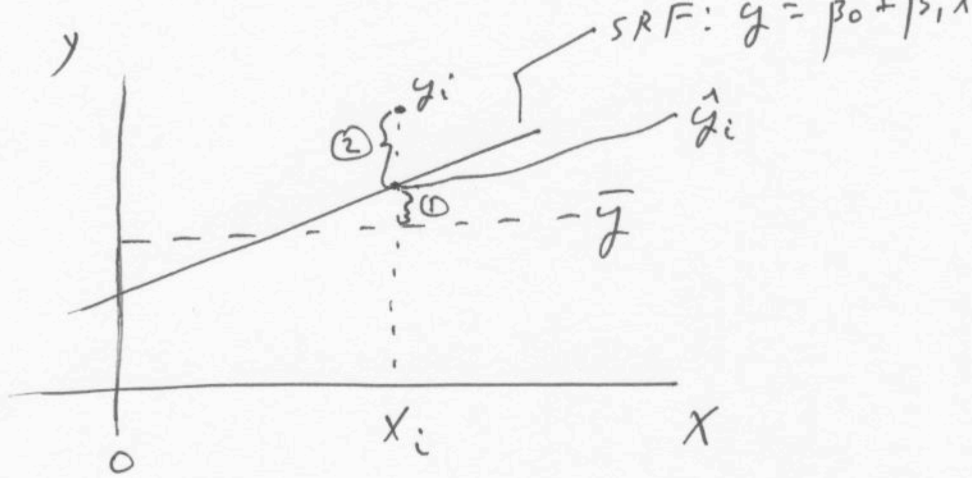
Lecture 9

Partitioning of Total Sum of Squares and Coefficient of Determination (R^2)



Total deviation from mean $= y_i - \bar{y} =$ 1 + 2

Deviation due to regression $= \hat{y}_i - \bar{y} =$ 1

Deviation due to error $= y_i - \hat{y}_i$

$\therefore$ Total deviation = deviation due to regression + deviation due to error

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

As it turns out

$$\sum_{1}^{N}(y_i-\bar{y})^2 = \sum_{1}^{N}(y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2$$

$$= \sum_{1}^{N}(y_i-\hat{y}_i)^2 + 2\sum_{1}^{N}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})$$

$$+ \sum_{1}^{N}(\hat{y}_i-\bar{y})^2$$

$$= \sum_{1}^{N}(y_i-\hat{y}_i)^2 + \sum_{1}^{N}(\hat{y}_i-\bar{y})^2$$

(because it can be shown that $\sum_{1}^{N}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) = 0$)

$$SST = SSR + SSE$$

Total sum of squares = sum of squared residuals + sum of squares explained.

The coefficient of determination R^2 is defined as

$$R^2 = \frac{SSE}{SST} = \frac{\sum_{1}^{N}(\hat{y}_i-\bar{y})^2}{\sum_{1}^{N}(y_i-\bar{y})^2}$$

and $0 \le R^2 \le 1$.

R^2 is interpreted as the percent (in decimal equivalent form) of the variation in y explained by the explanatory variables of a regression.

Analysis of Variance (ANOVA) Table

Source	DF	SS	MS	F
Explained	$K-1$	SSE	$\frac{SSE}{K-1}$	$\frac{MSSE}{MSSR}$
Error	$N-K$	SSR	$\frac{SSR}{N-K}$	
Total	$N-1$	SST		

The null hypothesis of interest for the ANOVA table is

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

$$H_1: \text{not } H_0.$$

This is called the <u>overall</u> test of significance of a multiple regression. The corresponding F-test is called the overall-F-test. This F test (in repeated samples) follows an F distribution with $K-1$ numerator degrees of freedom and $N-K$ denominator degrees of freedom under the truth of the null hypothesis (H_0). If the probability value of the F-statistic of the ANOVA table is greater than α (usually $\alpha = 0.05$) then we accept the null hypothesis that all of the explanatory variables are <u>jointly insignificant</u> and the sample mean provides an adequate description of the variation in the dependent variable. If the F-statistic's p-value is less than α, we reject the null hypothesis and accept the alternative hypothesis that one or more of the proposed explanatory variables provides significant explanatory power in describing the variation in the dependent variable y.