Regression Analysis and Ridge Regression

Outline of Talk

- Introduction
- Simple Linear Regression Model and Estimation of Parameters
- Multiple Linear Regression Model
- Least Squares Estimation
- Data Anomalies
- Multicollinearity
- Ridge Regression

Introduction

Regression analysis is a statistical tool for investigating the relationship between a dependent variable and one or more independent variables.

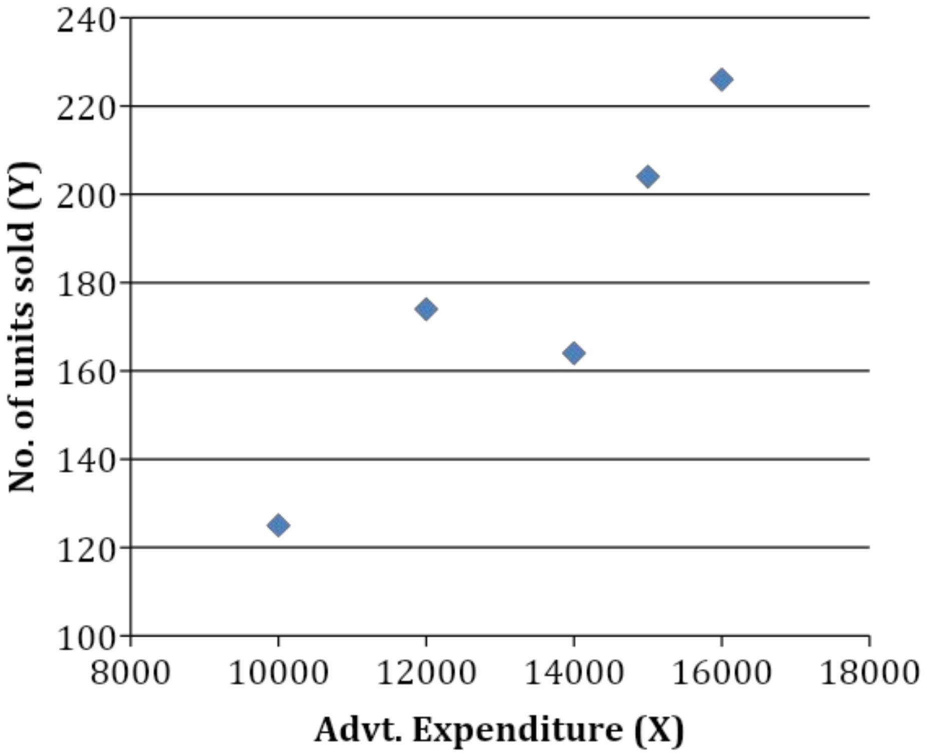
Suppose, as a business manager, you have a data regarding advertising expenditure (in Rs.) and sales (in number of units).

Advt. Expenditure (in Rs.)	10,000	12,000	14,000	16,000	15,000
Sales (no. of units)	125	174	164	226	204

In this case, Sales is depends upon the advertising expenditure. So sale is dependent variable also known as response variable and it is denoted by $\boldsymbol{Y}$. We can control the advertising expenditure so it is called independent or controlled or predictor of regressor variable. It is denoted by $\boldsymbol{X}$.

Introduction…

Now to investigate the relationship between a dependent variable and independent variable, we use **scatter plot**.



If the scatter plot indicates sort of linear relationship between the variables, we need to go for **linear model**.

But, if the scatter plot indicates sort of non linear relationship between X and Y then, we need to go for **non-linear model**.

Simple Linear Regression Model

In simple linear regression model, a single regressor variable (X) has linear relation with the response variable (Y). So, the mathematical form of simple linear regression model is

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Here, $Y =$ response variable

$X =$ regressor variable

$\beta_0 =$ intercept parameter

$\beta_1 =$ slope parameter

$\varepsilon =$ random error variable

Simple Linear Regression Model…

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i\,, \quad i = 1, 2, \ldots, n$$

Assumptions:

- ❑ ε_i is unobserved random variable with zero mean and unknown variance σ^2 i.e. $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{V}(\varepsilon_i) = \sigma^2$.

- ❑ ε_i and ε_j are uncorrelated i.e. $\mathrm{cov}(\varepsilon_i\,, \varepsilon_j) = 0$ for all $i \neq j$ $=1, 2, \ldots, n$.

- ❑ ε_i is normally distributed with zero mean and variance σ^2 i.e. $\varepsilon_i \sim \mathrm{N}\,(0, \sigma^2)$, $i = 1, 2, \ldots, n.$

Simple Linear Regression Model…

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i\,, \quad i = 1, 2, \ldots, n$$

$$\mathrm{E}\,(Y_i) = \mathrm{E}\,(\beta_0 + \beta_1 X_i + \varepsilon_i) = \beta_0 + \beta_1 X_i$$

$$\mathrm{V}\,(Y_i) = \mathrm{V}\,(\beta_0 + \beta_1 X_i + \varepsilon_i) = \mathrm{V}\,(\varepsilon_i) \;= \sigma^2$$

$$Y_i \sim \mathrm{N}(\beta_0 + \beta_1 X_i\,, \sigma^2)$$

Least Squares Estimation

Discovered independently by Gauss in Germany around 1795 and by Legendre in France around 1805 (Birkes and Dodge, 1993).

The regression parameters are obtained by minimizing the sum of squares of the differences between observed Y_i and the fitted line. That is $\sum_{i=1}^{n} e_i^2$ is minimum. Thus,

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \text{ is minimum}$$

After differentiating S partially with respect to β_0 and β_1, we get two normal equations. Hence, $\hat{\beta}_0$ and $\hat{\beta}_1$ are solutions of these two equations.

Least Squares Estimation…

$$\frac{\partial S}{\partial \beta_0} = 0 \Rightarrow -2\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right) = 0$$

$$\frac{\partial S}{\partial \beta_1} = 0 \Rightarrow -2\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)x_i = 0$$

After simplification, it gives

$$n\beta_0 + \beta_1\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\beta_0\sum_{i=1}^{n} x_i + \beta_1\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

The solution to above normal equations is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Illustrative Example

Sr. No.	X	Y	X^2	Y^2	XY
1	10000	125	100000000	15625	1250000
2	12000	174	144000000	30276	2088000
3	14000	164	196000000	26896	2296000
4	16000	226	256000000	51076	3616000
5	15000	204	225000000	41616	3060000
Total	67000	893	921000000	165489	12310000

$$\hat{\beta}_1 = \frac{5\times 12310000 - 67000\times 893}{5\times 921000000 - (67000)^2}$$

$$= 0.014819$$

$$\hat{\beta}_0 = \frac{893}{5} - 0.014819\times \left(\frac{67000}{5}\right)$$

$$= -19.9741$$

If we increase advt. expenditure by 1 rupee then the sale volume will be increased by 0.014819 units.

Multiple Linear Regression Model

Regression is most widely used statistical technique for investigation and modeling the relationship between variables.

The multiple linear regression model is

$$Y = X\beta + \varepsilon$$

where Y is an $n \times 1$ vector of response variable,
X is an $n \times k$ matrix of regressor variables with ones in the first column,
$\beta = (\beta_{0,} \beta_{1,} \beta_{2,} \ldots, \beta_{k-1})'$ is a vector of unknown regression coefficients and
ε is an $n \times 1$ vector of random errors.

Multiple Linear Regression Model…

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1(k-1)} \\ 1 & x_{21} & x_{22} & \vdots & x_{2(k-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n(k-1)} \end{bmatrix},$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{k-1} \end{bmatrix}, \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Multiple Linear Regression Model…

Assumptions

- ❑ **Form of the Model**
  - linear in regression parameters
- ❑ **Error**
  - E $(\varepsilon_i) = 0$
  - Var $(\varepsilon_i) = \sigma^2$
  - E $(\varepsilon_i, \varepsilon_j) = 0, \; i \neq j$
  - ε_i has normal distribution
- ❑ **Regressor Variables**
  - Non-random
  - Measured without error
  - Linearly independent

Chatterjee and Hadi (2006)

Multiple Linear Regression Model…

$$Y = X\beta + \varepsilon$$

$$\begin{aligned} \mathrm{E}\,(Y) &= \mathrm{E}\,(X\beta + \varepsilon) \\ &= X\beta \end{aligned}$$

$$\begin{aligned} \mathrm{COV}\,(Y) &= \mathrm{COV}\,(X\beta + \varepsilon) \\ &= \mathrm{COV}\,(\varepsilon) \\ &= \sigma^2 I \end{aligned}$$

$$Y \sim \mathrm{N_n}(X\beta\,,\, \sigma^2 I)$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_1 X_{i(k-1)} + \varepsilon_i$$

$$\begin{aligned} \mathrm{E}\,(Y_i) &= \mathrm{E}\,(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_1 X_{i(k-1)} + \varepsilon_i) \\ &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_1 X_{i(k-1)} \end{aligned}$$

$$\begin{aligned} \mathrm{V}\,(Y_i) &= \mathrm{V}\,(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_1 X_{i(k-1)} + \varepsilon_i) \\ &= \mathrm{V}\,(\varepsilon_i) \\ &= \sigma^2 \end{aligned}$$

$$Y_i \sim \mathrm{N}(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_1 X_{i(k-1)}\,,\, \sigma^2)$$

Least Squares Estimator

To find the vector of least squares estimator ☒, we minimize

$$☒☒(☒☒) = σ_{☒☒=1}^{☒☒} ☒☒_{☒☒}^{2} = ☒☒^{?}☒☒ = ☒☒☒☒ − ☒☒☒☒☒☒' ☒☒☒☒ − ☒☒☒☒☒☒$$

$$= ☒☒' ☒☒ − ☒☒' ☒☒' ☒☒ − ☒☒' ☒☒☒☒ + ☒☒' ☒☒' ☒☒☒☒$$

$$= ☒☒' ☒☒ − 2☒☒' ☒☒' ☒☒ + ☒☒ ☒☒' ☒☒☒☒$$

After differentiation ☒☒(☒☒) partially with respect to ☒ we get ☒ normal equations of the form

$$\frac{☒☒☒}{☒☒☒} = 0 \;☒\; −2☒☒' ☒☒ + 2☒☒' ☒☒☒ = 0$$

which simplifies to

$$☒☒?☒☒ = ☒☒?☒☒$$

Hence,

$$☒ = ☒☒?☒☒^{−1} ☒☒?☒☒.$$

Properties of Least Squares Estimator

- $\hat{\beta}$ is an unbiased estimator of the regression coefficient β.
- $\mathrm{Cov}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$.
- $\hat{\beta}$ is the best linear unbiased estimator (BLUE) in the class of all unbiased estimators of the regression coefficient β.

Unbiased estimator of σ^2

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - k - 1} = \frac{SS_{Res}}{n - k - 1} = MS_{Res}$$

Hat Matrix

The vector of fitted values $\hat{y}$ of the response variable y based on the LS estimator is defined as

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

where $H = X(X'X)^{-1}X'$ is the hat matrix also known as prediction matrix.

Some properties of H

- Hat matrix H is a symmetric and idempotent matrix.
- $(I - H)$ is also symmetric and idempotent matrix.
- $HX = X(X'X)^{-1}X'X = X$.
- $tr(H) = p$, $tr(\cdot)$ denote the trace of matrix.

Hypothesis Testing

Test for Significance of Regression

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_{k-1} = 0$$

$$H_1: \beta_j \neq 0 \text{ for at least one } j$$

Table 1 Analysis of Variance for Significance of Regression

Source of Variation	Sum of Squares	Degree of Freedom	Mean Squares	F_0 value
Regression	SS_R	k	MS_R	MS_R/MS_{Res}
Residual	SS_{Res}	n-k-1	MS_{Res}	
Total	SS_T	n-1		

$$SS_R = \hat{\beta}'X'y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}, \qquad SS_{Res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = y'y - \hat{\beta}'X'y,$$

$$SS_T = y'y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}.$$

Reject H_0 if $F_0 > F_{\alpha, k, (n-k-1)}$

Problems in the data

❑ **Form of the Model**

- linear in regression parameters

❑ **Error**

- E $(\varepsilon_i) = 0$
- Var $(\varepsilon_i) = \sigma^2$
- E $(\varepsilon_i, \varepsilon_j) = 0$, i $\neq$ j
- ε_i has normal distribution

❑ **Regressor Variables**

- Non-random
- Measured without error
- Linearly independent

Chatterjee and Hadi (2006)

Problems in the data…

- ❑ **Form of the Model**
  - linear in regression parameters
- ❑ **Error**
  - E $(\varepsilon_i) = 0$
  - Var $(\varepsilon_i) = \sigma^2$
  - E $(\varepsilon_i, \varepsilon_j) = 0$, i $\neq$ j
  - ε_i has normal distribution
- ❑ **Regressor Variables**
  - Non-random
  - Measured without error
  - Linearly independent

- Non-linearity
- Heteroscedasticity
- Autocorrelation
- Non-normality
- Outliers
- Multicollinearity

Chatterjee and Hadi (2006)

Some Illustrations

What is effect on regression analyses if predictors are perfectly uncorrelated?

x_1	x_2	y
2	5	52
2	5	43
2	7	49
2	7	46
4	5	50
4	5	48
4	7	44
4	7	43

Pearson correlation of x_1 and x_2 = 0.000

Some Illustrations…

Regress *Y* on *X_1*

The regression equation is y = 48.8 - 0.63 X_1

Predictor	Coef	SE Coef	T	P
Constant	48.750	4.025	12.11	0.000
X_1	-0.625	1.273	-0.49	0.641

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	3.13	3.13	0.24	0.641
Error	6	77.75	12.96		
Total	7	80.88			

Some Illustrations…

Regress Y on X_2

The regression equation is $y = 55.1 - 1.38\ X_2$

Predictor	Coef	SE Coef	T	P
Constant	55.125	7.119	7.74	0.000
X_2	-1.375	1.170	-1.17	0.285

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	15.13	15.13	1.38	0.285
Error	6	65.75	10.96		
Total	7	80.88			

Some Illustrations…

Regress *Y* on *X₁* and *X₂*

The regression equation is y = 57.0 - 0.63 X_1 - 1.38 X_2

Predictor	Coef	SE Coef	T	P
Constant	57.000	8.486	6.72	0.001
X_1	-0.625	1.251	-0.50	0.639
X_2	-1.375	1.251	-1.10	0.322

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	18.25	9.13	0.73	0.528
Error	5	62.63	12.53		
Total	7	80.88			

Source	DF	Seq SS
X_1	1	3.13
X_2	1	15.13

Some Illustrations…

Regress Y on X_2 and X_1

The regression equation is $y = 57.0 - 1.38\ X_2 - 0.63\ X_1$

Predictor	Coef	SE Coef	T	P
Constant	57.000	8.486	6.72	0.001
X_2	-1.375	1.251	-1.10	0.322
X_1	-0.625	1.251	-0.50	0.639

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	18.25	9.13	0.73	0.528
Error	5	62.63	12.53		
Total	7	80.88			

Source	DF	Seq SS
X_2	1	15.13
X_1	1	3.13

Some Illustrations…

Variables in model	b_1	b_2
X_1	-0.625	----
X_2	----	-1.375
X_1, X_2	-0.625	-1.375

$SSR(X_1) = 3.13$

$SSR(X_1|X_2) = \text{SSR(X1,X2)-SSR}(X_2)$

$= 18.26 - 15.13 = 3.13$

$SSR(X_2) = 15.13$

$SSR(X_2|X_1) = \text{SSR}(X_1,X_2)\text{-SSR}(X_1)$

$= 18.26 - 3.13 = 15.13$

Some Illustrations…

If predictors are perfectly uncorrelated, then…

- ❑ You get the same slope estimates.
- ❑ That is, the effect on the response ascribed to a predictor doesn't depend on the other predictors in the model.
- ❑ The sum of squares $SSR(X_1)$ is the same as the sequential sum of squares $SSR(X_1|X_2)$.
- ❑ The sum of squares $SSR(X_2)$ is the same as the sequential sum of squares $SSR(X_2|X_1)$.
- ❑ That is, the marginal contribution of one predictor variable in reducing the error sum of squares doesn't depend on the other predictors in the model.

Some Illustrations…

Hald(1952) presents data concerning the heat evolved in calories per gram of cement (Y) as a function of the amount of each of four ingredients in the mixture: tricalcium aluminate(X_1), tricalcium silicate (X_2), tetracalcium alumino ferite(X_3) and dicalcalcium silicate (X_4).

Y	X_1	X_2	X_3	X_4
78.5	7	26	6	60
74.3	1	29	15	52
104.3	11	56	8	20
87.6	11	31	8	47
95.9	7	52	6	33
109.2	11	55	9	22
102.7	3	71	17	6
72.5	1	31	22	44
93.1	2	54	18	22
115.9	21	47	4	26
83.8	1	40	23	34
113.3	11	66	9	12
109.4	10	68	8	12

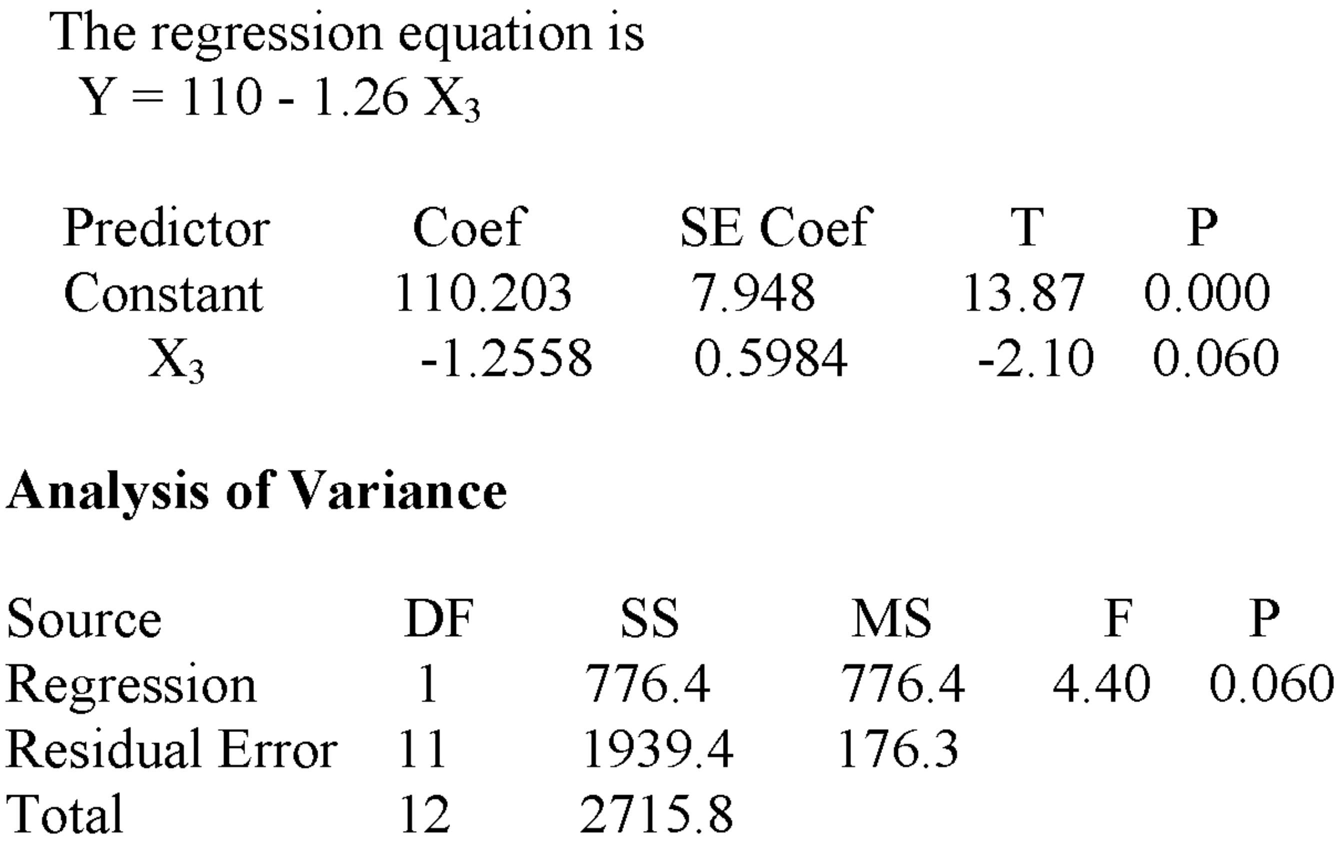
Some Illustrations…

Correlation Matrix

	X_1	X_2	X_3	X_4
X_1	1.00000	0.22858	-0.82413	-0.24545
X_2	0.22858	1.00000	-0.13924	-0.97295
X_3	-0.82413	-0.13924	1.00000	0.02954
X_4	-0.24545	-0.97295	0.02954	1.00000

Some Illustrations…

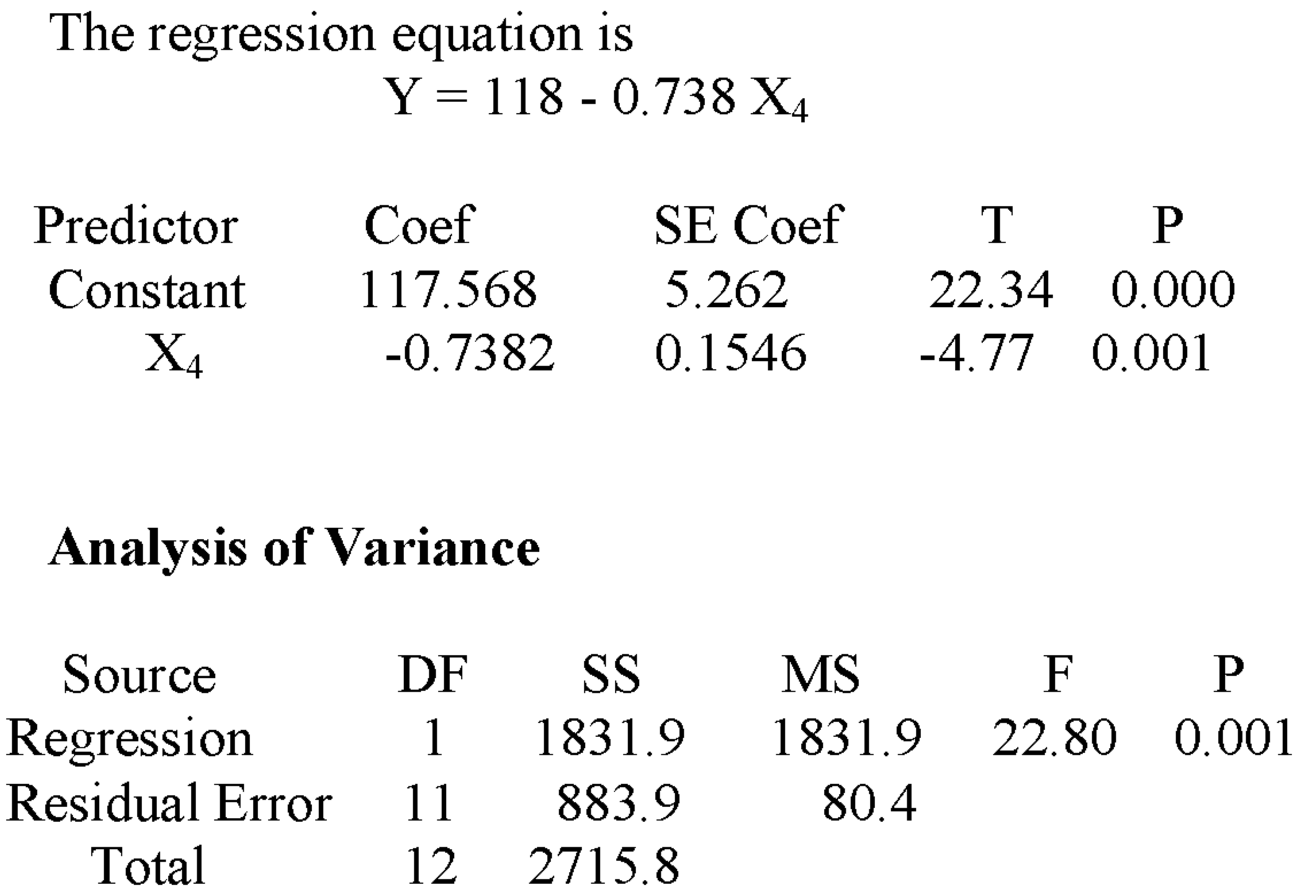
Regress *Y* on X_3



```
The regression equation is
 Y = 110 - 1.26 X3

Predictor       Coef      SE Coef       T       P
Constant     110.203      7.948     13.87   0.000
  X3         -1.2558      0.5984    -2.10   0.060

Analysis of Variance

Source           DF       SS         MS        F       P
Regression        1     776.4      776.4     4.40   0.060
Residual Error   11    1939.4      176.3
Total            12    2715.8
```

Some Illustrations…

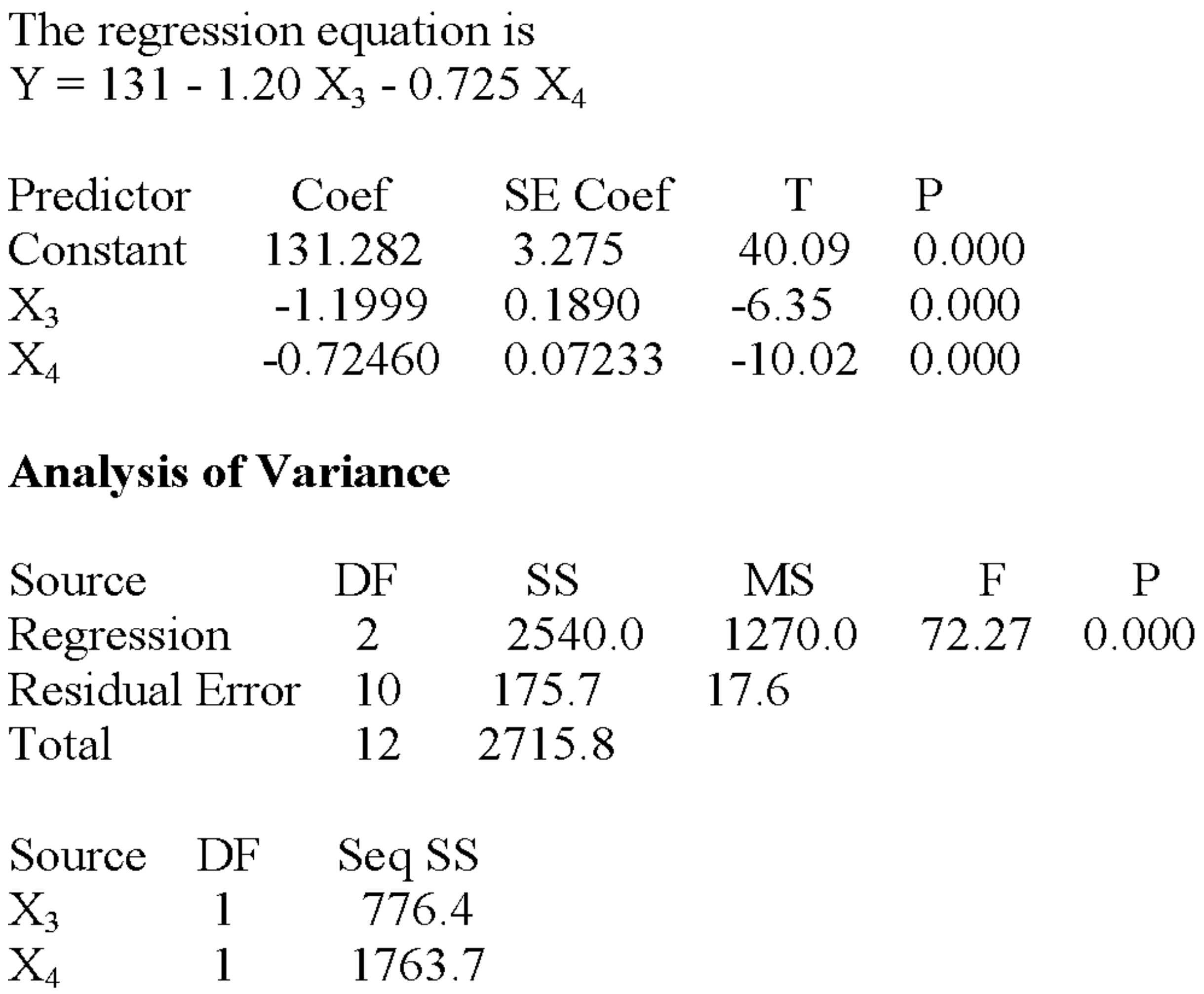
Regress *Y* on X_4



```
The regression equation is
              Y = 118 - 0.738 X4

Predictor       Coef        SE Coef        T        P
 Constant     117.568        5.262       22.34    0.000
      X4       -0.7382      0.1546      -4.77    0.001


 Analysis of Variance

  Source            DF      SS        MS         F        P
Regression           1    1831.9    1831.9     22.80    0.001
Residual Error      11     883.9      80.4
   Total            12    2715.8
```

Some Illustrations…

Regress Y on X_3 *and* X_4



```
The regression equation is
Y = 131 - 1.20 X3 - 0.725 X4

Predictor      Coef       SE Coef        T        P
Constant     131.282       3.275       40.09    0.000
X3           -1.1999      0.1890       -6.35    0.000
X4          -0.72460      0.07233     -10.02    0.000

Analysis of Variance

Source            DF        SS          MS         F        P
Regression         2      2540.0      1270.0     72.27    0.000
Residual Error    10       175.7        17.6
Total             12      2715.8

Source     DF      Seq SS
X3          1       776.4
X4          1      1763.7
```

Some Illustrations…

Variables in model	b_1	b_2
X_3	-1.2558	----
X_4	----	-0.7382
X_3, X_4	-1.1999	-0.7246

$$SSR(X_3) = 776.4$$

$$SSR(X_3|X_4) = \mathrm{SSR}(X_3, X_4)\text{-}\mathrm{SSR}(X_4)$$

$$= 708.1$$

$$SSR(X_4) = 1831.9$$

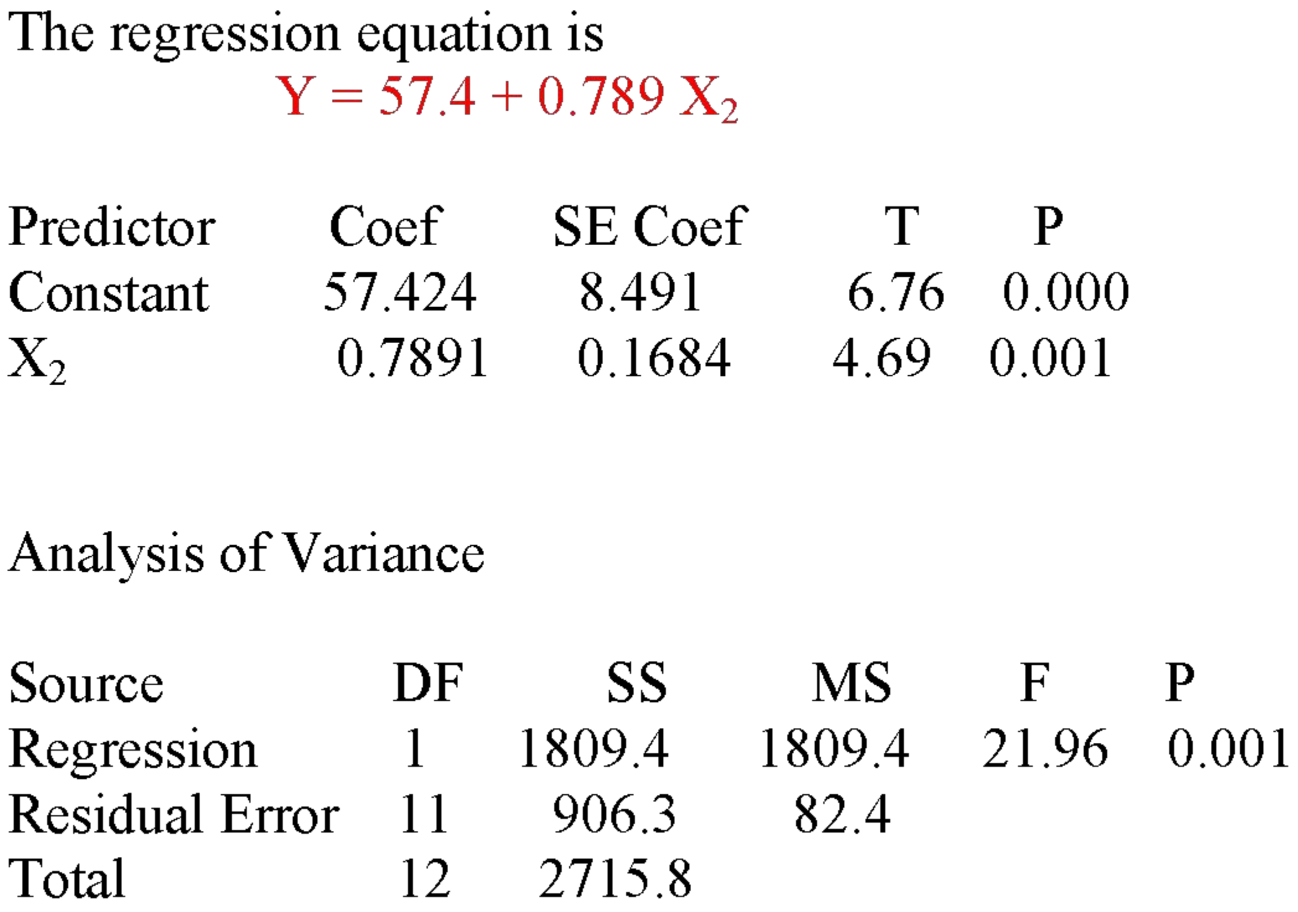
$$SSR(X_4|X_3) = \mathrm{SSR}(X_3, X_4)\text{-}\mathrm{SSR}(X_3)$$

$$= 1763.7$$

Some Illustrations…

What happens if predictors are highly correlated?

Regress *Y* on X_2



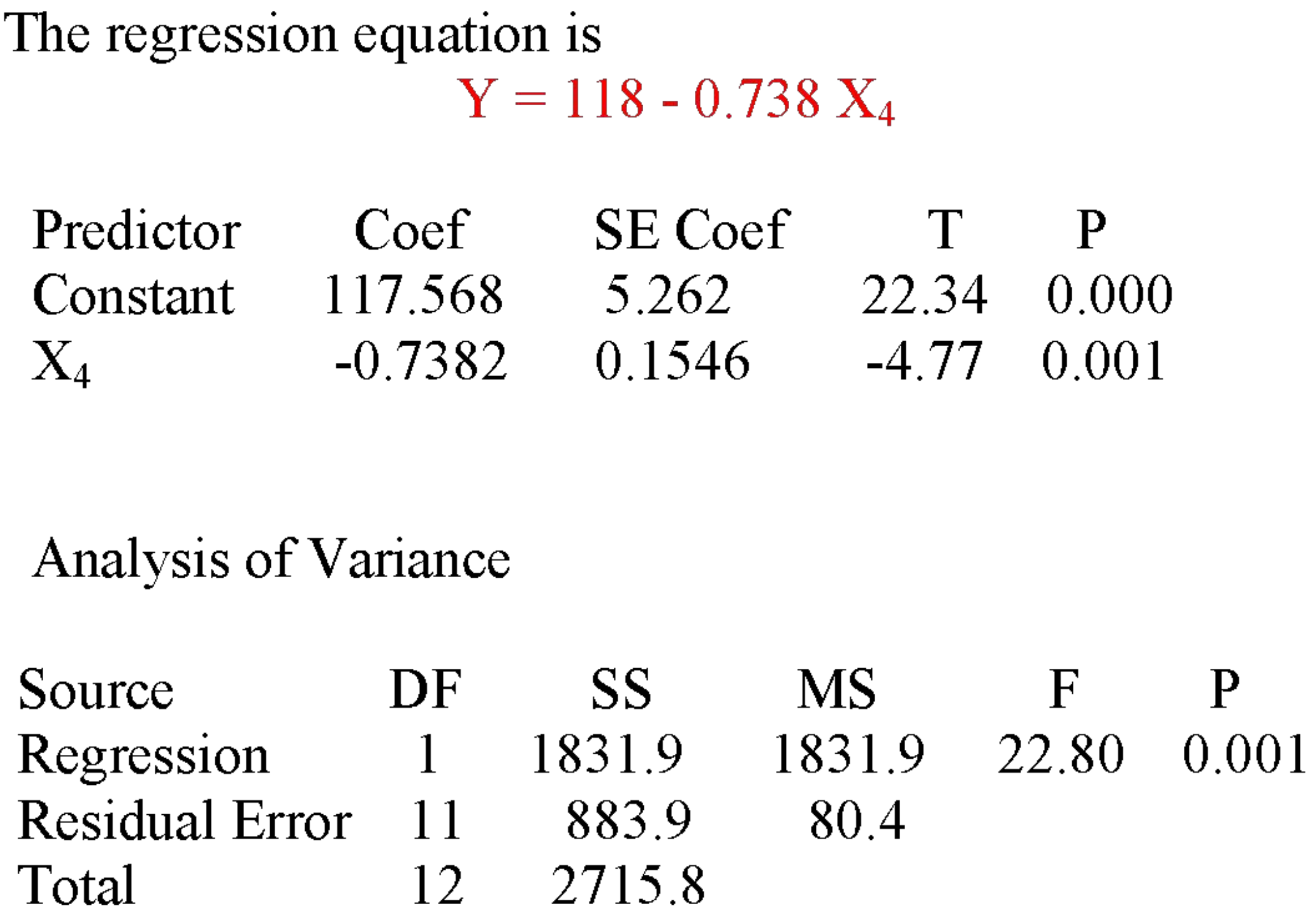
The regression equation is

$$Y = 57.4 + 0.789 X_2$$

```
Predictor      Coef      SE Coef       T       P
Constant      57.424      8.491       6.76    0.000
X2             0.7891     0.1684      4.69    0.001


Analysis of Variance

Source           DF       SS        MS       F       P
Regression        1     1809.4    1809.4   21.96   0.001
Residual Error   11      906.3      82.4
Total            12     2715.8
```

Some Illustrations…

Regress *Y* on X_4



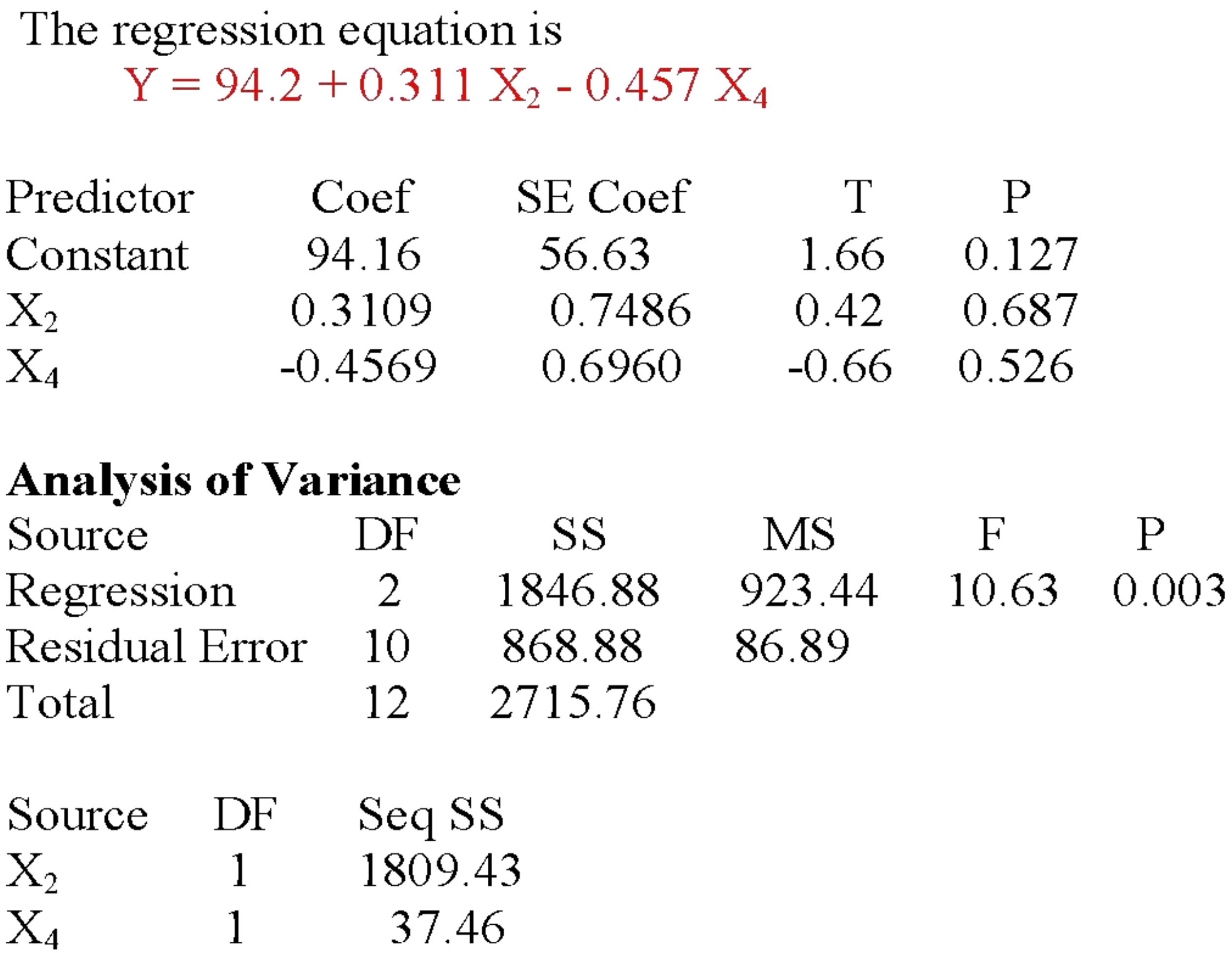
The regression equation is

$$Y = 118 - 0.738\ X_4$$

```
Predictor      Coef       SE Coef       T       P
Constant     117.568       5.262      22.34   0.000
X4           -0.7382      0.1546      -4.77   0.001


Analysis of Variance

Source           DF      SS        MS        F       P
Regression        1    1831.9    1831.9    22.80   0.001
Residual Error   11     883.9      80.4
Total            12    2715.8
```

Some Illustrations…

Regress *Y* on *X₂ and X₄*



The regression equation is

$$Y = 94.2 + 0.311\ X_2 - 0.457\ X_4$$

Predictor	Coef	SE Coef	T	P
Constant	94.16	56.63	1.66	0.127
X_2	0.3109	0.7486	0.42	0.687
X_4	-0.4569	0.6960	-0.66	0.526

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	1846.88	923.44	10.63	0.003
Residual Error	10	868.88	86.89		
Total	12	2715.76			

Source	DF	Seq SS
X_2	1	1809.43
X_4	1	37.46

Some Illustrations…

When predictor variables are correlated, the regression coefficient of any one variable depends on which other predictor variables are included in the model.

Variables in model	b_2	b_4
X_2	0.7891	----
X_4	----	-0.7382
X_2, X_4	0.3109	-0.4569

When predictor variables are correlated, the marginal contribution of any one predictor variable in reducing the error sum of squares varies, depending on which other variables are already in model.

$$SSR(X_2) = 1809.4$$

$$SSR(X_2|X_4) = 14.98$$

$$SSR(X_4) = 1831.9$$

$$SSR(X_4|X_2) = 43.48$$

Multicollinearity

In multiple linear regressions, if there is no linear relationship between the regressors then prediction, estimation and selection of an appropriate set of variables for the model can be made easily. When regressors are nearly or perfectly related then inference about prediction, estimation etc. based on the LS estimator of regression model can be misleading. If regressors are nearly or perfectly related to each other then problem of multicollinearity exist.

Multicollinearity…

In regression model, if two or more regressor variables are highly correlated to each other, or in other words in the X matrix two or more than two columns are linearly dependent to each other then the multicollinearity is said to exist among the regressors.

In terms of the linear dependency, the vectors X_1 , $X_{2,}$…, X_k are linearly dependent if there is a set of constants $t_{1,}$ $t_{2,}$…, t_k not all zero ,such that

$$\sum_{i=1}^{k} t_i X_i = 0 \qquad (1)$$

If above Eq. (1) holds exactly for a subset or approximately for some subset of the columns of X, then rank of $X'X$ is less than k and $(X'X)^{-1}$ does not exist. This implies that, linear dependency is present in $X'X$. Thus the problem of multicollinearity is said to exist.

Multicollinearity…

Consequences:

- The LS estimates become too large in absolute value.
- The LS estimates may have the incorrect algebraic sign.
- It tends to inflate the variance and covariance of the LS estimator.

Multicollinearity…

Methods of Detection of Multicollinearity

- ❑ Simple Correlation Matrix
- ❑ Variance Inflation Factor
- ❑ Condition Number and Condition Index

Multicollinearity…

Simple Correlation Matrix

If X_i and X_j are linearly dependent then , will be near to unity. It is helpful in detecting the near linear dependency between pairs of regressors only.

	X_1	X_2	X_3	X_4
X_1	1.00000	0.22858	-0.82413	-0.24545
X_2	0.22858	1.00000	-0.13924	-0.97295
X_3	-0.82413	-0.13924	1.00000	0.02954
X_4	-0.24545	-0.97295	0.02954	1.00000

Multicollinearity…

Variance Inflation Factor

Variance Inflation Factor is defined as (Marquardt, 1970)

$$VIF_j = \left(1 - R_j^2\right)^{-1}, \qquad j = 1, 2, \dots, k-1,$$

where, R_j^2 is the coefficient of multiple determination from the regression of X_j on the remaining regressor variables.

Variable	VIF
X_1	38.5
X_2	254.4
X_3	46.9
X_4	282.5

Multicollinearity…

Condition Number and Condition Index

The characteristics roots or eigen values are used to measure the degree of multicollinearity among the vectors in the matrix $X'X$. If one or more of the characteristic roots of matrix $X'X$ are small, then it indicates that there are near linear dependencies among the columns of X. Analysts prefer to consider a ratio of the range of these roots. Hence, the **condition number** is defined as

$$k_j = \frac{\lambda_{max}}{\lambda_{min}}, \qquad j = 1, 2, \dots, p - 1,$$

and **Condition Index** of the matrix $X'X$ is defined as

$$k_j = \frac{\lambda_{max}}{\lambda_j}, \qquad j = 1, 2, \dots, p - 1,$$

It is clear that, the largest condition index is the condition number. The value of $k_j > 1000$ indicates the severe multicollinearity in the data.

Multicollinearity…

Methods for dealing with Multicollinearity

In the literature, many techniques are available for dealing with the problem caused by multicollinearity. Some of these are useful in reducing the multicollinearity which occurs among the regressor variables where as some of these are useful in estimation of regression coefficients based on estimators other than LS. The approaches such as, **collecting additional data** and **model respecification** are useful in reducing the multicollinearity. Now, we review the well known **method for estimation other than LS** in the presence of multicollinearity.

Ridge Regression

Hoerl and Kennard (1970) proposed the use of ridge regression for estimation when the regressor variables are highly correlated. Ridge regression is based on the James-Stein estimator and the basic idea is to reduce the variance of estimation by shrinking the estimator, so that the MSE can be reduced. To achieve that in a ridge regression an additional parameter ‘r’ is added to the LS estimation problem. This approach allows a small bias on the estimates to obtain the parameter value. There are two types of ridge regression.

- ❑ Ordinary ridge regression
- ❑ Generalized ridge regression.

Ridge Regression…

Consider the general form of the linear model given in, then the **ridge regression estimator** of [illegible] is defined by

[illegible] = [illegible] + [illegible] $^{-}$ [illegible] [illegible]

where ‘r’ is the ridge constant or ridge parameter [Hoerl et al., (1975), recommended r = [illegible]^{2} / [illegible] , where [illegible] and [illegible]^{2} are the LS estimator of the [illegible] and [illegible]^{2} respectively]. [illegible] is also called as **ordinary ridge regression estimator** of [illegible].

Ridge Regression…

Canonical Form of Regression Model

Consider the multiple linear regression model as

$$y = \beta_0 1 + X^{*}\beta^{*} + \varepsilon$$

For convenience, it is assumed that the X^{*} variables are standardized so that $X^{*'}X^{*}$ has the form of correlation matrix.

Let $\lambda_1, \lambda_2, \dots, \lambda_{p-1}$ be the eigen values of $X^{*'}X^{*}$ and $T_1, T_2, \dots, T_{p-1}$ be the corresponding eigen vectors. Let $\Lambda = diag(\lambda_1, \lambda_2, \dots, \lambda_{p-1})$ and $T = (T_1, T_2, \dots, T_{p-1})$ such that $X^{*'}X^{*} = T\Lambda Q'$. The regression model can be written in the canonical form as

$$y = \beta_0 1 + Z\alpha + \varepsilon$$

where $Z = X^{*}T$ and $\alpha = T'\beta^{*}$.

Ridge Regression…

Since, ☒☒$^{⍰}$ is standardized, we can use ☒, the mean of response variable to estimate ☒☒$_{0}$. So we need only to consider the estimation of ☒☒ (see Liu, 1993).

The LS estimator of ☒☒ for the model is given by

$$\text{⍰}_{\text{☒☒☒}} = \Lambda^{-1}\text{☒☒}^{\text{⍰}}\text{☒☒}$$

Since ☒☒ = ☒☒$^{⍰}$☒☒$^{⍰}$ and ☒☒$^{⍰}$☒☒ = ☒, where ☒ denotes the identity matrix of order ☒ – 1 × ☒ – 1. The LS estimator of ☒☒$^{⍰}$ is given by

$$\text{⍰}^{\text{⍰}}_{\text{☒☒☒}} = \text{☒☒⍰}_{\text{☒☒☒}}.$$

Remark Note that, because of the relation ☒☒ = ☒☒$^{⍰}$☒☒$^{⍰}$, any estimator ⍰ of ☒☒ has a corresponding ⍰$^{⍰}$ = ☒☒⍰ and MSE(⍰$^{⍰}$) = MSE(⍰) (see Sakallioglu and Kaciranlar, 2008). Hence, it is sufficient to consider only a canonical form.

Ridge Regression…

The ordinary ridge regression (ORR) estimator proposed by Hoerl and Kennard (1970a, b) is defined as

$$\hat{\beta}_{ORR} = (C + kI)^{-1}X'y = (I - kW^{-1})\hat{\beta}_{OLS}$$

where $k > 0$ is a shrinkage parameter, $W = (C + kI)$.

Bias, Covariance and MSE:

$$Bias(\hat{\beta}_{ORR}) = E(\hat{\beta}_{ORR}) - \beta = -kW^{-1}\beta$$

$$Cov(\hat{\beta}_{ORR}) = Cov[(I - kW^{-1})\hat{\beta}_{OLS}] = \sigma^2 (I - kW^{-1})C^{-1}(I - kW^{-1})',$$

where $Cov(\hat{\beta}_{OLS}) = \sigma^2 C^{-1}$.

$$MSE(\hat{\beta}_{ORR}) = tr[Cov(\hat{\beta}_{ORR})] + [Bias(\hat{\beta}_{ORR})]'[Bias(\hat{\beta}_{ORR})]$$

$$= \sigma \sum_{j=1}^{p} \left[\frac{\sigma^2}{\lambda_j}\left(1 - \frac{k}{\lambda_j + k}\right)^2 + \left(\frac{k}{\lambda_j + k}\right)^2 \alpha_j^2\right]$$

where $\lambda_j + k$ is the jth diagonal element of matrix $W = \Lambda + kI$

Ridge Parameter Determination Methods

1 Method suggested by Hoerl, Kennard and Baldwin (1975)

$$r_{HKB} = \frac{k\hat{\sigma}^2}{\hat{\alpha}'\hat{\alpha}}$$

2 Method suggested by Lawless and Wang (1976)

$$r_{LW} = \frac{k\hat{\sigma}^2}{\sum_{i=1}^{k} \lambda_i \hat{\alpha}_i^2}$$

3 Method suggested by Masuo Nomura (1988)

$$r_{HMO} = k\hat{\sigma}^2 \Big/ \sum_{i=1}^{k} \left[\hat{\alpha}_i^2 \Big/ \left\{ 1 + \left(1 + \lambda i \left(\hat{\alpha}_i^2 / \hat{\sigma}^2 \right)^{1/2} \right) \right\} \right]$$

4 Method suggested by Khalaf and Shukur (2005)

$$r_{KS} = (\lambda_{\max} \hat{\sigma}^2) / ((n-k-1)\hat{\sigma}^2 + \lambda_{\max} \hat{\alpha}^2{}_{\max})$$