LECTURE 7 Linear Modelling

Objectives:

- ANOVA techniques for linear regression and model evaluation
- Confidence Intervals
- Coefficient of Determination
- Correlation Coefficient

ANOVA Technique

The hypothesis testing and accuracy determination of A and B can be conducted by the ANOVA technique as well.

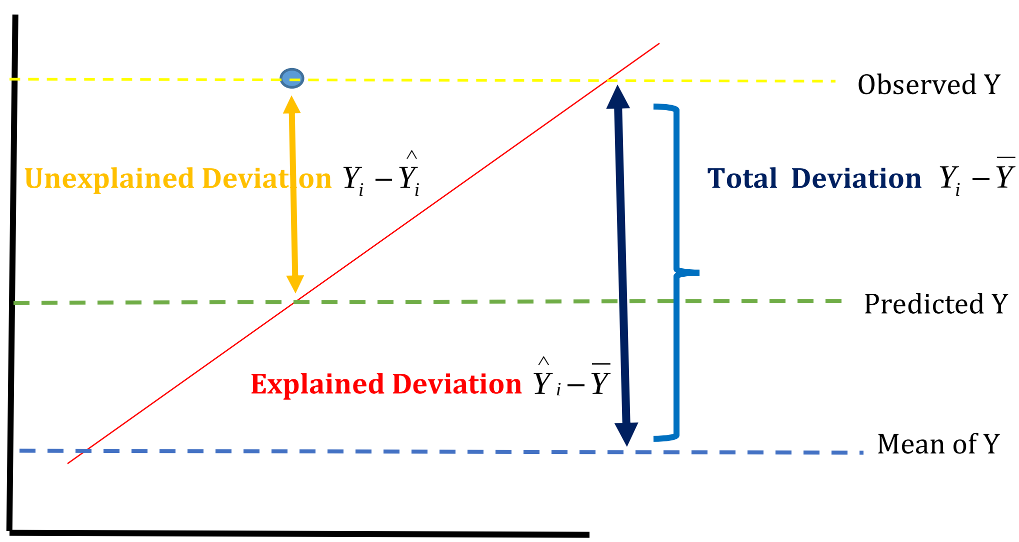
Total variation in the Response variable (Observed data about its mean) = $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ **……….1**

Some of the variation is explained by the explanatory/regressor variable (or the model: regression equation) and some of it left unexplained (residual or error).

For one observation ie for the ith observation :

$(Y_i - \overline{Y}) = (\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i)$ Total Deviation = Explained + Unexplained**………………2**

Taking a just one response variable.



For one observation:

$Y_i - \hat{Y}_i = (Y_i - \overline{Y}) - (\hat{Y}_i - \overline{Y})$ **…………………………………3**

Deviation for the ith observation from its predicted/fitted =

Deviation of the ith observation from its Overall mean --- Deviation of the ith predicted value from

We use the equation **2** to write out the sum of squares of deviation across all the Y_{is} (Observed Data). Equation 2 is for one observation. To obtain the total sum of squares of deviation of the observed data from its mean we apply summation across the entire equation (both sides of the equality) from 1 to n.

$$\sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n}[(\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i)]^2 \qquad \text{.......................................}\mathbf{4}$$

TSS = RegSS + RSS

Sum of squared	Sum of Squared	Sum of squared
total variation	regression	residual
Total Variation =	Explained by +	Unexplained by
	model	model

Here the second term on the right side is the SSR Sum of Square Residuals. This the unexplained part of the variation and has to be minimized. We wish to maximize the part explained by the model.

$$\sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n}[(\hat{Y}_i - \overline{Y})]^2 + \sum_{i=1}^{n}[(Y_i - \hat{Y}_i)]^2 + 2\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})(Y_i - \hat{Y}_i) \text{ ....}\mathbf{5}$$

Evaluating the third term (cross product term) is

$$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})(Y_i - \hat{Y}_i) \text{ ....}\mathbf{6}$$

$$(\hat{Y}_i - \overline{Y}) = A + BX_i - A - B\overline{X} \text{ ....}\mathbf{7} \quad \textbf{since} \quad \overline{Y} = A + B\overline{X}$$

$$(\hat{Y}_i - \overline{Y}) = B(X_i - \overline{X}) \text{ ....8}$$

And $(Y_i - \hat{Y}_i) = Y_i - (A + BX_i)$

$$(Y_i - \hat{Y}_i) = Y_i - (\overline{Y} - B\overline{X} + BX_i) \quad \text{since } A = \overline{Y} - B\overline{X}$$

$$(Y_i - \hat{Y}_i) = Y_i - \overline{Y} - B(X_i - \overline{X}) \text{ ....}\mathbf{9}$$

Substituting in 8 and 9 in 6 we obtain:

$$\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})(Y_i-\hat{Y}_i)=\sum_{i=1}^{n} B(X_i-\overline{X})[(Y_i-\overline{Y})-B(X_i-\overline{X})] \text{.................................} \mathbf{9}$$

$$\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})(Y_i-\hat{Y}_i) = Bs_{xy}-B^2s_{xx}=0 \text{ Since } B=\frac{s_{xy}}{s_{xx}} \text{..} \mathbf{10}$$

Therefore applying the result 5 onto 10:

$$\sum_{i=1}^{n} (Y_i-\overline{Y})^2=\sum_{i=1}^{n}[(\hat{Y}_i-\overline{Y})]^2+\sum_{i=1}^{n}[(Y_i-\hat{Y}_i)]^2 \text{...} \mathbf{11}$$

TSS	=	RegSS	+	RSS
Sum of squared total variation		Sum of Squared regression		Sum of squared residual

$$\text{RegSS}=\sum_{i=1}^{n}[(\hat{Y}_i-\overline{Y})]^2=\sum B^2(X_i-\overline{X})^2=B^2 s_{xx} \quad \text{(By applying equation 7)........................} \mathbf{12}$$

$$RSS=\sum_{i=1}^{n}[(Y_i-\hat{Y}_i)]^2=\sum_{i=1}^{n}E_i^2 \sim \chi^2_{n-2} \text{..} \mathbf{13}$$

$$TSS=\sum_{i=1}^{n} (Y_i-\overline{Y})^2 \text{ has the n elements } Y_1-\overline{Y}, Y_2-\overline{Y} \text{......... ....} Y_n-\overline{Y} \text{...............................} \mathbf{14}$$

$$\text{With the constraint } \sum_{i=1}^{n}(Y_i-\overline{Y})=0 \quad \text{(one constraint)}$$

Therefore TSS had the degree of freedom n-1 as n-1 elements can be selected independently but the nth is dependent on the constraint given above.

Degree of freedom has a additive property.

$$TSS=\operatorname{Re}gSS+RSS$$
$$DF_{TSS}=DF_{\operatorname{Re}gSS}+DF_{RSS} \text{ ....} \mathbf{15}$$
$$n-1=\text{.......... ...}+n-2$$

Therefore $DF_{\operatorname{Re}gSS}=1$

ANOVA TABLE

Sources of variation	Sum of Squares	Degree of Freedom	Mean Square	F
Regression	**RegSS**	**1**	$RegMS = \frac{RegSS}{1}$	$F = \frac{RegMS}{RMS}$
Residual	**RSS**	**n-2**	$RMS = \frac{RSS}{n-2}$	
Total	**TSS**	**n-1**		

We have proved that :

$E(RMS) = \sigma_{\varepsilon}^{2}$ ……………………………………………………………….**16**

Similarily:

$E(RegMS) = \sigma_{\varepsilon}^{2} + \beta^2 s_{xx}$ ………………………………………………………….**17**

We have proved that :

$$\frac{(n-2)RMS}{\sigma^2} \sim \chi^2_{n-2}$$

$$\frac{RegMS}{\sigma^2} \sim \chi^2_{1}$$

……………………**18**

are function of random variable Y and independent

Under H_0: $\beta=0$

Statistical Theorem: $\begin{matrix} X \sim \chi^2_m \\ Y \sim \chi^2_n \end{matrix}$ and these are independent then

$\frac{X/m}{Y/n} \sim F_{m,n}$ Ratio of two chi squares follow the F-distribution……………………………………..**19**

Therefore from this theorem we can derive (using 18 and 19):

$$F = \frac{\mathrm{Re}\,gSS}{RMS} \sim F_{1,n-2} \quad \textbf{20}$$

Seeing the expected values $E(\frac{RSS}{n-2}) = \sigma_\varepsilon{}^2$

$$E(RMS) = \sigma_\varepsilon{}^2 \quad \textbf{21}$$

Similarily:

$$E(\mathrm{Re}\,gMS) = \sigma_\varepsilon{}^2 + \beta^2 s_{xx} \quad \textbf{22}$$

If β=0 or close to zero then the ratio F is close to 1 ie the Mean Sum of squared deviation of regression=Sum of squared deviation of Residuals

but if β is not zero then this ratio will be more than zero.

The hypothesis H_0:β=0 can be tested and rejected if $F>F_{\alpha,1,n-2}$

The ANOVA and t test can be used interchangeably for Simple Linear regression.

It can be proved that $F=t^2$

Usually for Multiple Regression ANOVA approach is used.

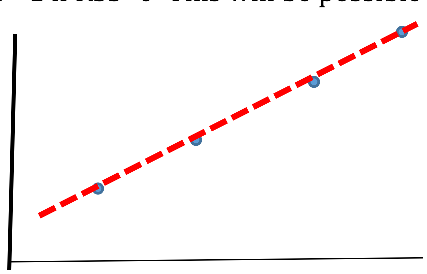
Coefficient of Determination R^2

$R^2 = \frac{\mathrm{Re}\,gSS}{TSS}$ This is one way to evaluate the performance of fitted model ie Goodness of fit **...23**

$R^2 = 1 - \frac{RSS}{TSS}$ $\quad 0 \le R^2 \le 1$ **24**

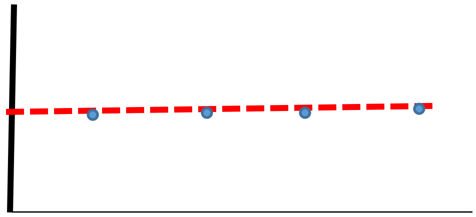
R^2 is the proportion (percentage) of variability in the response that can be explained by the Model (due to Regressor variable).

Case 1) $R^2=1$ if RSS=0 This will be possible if the fitted model explains 100% of the variability in Y.



Case2) $R^2=0$ TSS=RSS therefore $\sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n}[(Y_i - \hat{Y}_i)]^2$ **25**

This is possible when $\hat{Y}_i = \overline{Y}$



This will happen when $\hat{Y} = \overline{Y}$ the fitted model does not depend on the regressor variable. The response variable does not get effected by regressor variable. There is no relationship between X and Y .In other words β=0

$R^2 = \frac{\text{Re}\,gSS}{TSS} = \frac{\beta^2 s_{xx}}{TSS}$ This equation makes sense because R^2 is zero when β=0……………………**26**

CONFIDENCE INTERVAL OF β

The population Linear Regression Equation is : $Y = \alpha + \beta X + \varepsilon$

The least square estimator of β is

$B = \frac{s_{xy}}{s_{xx}}$ ……………………………………………..**27**

This is the **point estimate** of β (population parameter) which will be used to find the sampling distribution of the point estimate.

We have proved that B is an unbiased estimate of β ie $E(B) = \beta$ and that $V(B) = \frac{\sigma^2}{s_{xx}}$ …………**28**

We had also proved that $B = \frac{s_{xy}}{s_{xx}} = \sum K_i Y_i$ so B is a linear combination of Y_is. We had assumed that Yis are normally distributed therefore B is normally distributed as well.

Sampling distribution of B

$B \sim N(\beta, \frac{\sigma_\varepsilon^2}{s_{xx}}) = N(\beta, \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2})$ (Page110)………………………………**29**

$$\frac{B-\beta}{\sqrt{\frac{\sigma_{\varepsilon}{}^{2}}{s_{xx}}}} \sim N(0,1)$$ ……………………………………..**30**

Usually σ is not known therefore we have to replace the σ by the unbiased estimator which is sum of squared residuals.

We have prove that the unbiased estimator of σ^2 is $\frac{RSS}{n-2}$ or $s_E^2 = \frac{\sum_{i=1}^{n} E_i{}^2}{n-2}$ ………………………..**31**

therefore Replacing it in the equation of V(B) (from equation 29)

$$\hat{V}(B) = \frac{s_E^2}{s_{xx}} = \frac{s_E^2}{\sum_{i=1i}^{n} (X_i - \overline{X})^2}$$ ……………………………………**32** (Page 111)

Sampling Distribution of B

$$\frac{B-\beta}{\sqrt{\frac{s_E^2}{s_{xx}}}} \sim t_{n-2}$$ ……………………………………………………**33**

Or

$$\frac{B-\beta}{\sqrt{\frac{RMS}{s_{xx}}}} \sim t_{n-2}$$ **…………………………………………..34** Page 111 in the book

Creating the confidence intervals:

$$P\{-t_{\alpha/2,n-2} \le \frac{B-\beta}{SE(B)} \le t_{\alpha/2,n-2}\} = 1-\alpha$$ …………………………………………..**35**

To make this probability high the alpha value has to be chosen appropriately

Therefore the 100(1-α)% confidence of β is

$$P\{B - t_{\alpha/2,n-2} SE(B) \le \beta \le B + t_{\alpha/2,n-2} SE(B)\} \qquad \textbf{36}$$

Where

$$SE(B) = \sqrt{\frac{RMS}{s_{xx}}}$$

Correlation Coefficient

The correlation coefficient quantifies the linear relationship between two variables. If the correlation coefficient is zero it might be possible that the relationship between two variables is nonlinear instead of linear.

$R^2 = \frac{\mathrm{Re}\,gSS}{TSS} = \frac{\beta^2 s_{xx}}{TSS}$ Usually the correlation coefficient of simple linear regression is represented by r

Correlation can also be represented by the covariance terminology

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

σ_{XY} Is the covariance of random variable X and Y

σ_X Is the standard deviation of X

σ_Y Is the standard deviation of Y

For sample the sample covariance is used

$$S_{XY} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2}}$$