MAT 3103: Statistics and Probability **Semester: Summer 2018**

Chapter 10: Regression Analysis

Dr Atikur R. Khan (Co-ordinator), Atika Farzana Urmi, Md. Mortuza Ahmmed, Ashfia Tasnim Munia

10.1 Regression Model

Linear regression analysis measures average relationship between dependent and explanatory variables, with an objective to learn how explanatory variables affect the dependent variable. When only one explanatory variable is used to measure this relationship, then it is known as simple linear regression. If more than one explanatory variables are used to measure this relationship, then it is known as multiple linear regression.

- Dependent variable: For example, in an experiment, height (Y) of a baby depends on baby's age (X). So, Y is a dependent variable, because it depends on X. Also, known as response variable, because we observe values of Y in response to values of X.
- Explanatory variable: It explains the dependent variable. For example, age of a baby (X) explains the height of a baby (Y). It is also known as predictor variable, because it can be used to predict the dependent variable. For example, when you go to the shopping centre to buy clothes, you essentially predict size by the age of a baby.

10.2 Simple Linear Regression Model

A simple linear regression model consists of a single explanatory (or predictor) variable X to explain (or predict) the dependent (or response) variable Y. This model is written as

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

where β_0 (known as intercept) and β_1 (known as regression coefficient) are parameters, and ϵ is an error term.

Let us assume that we have data from n subjects (say, for example, n babies) as $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$, then the above regression model can be written as

[1]Prepared by Dr. Atikur R. Khan (atikur@aiub.edu).

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_n + \epsilon_n
\end{aligned}
$$

In matrix and vector form, we may write

$$
\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \tag{10.1}
$$

$$
\Rightarrow \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{10.2}
$$

Here $\boldsymbol{\beta}$ is the vector of parameters and we want to estimate this parameter vector by minimizing the sum of squared error. Since $\boldsymbol{\epsilon}$ is the vector of errors in the model, we may write

$$
\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}
$$

and

$$
\boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}
$$

where $\boldsymbol{\epsilon}'\boldsymbol{\epsilon}$ is essentially the sum of squared error in the model. The least square method, minimizes this sum of squared error with respect to $\boldsymbol{\beta}$ to estimate parameters. We find that

$$
\frac{\delta\left(\boldsymbol{\epsilon}'\boldsymbol{\epsilon}\right)}{\delta\boldsymbol{\beta}} = 0
$$

provides

$$
-2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = 0 \Rightarrow \mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}
$$

which is known as the normal equation and by solving this equation we get

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

Least squares estimator is obtained by minimizing the sum of squares error, $\boldsymbol{\epsilon}'\boldsymbol{\epsilon}$. Thus we must make sure that the minimization is happened when we estimate $\boldsymbol{\beta}$. Let us check that

$$
\frac{\delta^2\left(\boldsymbol{\epsilon}'\boldsymbol{\epsilon}\right)}{\delta\boldsymbol{\beta}\delta\boldsymbol{\beta}'} = 2\mathbf{X}'\mathbf{X}
$$

which is a positive definite matrix. Thus $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is obtained by minimizing the sum of squares error (SSE) and is known as a least squares estimator (LSE) of $\boldsymbol{\beta}$.

10.2.1 Mean and Variance of Estimator

We may write that

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\right) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}$$

Thus the mean of $\hat{\boldsymbol{\beta}}$ is

$$E(\hat{\boldsymbol{\beta}}) = E\left(\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\right) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\epsilon}) = \boldsymbol{\beta}$$

under the assumption that

(i) $E(\boldsymbol{\epsilon}) = \mathbf{0}$

(ii) $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ with independently and identically distributed $\epsilon_i \sim N(0, \sigma^2)$

(iii) $\mathbf{X}$ is the matrix of known fixed values of uncorrelated variable (or variables), that is, $\mathbf{X}$ has rank equal to number of its columns.

$$\begin{aligned}
V(\hat{\boldsymbol{\beta}}) &= E\left[\left(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right)\left(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right)'\right] \\
&= E\left[\left(\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} - \boldsymbol{\beta}\right)\left(\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} - \boldsymbol{\beta}\right)'\right] \\
&= E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right] \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

where σ^2 is unknown parameter and we also estimate this parameter.

10.2.2 Fitted Line and Residuals

The fitted regression line is written as

$$\begin{aligned}
&\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \\
\Rightarrow &\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_i \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \\
&\Rightarrow \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad \text{which is the fitted line}
\end{aligned}$$

Since $\mathbf{y}$ is the vector of original observations and $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ is the vector of estimated observations, we compute

vector of residuals

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

$$\Rightarrow \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$$

$$\Rightarrow e_i = y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad \text{which is the estimated error (residual)}$$

Now, the error variance σ^2 can be estimated by using these residuals

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-k} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-k} = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n-k}$$

where k is the number of X variable included in the model. For simple linear regression model, number of X variable is one.

Thus the estimated variance of the estimator is

$$\widehat{V}(\boldsymbol{\beta}) = \hat{\sigma}^2(\mathbf{XX})^{-1}$$

where diagonal elements are variances of estimators. Thus $V(\hat{\beta}_0)$ is the first diagonal element of $\hat{\sigma}^2(\mathbf{XX})^{-1}$ and $V(\hat{\beta}_1)$ is the second diagonal element of $\hat{\sigma}^2(\mathbf{XX})^{-1}$.

10.2.3 Testing of Significance

We are interested in testing whether the X variable(s) are significant or not. This leads to testing the null hypothesis against an alternative hypothesis of the form

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

Test statistic is

$$t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}$$

where $se(\hat{\beta}_j)$ is obtained from the diaginal elements of $\widehat{V}(\boldsymbol{\beta}) = \hat{\sigma}^2(\mathbf{XX})^{-1}$, and k is the number of X variable in the model.

10.2.4 Data Analysis

We want to fit a linear regression model of the form

$$Price_i = \beta_0 + \beta_1 Size_i + \epsilon_i$$

Thus, we may write in the form that

Table 10.1: Price for size of residential flats

Flat Number	Price (in million taka)	Size (in thousand square foot)
1	12.5	2.5
2	7.5	1.5
3	11.0	2.4
4	8.5	2.0
5	11.5	2.2
6	12.0	2.5
7	6.5	1.2

$$\begin{pmatrix} 12.5 \\ 7.5 \\ 11.0 \\ 8.5 \\ 11.5 \\ 12.0 \\ 6.5 \end{pmatrix} = \begin{pmatrix} 1 & 2.5 \\ 1 & 1.5 \\ 1 & 2.4 \\ 1 & 2.0 \\ 1 & 2.2 \\ 1 & 2.5 \\ 1 & 1.2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$\Rightarrow \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Thus $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ where $\mathbf{y} = \begin{pmatrix} 12.5 \\ 7.5 \\ 11.0 \\ 8.5 \\ 11.5 \\ 12.0 \\ 6.5 \end{pmatrix}$, $\mathbf{X} = \begin{pmatrix} 1 & 2.5 \\ 1 & 1.5 \\ 1 & 2.4 \\ 1 & 2.0 \\ 1 & 2.2 \\ 1 & 2.5 \\ 1 & 1.2 \end{pmatrix}$ and $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$.

```
R Codes:
y = c(12.5, 7.5, 11.0, 8.5, 11.5, 12.0, 6.5) # for price
x1 =rep(1, times = length(y)) # for intercept term
x2 = c(2.5,1.5,2.4,2.0,2.2,2.5,1.2) # for size
X = cbind(x1,x2) # forms the X matrix
XX = t(X)%*%X # computes X'X
XX  # produces following output
             x1    x2
         x1  7.0 14.30
         x2 14.3 30.79
IXX = solve(XX) # produces inverse of matrix X'X
IXX # shows following output
                  x1        x2
           x1  2.788949 -1.295290
           x2 -1.295290  0.634058
betavector = IXX%*%t(X)%*%y # estimates beta vector
betavector # shows following result
        [,1]
           x1 0.8337862
           x2 4.4519928
betavector[1] # shows the first coefficient beta_0
```

```
[1] 0.8337862

betavector[2] # shows the second coefficient
[1] 4.451993

resid = y - X%*%betavector # vecror of residuals
n = length(y) # number of observation
k = 1 # number of variables
sigma2 = sum(resid^2)/(n-k) # estimate of noise variance
sigma2 # shows the result
[1] 0.4924894
VarBeta = sigma2*IXX
VarBeta # shows variance matrix as below
           x1         x2
    x1  1.3735280 -0.6379166
    x2 -0.6379166  0.3122669
varbeta_0 = VarBeta[1,1] # variance of beta_0
varbeta_1 = VarBeta[2,2] # variance of beta_1
tstat = betavector[2]/sqrt(varbeta_1) # t-statistic for testing H_0: beta_1 = 0
tstat
[1] 7.966939
# now compare this test statistic with critical value, hats off!!!
```

10.3 Multiple Linear Regression Model

Let us now consider that there are more than one (say, k) independent (or explanatory) variables that can be used to predict the dependent variable Y. Thus the regression model can be written as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \epsilon_i, \text{ where } \epsilon_i \sim N(0, \sigma^2)$$

Thus for n observations, we may write

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \ldots \beta_k x_{k1} \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{12} + \ldots \beta_k x_{k2} \epsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{1n} + \ldots \beta_k x_{kn} \epsilon_n \end{aligned}$$

In matrix and vector form, we may write

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{k1} \\ 1 & x_{12} & \dots & x_{k2} \\ \vdots & & & \\ 1 & x_{1n} & \dots & x_{kn} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$
$$\Rightarrow \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

which is similar to the equation in (10.1). Thus the overall estimating equations are similar to those shown for simple linear regression model.

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ E(\hat{\boldsymbol{\beta}}) &= \boldsymbol{\beta} \\ V(\hat{\boldsymbol{\beta}}) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \\ \hat{\mathbf{y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ \mathbf{e} &= \mathbf{y} - \hat{\mathbf{y}} \\ \hat{\sigma}^2 &= \frac{\sum_{i=1}^{n} e_i^2}{n-k} = \frac{\mathbf{e}'\mathbf{e}}{n-k} \\ \hat{V}(\hat{\boldsymbol{\beta}}) &= \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

Thus the test statistic to test the hypothesis

$$H_0 : [\boldsymbol{\beta}]_i = 0 \ \ vs. \ \ H_0 : [\boldsymbol{\beta}]_i \neq 0$$

is

$$t = \frac{[\hat{\boldsymbol{\beta}}]_i}{sd([\hat{\boldsymbol{\beta}}]_i)} = \frac{[\hat{\boldsymbol{\beta}}]_i}{\sqrt{[\hat{V}(\hat{\boldsymbol{\beta}})]_{ii}}} \sim t_{n-k}$$

where $[\hat{\boldsymbol{\beta}}]_i$ is the ith element of $\hat{\boldsymbol{\beta}}$ and $[\hat{V}(\hat{\boldsymbol{\beta}})]_{ii}$ is the (i,i)th element of $\hat{V}(\hat{\boldsymbol{\beta}})$.

10.3.1 Data Analysis

Table 10.2: Price for size of residential flats and green rating

Flat Number	Price (in million taka)	Size (in thousand square foot)	Rating (green rating)
1	12.5	2.5	0.95
2	7.5	1.5	0.65
3	11.0	2.4	0.75
4	8.5	2.0	0.70
5	11.5	2.2	0.80
6	12.0	2.5	0.85
7	6.5	1.2	0.5

`R Codes:`

```
y = c(12.5, 7.5, 11.0, 8.5, 11.5, 12.0, 6.5) # for price
x1 =rep(1, times = length(y)) # for intercept term
x2 = c(2.5,1.5,2.4,2.0,2.2,2.5,1.2) # for size
x3 = c(0.95, 0.65, 0.75, 0.70, 0.80, 0.85, 0.50)
X = cbind(x1,x2,x3) # forms the X matrix
XX = t(X)%*%X # computes X'X
IXX = solve(XX) # produces inverse of matrix X'X
betavector = IXX%*%t(X)%*%y # estimates beta vector
betavector[1] # shows the first coefficient beta_0
betavector[2] # shows the second coefficient beta_1
betavector[3] # shows the second coefficient beta_2

resid = y - X%*%betavector # vecror of residuals
n = length(y) # number of observation
k = 2 # number of variables
sigma2 = sum(resid^2)/(n-k) # estimate of noise variance
VarBeta = sigma2*IXX
varbeta_0 = VarBeta[1,1] # variance of beta_0
varbeta_1 = VarBeta[2,2] # variance of beta_1
varbeta_2 = VarBeta[3,3] # variance of beta_2
tstat1 = betavector[2]/sqrt(VarBeta[2,2]) # t-statistic for testing H_0: beta_1 = 0
tstat2 = betavector[3]/sqrt(VarBeta[3,3]) # t-statistic for testing H_0: beta_2 = 0
# now compare this test statistic with critical value, hats off!!!
```

10.4 Exercises

1. For simple linear regression model verify whether following results are correct

   (i) $\mathbf{X'X} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{bmatrix}$, $(\mathbf{X'X})^{-1} = \frac{1}{n\sum x_i^2 - n^2\bar{x}^2} \begin{bmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} = \frac{1}{nSS_x} \begin{bmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix}$
   and $\mathbf{X'y} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} = \begin{bmatrix} n\bar{y} \\ \sum x_i y_i \end{bmatrix}$ where $SS_x = \sum x_i^2 - n\bar{x}^2$

   (ii) $(\mathbf{X'X})^{-1}\mathbf{X'y} = \frac{1}{nSS_x} \begin{bmatrix} \sum x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix} \begin{bmatrix} n\bar{y} \\ \sum x_i y_i \end{bmatrix} = \frac{1}{nSS_x} \begin{bmatrix} \sum x_i^2 (n\bar{y}) - n\bar{x}\sum x_i y_i \\ -n^2\bar{x}\bar{y} + n\sum x_i y_i \end{bmatrix}$.

   $\hat{\boldsymbol{\beta}} = \frac{1}{nSS_x} \begin{bmatrix} n\bar{y}SS_x + n\bar{x}^2 n\bar{y} - n\bar{x}\sum x_i y_i \\ nSP_{xy} \end{bmatrix} = \frac{1}{nSS_x} \begin{bmatrix} n\bar{y}SS_x - n\bar{x}SP_{xy} \\ nSP_{xy} \end{bmatrix} = \begin{bmatrix} \bar{y} - \frac{SP_{xy}}{SS_x}\bar{x} \\ \frac{SP_{xy}}{SS_x} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$

2. For simple linear regression model derive test statistic for testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$.

3. Assume that your company's profit in recent 5 years compared to that of year 2010 are 1.1, 1.4, 1.3, 1.6, 1.8. Fit a linear trend model and comment on the results. (Hints: put Y values y = c(1.1, 1.4, 1.3, 1.6, 1.8) and X values are time trend t = c(1,2,3,4,5) to fit a simple linear regression model).