

LECTURE 6 Linear Modelling

Objectives:

- **Unbiasedness** of least square estimators **A** and **B**
- Variance of **A** and **B** and Distribution of **RMS**
- Evaluation of a **model**.
- **Residuals** and **properties** of Regression Models.

STATISTICAL PROPERTIES OF LEAST SQUARE ESTIMATORS (REGRESSION COEFFICIENTS A AND B)

MRS SAHARAN WILL TEACH

1) A and B are the unbiased estimators of α and β

- The **fitted line** created by least square technique using the **sample** is $Y = A + BX + E$ This is estimation for the true **population** regression line $Y = \alpha + \beta X + \varepsilon$
- There are **two steps** to accomplish this proof, first prove that these **estimators** are **linear** estimators (Linear combinations) and **second** prove that they are **unbiased**.

STEP 1 A and B are Linear Estimators

$$B = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2} = \sum K_i Y_i$$

$$\text{For } K_i = \frac{X_i - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

Therefore **B** is a **linear combination** of Y_i s.

- Similarly for B

$$A = \bar{Y} - B\bar{X} = \frac{1}{n} \sum_{i=1}^n Y_i - B\bar{X} \quad \text{A is also a linear combinations of } Y_i\text{s}$$

Because the first term consisting of Y_i is a linear combination of Y_i s and B is a linear combinations Y_i s. Therefore A and B are Linear estimators.

STEP 2

- To prove that the **estimators** are **unbiased**:

To prove $E(A) = \alpha$ and $E(B) = \beta$ where

$$Y_i = \alpha + \beta X_i + \varepsilon_i \quad \dots\dots\dots 1$$

Dividing both sides by n and taking a summation from 1 to n:

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i = \alpha + \beta \bar{X} + \bar{\varepsilon} \quad \dots\dots\dots 2$$

Where $\bar{Y} = \frac{1}{n} \sum Y_i$ $\bar{X} = \frac{1}{n} \sum X_i$ $\bar{\varepsilon} = \frac{1}{n} \sum_{i=1}^n \varepsilon_i$

Subtracting the second equation from the first (1-2) results:

$$Y_i - \bar{Y} = \beta(X_i - \bar{X}) + \varepsilon_i - \bar{\varepsilon} \quad \dots\dots\dots 3$$

Taking expectation on both sides:

$$E(Y_i - \bar{Y}) = \beta(X_i - \bar{X}) + E(\varepsilon_i - \bar{\varepsilon}) \quad \dots\dots\dots 4$$

Here **Y and ε are random variables** but **X is not a Random variable as it is a controlled** variable. Therefore $(X_i - \bar{X})$ is **just a constant**. Remember we had assumed $\varepsilon_i \sim N(0, \sigma^2)$ so the expected value of ephsilons $(\varepsilon, \bar{\varepsilon}) = 0$ and also $E(\varepsilon_i - \bar{\varepsilon}) = 0$

Therefore 4 results:

$$E(Y_i - \bar{Y}) = \beta(X_i - \bar{X}) \quad \dots\dots\dots 5$$

We had already **proved** the following **formula for B**:

$$B = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad \dots\dots\dots 6$$

$$E(B) = \frac{\sum_{i=1}^n (X_i - \bar{X}) * E(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \beta \quad \dots\dots 7$$

Taking expectation of 6 and the result of 5:

$$E(B) = \frac{\sum_{i=1}^n (X_i - \bar{X})\beta * E(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \beta$$

$$E(B) = \beta$$

Next to prove $E(A) = \alpha$

We have previously proved that :

$$A = \bar{Y} - B\bar{X}$$

$$E(A) = E(\bar{Y} - B\bar{X}) \dots\dots\dots 1$$

$$\text{We know that } \bar{Y} = \alpha + \beta\bar{X}$$

$$E(A) = E(\alpha + \beta\bar{X} - B\bar{X}) \dots\dots\dots 2$$

Using the **previous proof** : $E(B) = \beta$

$$E(A) = \alpha + \beta\bar{X} - \beta\bar{X}$$

$$E(A) = \alpha$$

VARIANCE OF A AND B **GS1 WILL TEACH**

Variance of B

$$V(B) = V\left(\frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) = V\sum_{i=1}^n K_i Y_i \dots\dots\dots 1$$

Note: This K is different from previously mentioned.

Where $K_i = \left(\frac{X_i - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) \dots\dots\dots 2$

Since Y_i are **independent** therefore Variance can be computed as follows:

$$V(B) = V\left(\sum_{i=1}^n K_i Y_i\right) = \sum_{i=1}^n K_i^2 V(Y_i) \dots\dots\dots 3$$

$$V(B) = \sum_{i=1}^n K_i^2 \sigma^2 \dots\dots\dots 4$$

Since **K is a function X_i** therefore X_i s are fixed quantities because k are fixed.

Therefore using $\sum K_i^2 = \frac{1}{\sum (X_i - \bar{X})^2}$ since

Proof of this Property

$$K_i = \left(\frac{X_i - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) \text{ therefore } \sum K_i^2 = \sum \left[\frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right]^2 = \frac{1}{\sum (X_i - \bar{X})^2}$$

Using this property of k 4 becomes:

$$Var(B) = \frac{\sigma^2}{\sum (X_i - \bar{X})^2} = \frac{\sigma^2}{s_{xx}} \dots\dots\dots 5$$

Now we know $Y = \alpha + \beta X + \varepsilon$ where $Y \sim N(\alpha + \beta X, \sigma^2)$ and $\varepsilon \sim N(0, \sigma^2)$

$$Var(Y) = Var(Y - \alpha - \beta X) = Var(\varepsilon) \text{ since } \alpha + \beta X \text{ is a constant}$$

Therefore we can **replace** σ by σ_ε

$$Var(B) = \frac{\sigma_\varepsilon^2}{s_{xx}} \dots\dots\dots \text{Page 109}$$

VARIANCE OF A

$$V(A) = V(\bar{Y} - B\bar{X}) = V(\bar{Y}) + V(B\bar{X}) - 2\bar{X}Cov(\bar{Y}, B) \dots\dots\dots 1$$

There are three terms on the RHS

First term and second term by using the variance rule:

$$V(\bar{Y}) = V\left(\frac{1}{n} \sum_{i=1}^n Y_i\right) = \frac{1}{n^2} \sum_{i=1}^n V(Y_i) = \frac{\sum_{i=1}^n \sigma^2}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \dots\dots\dots 2$$

$$V(B\bar{X}) = \bar{X}^2 V(B) = \frac{\bar{X}^2 \sigma^2}{s_{xx}} \quad \text{Xbar is a constant and we calculated V(B)} \dots\dots\dots 3$$

Third term is as follows:

$$2\bar{X}Cov(\bar{Y}, B) = 2\bar{X}Cov\left(\frac{\sum_{i=1}^n Y_i}{n}, \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) \dots\dots\dots 4$$

$$2\bar{X}Cov(\bar{Y}, B) = \frac{2\bar{X} \sum \sum (X_i - \bar{X})Cov(Y_i, Y_i)}{n(\sum_{i=1}^n (X_i - \bar{X})^2)} \dots\dots\dots 5$$

Y_i are independent there fore $Cov(Y_i, Y_j) = 0$ for $i \neq j$

$$2\bar{X}Cov(\bar{Y}, B) = 0 \dots\dots\dots 6$$

Therefore using the addition of the three terms in equation 1,3 and 6

$$V(A) = \frac{\sigma^2}{n} + \frac{\bar{X}^2 \sigma^2}{s_{xx}} + 0 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{s_{xx}} \right) \dots\dots\dots 7$$

$$V(A) = \sigma^2 \left(\frac{1}{n} + \frac{\sum X_i^2}{\sum (X_i - \bar{X})^2} \right) \dots\dots\dots 8$$

$$V(A) = \sigma^2 \left(\frac{\sum X_i^2 + \sum \bar{X}_i^2 - 2 \sum X_i \bar{X} + n \sum X_i^2}{n \sum (X_i - \bar{X})^2} \right) \dots\dots\dots 9$$

Changing all Numerators to X_i s:

$$V(A) = \sigma^2 \left(\frac{\sum X_i^2 + \sum n X_i^2 - 2 \sum X_i n X_i + n \sum X_i^2}{n \sum (X_i - \bar{X})^2} \right) \dots\dots\dots 10$$

Finally replacing σ by σ_ε (same logic as given for B) we obtain

$$V(A) = \frac{\sigma_\varepsilon^2 \sum X_i^2}{n \sum (X - \bar{X})^2}$$

ESTIMATION OF σ or σ_ε **GSI WILL TEACH**

Since we do not know the value of σ^2 we have to estimate it. In many books Sum of squared residuals is denoted by RSS instead of $S(A, B)$ or $\sum_{i=1}^n E_i^2$

We will prove that the unbiased estimator of σ^2 is $\frac{RSS}{n-2}$

$$\text{ie } E\left(\frac{RSS}{n-2}\right) = \sigma^2$$

Proof:

Since we know $\hat{Y} = A + B X_i$

$$RSS = \sum_{i=1}^n E_i^2 = \sum_{i=1}^n (Y_i - \hat{Y})^2 = \sum_{i=1}^n (Y_i - A - B X_i)^2 \dots\dots\dots 1$$

We know from our previous proof: $A = \bar{Y} - B \bar{X}$

Substituting for A and squaring with the first two values as the first term and second third value as the second term of the square.

$$RSS = \sum_{i=1}^n (Y_i - \bar{Y} + B\bar{X} - BX_i)^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 + B^2 \sum_{i=1}^n (\bar{X} - X_i)^2 - 2 \sum_{i=1}^n B(\bar{X} - X_i)(Y_i - \bar{Y})$$

.....2

The equation 2 can be written as

$$RSS = s_{yy} + B^2 s_{xx} - 2Bs_{xy} \text{3}$$

We know by the equation of B

$$B = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \text{4}$$

4 can be written as follows:

$$B = \frac{s_{xy}}{s_{xx}} \Rightarrow s_{xy} = Bs_{xx} \text{5}$$

Substituting 5 into 3 we obtain the following result:

$$RSS = s_{yy} + B^2 s_{xx} - 2B^2 s_{xx} \text{6}$$

$$RSS = s_{yy} - B^2 s_{xx} \text{7}$$

Taking expectation of both sides:

$$E(RSS) = E(s_{yy}) - E[B^2(s_{xx})] \text{8}$$

Now we will compute $E(s_{yy})$ and $E[B^2(s_{xx})]$

$$E(s_{yy}) = E \sum (Y_i - \bar{Y})^2 \text{9}$$

$$E(s_{yy}) = E(\sum Y_i^2 + \sum \bar{Y}^2 - 2 \sum Y_i \bar{Y}) \text{10}$$

$$E(s_{yy}) = E \sum Y_i^2 + E(n\bar{Y}^2 - 2n\bar{Y}\bar{Y}) \text{11}$$

$$E(s_{yy}) = E \sum Y_i^2 - E(n\bar{Y}^2) \dots\dots\dots 12$$

$$E(s_{yy}) = E \left[\sum_{i=1}^n Y_i^2 \right] - nE[\bar{Y}^2] \dots\dots\dots 13$$

$$Y_i = A + BX_i + E_i \dots\dots\dots 14$$

Finding the expectation of both sides:

$$E(Y_i) = A + BX_i \quad \text{Since } E(E_i) = 0$$

$$V(Y_i) = \sigma^2 \quad \text{Since } V(\varepsilon_i) = \sigma^2 \quad \text{as } \varepsilon \text{ and } Y \text{ are random and } X \text{ is non random.}$$

By formulas for variance :

$$E[Y_i^2] = V(Y_i) + [E(Y_i)]^2 = \sigma^2 + (A + BX_i)^2 \dots\dots\dots 15$$

$$E[\bar{Y}^2] = V(\bar{Y}) + [E(\bar{Y})]^2 = \frac{\sigma^2}{n} + [A + B\bar{X}]^2 \dots\dots\dots 16$$

Therefore substituting the values of 15,16 in 13

$$E(s_{yy}) = \left[\sum_{i=1}^n E[Y_i^2] - nE[\bar{Y}^2] \right] = \sum_{i=1}^n [\sigma^2 + (A + BX_i)^2] - n \frac{\sigma^2}{n} - n[A + B\bar{X}]^2 \dots\dots\dots 16$$

$$E(s_{yy}) = n\sigma^2 + \sum_{i=1}^n (A + BX_i)^2 - \sigma^2 - n(A + B\bar{X})^2 \dots\dots\dots 17$$

$$E(s_{yy}) = (n-1)\sigma^2 + B^2 \left[\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right] \dots\dots\dots 18$$

$$E(s_{yy}) = (n-1)\sigma^2 + B^2 s_{xx} \dots\dots\dots 19$$

Now computing $E[B^2(s_{xx})]$

By the formula for expectation:

$$E[B^2(s_{xx})] = s_{xx} E(B^2) \dots\dots\dots 20$$

$$E(B) = \beta$$

We have already proved that

$$V(B) = \frac{\sigma^2}{s_{xx}}$$

$$\text{Also we know } E(B^2) = V(B) + [E(B)]^2 = \frac{\sigma^2}{s_{xx}} + \beta^2 \dots\dots\dots 21$$

Substituting $E(B^2)$ from 21 into 20

$$E[B^2(s_{xx})] = s_{xx} E(B^2) = s_{xx} \left[\frac{\sigma^2}{s_{xx}} + \beta^2 \right] = \sigma^2 + \beta^2 s_{xx} \dots\dots\dots 23$$

Now substituting these values of $E[B^2(s_{xx})]$ from equation 23 and $E(s_{yy})$ from equation 19 in $E(RSS)$ in equation 8 we obtain:

$$E(RSS) = E(s_{yy}) - E[B^2(s_{xx})] = (n-1)\sigma^2 + \beta^2 s_{xx} - \sigma^2 - \beta^2 s_{xx} = (n-2)\sigma^2$$

$$E\left(\frac{RSS}{n-2}\right) = \sigma^2 \quad \text{We have found the unbiased estimator of } \sigma^2$$

MRS SAHARAN WILL TEACH

Residual Standard Error

If wish to find **how well the line fits the data** we will need the **standard deviation of the residuals** which is called the **standard error of the residuals**. Therefore we will need the **sampling distribution of the residuals**. Since **we do not have the variance of the population** we use the sample variance within the estimate that we just proved for the formula

$$s_E = \text{Residual Standard Error}$$

$$s_E^2 = \left(\frac{RSS}{n-2} \right)$$

If we denote $RSS = \sum_{i=1}^n E_i^2$ the formula becomes

$$s_E^2 = \left(\frac{\sum_{i=1}^n E_i^2}{n-2} \right) \dots\dots\dots \text{Page 87 in text}$$

$$\frac{RSS}{n-2} = \text{RMS} \quad \text{is called } \text{residual mean squared}.$$

DISTRIBUTION OF THE RMS

- Assumptions of the population model can be applied to the fitted model.

We know Sum of Squared Residual = $RSS = \sum_{i=1}^n E_i^2$ where $E_i = Y_i - \hat{Y}_i$

It can be proved that $E(E_i)=0$ $V(E_i)=\sigma^2$ (Like we proved in the population case)

Where $E_i = \hat{\varepsilon}_i$ where E_i is the estimate of the i th population error term.

- By the model assumptions $\varepsilon_i \sim N(0, \sigma^2)$ and ε_i s are independent

Since we know ε_i s are normal and independent

this implies the observations Y_i s are also Normal with the following distribution:

$$Y_i \sim N(A + BX_i, \sigma^2) \text{ where } Y_i = E_i + A + BX_i$$

- $E_i = Y_i - \hat{Y}_i$ $E_i = Y_i - A - BX_i$ is a linear combination of the Y_i s (observations) since all the terms are linear combinations of Y_i s. See our previous proof of how A and B are Linear combinations of Y_i s.
- Since Y_i are normally distributed therefore E_i is also normally distributed as linear combination of normal variables is normal.

$$E_i \sim N(0, \sigma^2)$$

$$\frac{E_i}{\sigma} \sim N(0,1) \text{ Standard Normal}$$

$$\text{Therefore } \frac{E_i^2}{\sigma^2} \sim \chi_1^2$$

- RSS Sum of Squared residual $RSS = \sum_{i=1}^n E_i^2$ is the sum of E_i squared but their distribution

is not chi squared(n) because all the E_i s are not independent.

E are bound by the following constraints. We already know that A and B are Least Square Estimators of α and β .

And $E_i = Y_i - \hat{Y}_i$ satisfies the following two constraints

$$1) E_1 + E_2 + \dots + E_n = 0$$

$$2) E_1 X_1 + E_2 X_2 + \dots + E_n X_n = 0$$

These are the two normal equations

There are n-2 degree of freedom for the residuals. All the E_i s are not independent. The first n-2 E_i s can be chosen independently but the final two have to be chosen such that the two constraints are satisfied. Thus this distribution of SSR follows χ^2_{n-2}

$$\text{Therefore } \frac{SSR}{\sigma^2} = \frac{\sum_{i=1}^n E_i^2}{\sigma^2} \sim \chi^2_{n-2}$$

$$\frac{SSR}{\sigma^2} \sim \chi^2_{n-2} \Rightarrow \frac{(n-2)RMS}{\sigma^2} \sim \chi^2_{n-2} \quad \text{Since } RMS = \frac{SSR}{n-2}$$

This result is used in Testing Hypothesis

EVALUATION OF MODEL

- For a data set we have estimated the regression coefficients. We have fitted a regression model onto the data.
- Now for this fitted Linear model we will evaluate the goodness of the fit.
- We need to test the significance of A and B

1) Testing the significance of Slope Coefficient B : Hypothesis Formulation and Testing

$H_0: \beta=0$ (There is no linear relationship) Null Hypothesis

$H_1: \beta \neq 0$ (There is a linear relationship) Alternative Hypothesis

H_0 This signifies that the explanatory variable does not effect the value of the response variable.

To test this hypothesis the test statistics has to be computed.

$$B = \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2} = \sum K_i Y_i$$

So B is the linear combination of Y_i and we know $Y_i \sim N(\alpha + \beta X, \sigma^2)$

Therefore B is the linear combination of normal variables. The mean of B (unbiased estimator of β) is β and variance is $\frac{\sigma^2}{S_{xx}}$ (as per earlier proof) . Therefore $B \sim N(\beta, \frac{\sigma^2}{S_{xx}})$

This is the sampling distribution for B used to find the critical value.

$$z = \frac{B - \beta}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$$

- **Test statistics:**

$$z = \frac{B}{\sqrt{\frac{\sigma^2}{s_{xx}}}} \quad \text{under Null Hypothesis for which } H_0: \beta=0 \text{ (There is no linear relationship)}$$

If σ^2 is known then z test can be used, Reject H_0 if $|Z| > Z_{\alpha/2}$

Usually σ^2 is not known Then we proved that the unbiased estimator of σ^2 was $SSR/n-2 = \text{RMS}$

Therefore the test statistics now becomes

$$t = \frac{B - \beta}{\sqrt{\frac{\text{RMS}}{s_{xx}}}} \quad \text{This does not follow normal distribution. now derive its distribution}$$

$$\text{We know that } B \sim N\left(\beta, \frac{\sigma^2}{s_{xx}}\right)$$

$$\text{therefore } \frac{B - \beta}{\sqrt{\frac{\sigma^2}{s_{xx}}}} \sim N(0,1)$$

$$\text{We know that } \frac{(n-2)\text{RMS}}{\sigma^2} \sim \chi_{n-2}^2$$

It can be proved that these two distributions are independent.

Standard result in sampling distribution:

If $X \sim N(0,1)$ and $Y \sim \chi_n^2$ and these distributions are independent then:

$$\frac{X}{\sqrt{\frac{Y}{n}}} \sim t_n$$

Using this conclusion

$$\frac{\frac{B - \beta}{\sqrt{\sigma^2}}}{\sqrt{\frac{(n-2)\sigma^2 RMS}{(n-2)}}} \sim t_{n-2}$$

$$\frac{\frac{B - \beta}{\sqrt{RMS}}}{\sqrt{s_{xx}}} \sim t_{n-2}$$

Therefore the test statistics is

$$t = \frac{B - \beta}{\sqrt{\frac{RMS}{s_{xx}}}}$$

$$t = \frac{B}{\sqrt{\frac{RMS}{s_{xx}}}} \text{ under } H_0$$

This is a 2 sided test We will reject H_0 if $|t| > t_{\alpha/2, n-2}$

Example Is the number of hours of work in a student life affecting the number of time spent with family in a day.

X	Y	$X_i Y_i$	X_i^2	\hat{Y}_i	$E_i = Y_i - \hat{Y}_i$	E_i^2
2	3					
3	1					
1	1					
4	1					
2	3					
1	1					

Find the Fitted Equation for Least Square estimation also is there a relationship between the two variables? Alpha=.05

STEP 1 Evaluate A and B

$$B = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^n X_i^2 - n \bar{X}^2} \quad \text{and} \quad A = \bar{Y} - B \bar{X}$$

Then substitute to find the fitted equation $\hat{Y} = A + BX$

Step 2 Find \hat{Y}_i , $E_i = Y_i - \hat{Y}_i$, E_i^2 and then find $SSR = \sum_{i=1}^n E_i^2$ and $RMS = \frac{SSR}{n-2}$

Find the t test $t = \frac{B}{\sqrt{\frac{RMS}{s_{xx}}}}$ where an easier formula for s_{xx} is

$$s_{xx} = \sum_{i=1}^n X_i^2 - n \bar{X}^2$$