Obs	x1	x2	y	yhat	ehat
1	105	10	75	71.3703	3.62967
2	110	12	79	77.9829	1.01714
3	120	6	68	70.1164	-2.11642
4	116	13	85	82.9651	2.03491
5	122	16	91	92.1656	-1.16562
6	130	8	79	79.1232	-0.12318
7	114	20	98	96.7714	1.22855
8	102	15	76	80.4795	-4.47955

Model Selection:

Any dependent variable may be affected by many explanatory variables. The goal in model building is to select the best set of explanatory variables (in statictically or economically). There are many statistical methods to select such a set of explanatory variables bu here we are going to investigate the simple and applicable one. Having more explanatory variables in the model may cause many problems. Therefore, it is important to choose the best set of explanatory variables in the model. Adding a new explanatory variable to the model, the percentage of the variability increases (the value of R^2 increases). However, adding a new variable to the model will have a cost (either economically or statistically) to pay. That's why we need to built models with a minimum cost. There are many statistical methods to built such models (for example, choose the models with have the smallest value of AIC statistic, or SBC statistic. These statistical techniques choose models with minimize the cost or penalty). In this class we are not going to discuss such techniques.

In some cases, adding a new explanatory variable to the model may cause statistical problem. For example, even the value of percentage of variablility increases a significant parameter may turn out to be insignificant (or wise versa). Therefore, adding such an explanatory variable is not meaningful.

Consider a linear regression equation with 3 explanatory variables

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + e_i \ , \ i = 1,2,3,...,n \ . \qquad \textbf{Model I}$$

If we run this regression we can calculate the value of R^2 and the values of OLS estimators of the regression parameters β_i's. Suppose we add a new explanatory variable to the model and write as

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,i} + e_i \ , \ i = 1,2,3,...,n \ . \qquad \textbf{Model II}$$

We can either start from Model II by eliminating the expanatory variables (backward selection) or we can start from Model I by adding a new variable (forward selection) to get sa significant model. The most practical way of selection of a suitable model, we start with the first explanatory variable and start to add a new variable and notice all the statistical properties. It is very similar to do same analyses starting with all explanatory variables and we eliminate insignificant explanatory variablesstep-by-step. This technique is known as the *stepwise* regression approach. As it is mentioned, there are many model selection (selection of significant explanatory variables) criteria.

Consider a multiple regression equation with p explanatory variables as

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_p x_{p,i} + e_i \ , \ i = 1,2,3,...,n \ .$$

When we run this regression equation some of the explanatory variables may not be significat. Let x_2 , x_5 and x_8 be the variables seem to be insignificant. As it is seen in the above example, we may fail to reject these hypotheses individually eventhough the whole model is significant. That is, we may fail to reject $H_0 : \beta_2 = 0$, $H_0 : \beta_5 = 0$ and $H_0 : \beta_8 = 0$ individually but we may reject $H_0 : \beta_2 = \beta_5 = \beta_8 = 0$ at the same significance level. To make it more clear let us consider the model (call this full model)

$$Full\ Model: \quad Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,i} + e_i \ , i = 1,2,3,...,n$$

and check whether x_2 and x_2 are significant or not at the same time. That is, we may want to test $H_0 : \beta_2 = \beta_3 = 0$ or not. We may or may not reject (or fail to reject) $H_0 : \beta_2 = 0$ and $H_0 : \beta_3 = 0$ separately. In order to test the null hypothesis $H_0 : \beta_2 = \beta_3 = 0$ we write the reduced model

$$Reduced\ Model: \quad Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_4 x_{4,i} + e_i \ , i = 1,2,3,...,n$$

The ANOVA table can be constructed according to full and reduced models. Let $SSR(X_1, X_2, X_3, X_4)$ and $SSE(X_1, X_2, X_3, X_4)$ denote the regression sum of squares and error sum of squares under the full model (re-name these SS's as $SSR(full)$ and $SSE(full)$). Similarly, we can calculate the regression sum of squares and error sum of squares according to reduced model (say $SSR(red)$ and $SSE(red)$) and in order to test $H_0 : \beta_2 = \beta_3 = 0$ we define the F statistic

$$F = \frac{[SSE(red) - SSE(fell)]/2}{MSE(full)} \ .$$

Under the null hypothesis $H_0 : \beta_2 = \beta_3 = 0$, the F statistic is distributed as F with 2 and $(n-5)$; therefore we reject the null of $H_0 : \beta_2 = \beta_3 = 0$ at the level α if $F_h > F^{\alpha/2}(2, n-5)$. If we have only one parameter to estimate, the value of F statistic is the same as square of the t statistic (that is, if $X \sim t_p$ then $X^2 \sim F(1, p)$).

Example: A personnel officer in a governmental agency administered four newly developed attitute test to each of 25 applicants for entry-level clerical positions in the agency. For purposes of the study, all 25 applicants were accepted for positions irrrespective of their test scores. After a probationary period, each applicant was rated for proficiency on the job. It is expected that a regression model containing only first-order terms and no interaction term will be appopriate. That is we want to consider a regression model as

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,i} + e_i \ , i = 1,2,3,...,25. \quad (1)$$

Here we want to find the best possible set of explanatory variables. Actually, according to the ANOVA tables given below the second test seems to be insingificant. Therefore, if we can afford three variables in the model best model is the appropriate one. That is, the possible model is

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_3 x_{3,i} + \beta_4 x_{4,i} + e_i \ , i = 1,2,3,...,25 \quad (2)$$

because, $SSE(full) \cong 335.98$ and $MSE(full) \cong 16.8$, when we run the reduced model we have $SSE(red) \cong 348.197$. In order to test wherher model (2) is significant againt the alternative the oppropriate model is (2) the value of F statistic

$$F_h = \frac{[SSE(red) - SSE(fell)]/1}{MSE(full)} = \frac{(348.197 - 335.98)}{16.80} \cong 0.73$$

and thus we fail to reject the null hypothesis of $H_0 : \beta_2 = 0$ because $F_h = 0.73 < F^{0.05}(1,20)$. That is model (2) is significant.

The scores on the four tests (X_1, X_2, X_3, X_4) and the job proficiency score (Y) for 25 employees were as foloows:

	Test Score				Job Proficiency Score
Subject	X_1	X_2	X_3	X_4	Y
1	86	110	100	87	88
2	62	97	99	100	80
3	110	107	103	103	96
4	101	117	93	95	76
5	100	101	95	88	80
6	78	85	95	84	73
7	120	77	80	74	58
8	105	122	116	102	116
9	112	119	106	105	104
10	120	89	105	97	99
11	87	81	90	88	64
12	133	120	113	108	126
13	140	121	96	89	94
14	84	113	98	78	71
15	106	102	109	109	111

16	109	129	102	108	109
17	104	83	100	102	100
18	150	118	107	110	127
19	98	125	108	95	99
20	120	94	95	90	82
21	74	121	91	85	67
22	96	114	114	103	109
23	104	73	93	80	78
24	94	121	115	104	115
25	91	129	97	83	83

In the obe discussion, we observe that the possible model is model (2) if we can efford three variables in the model.

Suppose, if it is possible we want to eliminate one more explanatory variable from the model. The first variable is the second test we eliminate. That is, we need to test

- $H_0 : \beta_1 = \beta_2 = 0$ (to eliminate X_1 and X_2)
- $H_0 : \beta_2 = \beta_3 = 0$ (to eliminate X_2 and X_3)
- $H_0 : \beta_2 = \beta_4 = 0$ (to eliminate X_2 and X_4).

The correpondig sum of squares to calculate the value F statistics are

$SSE(X_3, X_4) = 1111.3126$, $SSE(X_1, X_4) = 1672.58526$, $SSE(X_1, X_3) = 606.65745$.

The values of the F statistics are

$$F_{1,h} = \frac{[SSE(red) - SSE(fell)]/2}{MSE(full)} = \frac{(1111.31 - 335.98)/2}{16.80} \cong 23.07$$

$$F_{2,h} = \frac{[SSE(red) - SSE(fell)]/2}{MSE(full)} = \frac{(1672.59 - 335.98)/2}{16.80} \cong 39.78$$

$$F_{3,h} = \frac{[SSE(red) - SSE(fell)]/2}{MSE(full)} = \frac{(606.66 - 335.98)/2}{16.80} \cong 8.056$$

and the critical value is $F^{0.05}(2,20) = 3.49$. Since, $F_{i,h} > F^{0.05}(2,20)$ we reject all these three hypothesis. This means we can not eliminate one more explanatory variable. Finally, the most appopriate model is

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_3 x_{3,i} + \beta_4 x_{4,i} + e_i \ , \ i = 1,2,3,...,25 .$$

As it is mentioned above, adding a new variable to the model the value of R^2 increases. However, adding a new variable may cause some statistical problems. First, we consider the full model given below.

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,i} + e_i \ , \ i = 1,2,3,...,25 \quad (3)$$

and the results of the regression analysis of data is summarized in Table 1.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	8718.02248	2179.50562	129.74	<.0001
Error	20	335.97752	16.79888		
Corrected Total	24	9054.00000			

****

Root MSE	4.09864	R-Square	**0.9629**
Dependent Mean	92.20000	Adj R-Sq	0.9555
Coeff Var	4.44538		

****

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	-124.38182	9.94106	-12.51	<.0001	212521	2629.83427
x1	1	0.29573	0.04397	6.73	<.0001	2395.85466	759.83030
x2	1	0.04829	0.05662	0.85	0.4038	1806.96541	12.21949
x3	1	1.30601	0.16409	7.96	<.0001	4254.45924	1064.15000
x4	1	0.51982	0.13194	3.94	0.0008	260.74317	260.74317

Table 1.

First it is important to note that the model is significant. That is, we reject the null hypothesis of $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ at 5% level (the value of F statistic is large or the corresponding p values is very small). The value of R^2 is very high (R^2=0.9626). All the parameters except β_2 are significant. Therefore we eliminate the second test (X_2) and consider a new model

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_3 x_{3,i} + \beta_4 x_{4,i} + e_i \ , \ i = 1, 2, 3, ..., 25 \tag{4}$$

the results according to the model given in (2) are in the Table 2.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	8705.80299	2901.93433	175.02	<.0001
Error	21	348.19701	16.58081		
Corrected Total	24	9054.00000			

****

Root MSE	4.07195	R-Square	0.9615
Dependent Mean	92.20000	Adj R-Sq	0.9560
Coeff Var	4.41644		

****

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	-124.20002	9.87406	-12.58	<.0001	212521	2623.35826
x1	1	0.29633	0.04368	6.78	<.0001	2395.85466	763.11559
x3	1	1.35697	0.15183	8.94	<.0001	6051.48790	1324.38825
x4	1	0.51742	0.13105	3.95	0.0007	258.46044	258.46044

Table 2.

According to Table 2. all the parameters are now significant and the model is again significant (the value of F statistics is large and corresponding p-value is small). Moreover, the value of R^2=0.9615 which is very close the the value for the full model. Therefore we can say that the second test (the X_2 variable) has no contribution to the model. That is, there is no statistical problem for eliminating the second test from the model.

Consider the folowing models:

Model I : $Y_i = \beta_0 + \beta_1 x_{1,i} + e_i$, $i = 1,2,3,...,25$

Model II : $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + e_i$, $i = 1,2,3,...,25$

Model III : $Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + e_i$, $i = 1,2,3,...,25$.

First, we run the Model I. The ANOVA table and related statistics are given in Table 3.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2395.85466	2395.85466	8.28	**0.0085**
Error	23	6658.14534	289.48458		
Corrected Total	24	9054.00000			

****

Root MSE	17.01425	R-Square	0.2646
Dependent Mean	92.20000	Adj R-Sq	0.2326
Coeff Var	18.45363		

****

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	41.32156	18.00985	2.29	0.0312
x1	1	0.49224	0.17111	2.88	**0.0085**

Table 3.

According to Table 3, the model is significant (the value of F statistic is big and the corresponding p-value is small). Moreover, the parameter β_1 is also significant. That is test 1 seems to be significant in the model. It is important to note that the value of R^2 is vary low even the parameter is significant. Therefore, the variable X_1 is significant but in order to improve the percentage of the variability we need to add new explanatory variable to the model. Thus, we consider Model II

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + e_i \ , \ i = 1,2,3,\ldots,25 \, .$$

The corresponding ANOVA table and some statistical values are given in Table 4 below.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4202.82007	2101.41003	9.53	0.0010
Error	22	4851.17993	220.50818		
Corrected Total	24	9054.00000			

****

Root MSE	14.84952	R-Square	0.4642
Dependent Mean	92.20000	Adj R-Sq	0.4155
Coeff Var	16.10577		

****

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-7.96667	23.31365	-0.34	0.7358
x1	1	0.44830	0.15012	2.99	0.0068
x2	1	0.50441	0.17621	2.86	0.0090

Table 4.

The investigation of Table 4 indicates that the model is still significant (the value of F statistic is large and the p-value is very small). Moreover, both variables seem to be significat. However there is a slight increas in the value of R^2. ***Here, it is important to note that in Model I, the intercept term was significant. However, when we add the second variable to the model the intercept term turned out to be insignificat***. Therfore the model needs to be improved. That is, we need to add a new explanatory variable to the model.

And thus, we consider model III

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + e_i \ , \ i = 1,2,3,...,25 .$$

The ANOVA table and some statistical values are given in Table 5. below. When we add the third variable to the model there is a significant increas in the value of R^2 (from 46% to 93%). that is, Model III has a large percentage of variability in the dependent variable Y. However, the significant variable X_2 in Model II turned out to be insignificant.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	8457.27931	2819.09310	99.21	<.0001
Error	21	596.72069	28.41527		
Corrected Total	24	9054.00000			

****

Root MSE	5.33060	R-Square	**0.9341**
Dependent Mean	92.20000	Adj R-Sq	0.9247
Coeff Var	5.78156		

****

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-127.77378	12.88053	-9.92	<.0001
x1	1	0.34813	0.05451	6.39	<.0001

x2	1	0.04353	0.07362	0.59	0.5606
x3	1	1.77921	0.14541	12.24	<.0001

Table 5.

Moreover, the intercept term is now significant. Therefore, the explanatory variable X_3 should be in the model. Moreover, if when we consider a regression model of Y on X_1 and X_3, both variables are significant and the value of R^2 is almost the same as in the Model (the value of R^2=0.9341 decreased to R^2=0.9330, a slight decrease). The ANOVA table and related statistical results are given in Table 6. below for the model

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_3 x_{3,i} + e_i \ , \ i = 1,2,3,...,25$$

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	8447.34255	4223.67128	153.17	<.0001
Error	22	606.65745	27.57534		
Corrected Total	24	9054.00000			

****

Root MSE	5.25122	R-Square	**0.9330**
Dependent Mean	92.20000	Adj R-Sq	0.9269
Coeff Var	5.69547		

****

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-127.59569	12.68526	-10.06	<.0001
x1	1	0.34846	0.05369	6.49	<.0001
x3	1	1.82321	0.12307	14.81	<.0001

Table 6.

According to Table 6. if we can efford two explanatory variables in the model, these variables should be X_1 and X_3. If we can efford one more variable, we can consider the full model as it is given in the equation (3), the value of is observed as R^2=0.9555 and all the parameters are significant except β_2.

As a conclution, if we want to use three explanatory variables in the multiple regression it should be

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_3 x_{3,i} + \beta_4 x_{4,i} + e_i \ , \ i = 1,2,3,...,25 .$$

However, if we have to eliminate one more explanatory variable the model should include X_1 and X_3 namely,

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_3 x_{3,i} + e_i \ , \ i = 1,2,3,...,25 .$$

In a summary, there is no contribution of X_2. In the multiple regression model a set of explanatory variables is $\{x_1, x_3\}$ or $\{x_1, x_3, x_4\}$ depending on the the number of explanatory variables desired to be used.