Measures of Linear Association between X and Y

Measuring the utility of the regression model involves quantifying the contribution of X in predicting Y. To do this, we compute how much the prediction errors of Y were reduced by using the information provided by X.

Coefficient of Determination (A measure of "Goodness of Fit")

Coefficient of determination is a natural measure of the effect of X in reducing the variation in Y_i. In reducing the uncertainty in predicting Y, is to express the reduction variation ($SSTO - SSE = SSR$) as a proportion of the total variation:

$$r^2 = \frac{SSR}{SSTO} = \frac{SSTO - SSE}{SSTO} = 1 - \frac{SSE}{SSTO} \qquad \textbf{(3.27)}$$

Since $0 \leq SSE \leq SSTO$, it follows that $0 \leq r^2 \leq 1$.

We may interpret r^2 as *the proportionate reduction of total variation associated with the use of the independent variable X*. Thus, the larger is r^2 the more is the total variation of Y reduced by introducing the independent variable X.

Take Note!!!

The coefficient of determination r^2 (*two variable case*) or r^2 (*multiple regression*) is a summary measure that tells how well the sample regression line fits the data.

Practical Interpretation of the Coefficient of Determination, r^2

"About 100(r^2)% of the sample variation in Y can be explained by (or attributed to) X using to predict Y in the linear regression model".

Properties of r^2

1. It is nonnegative quantity.
2. Its limit are $0 \leq r^2 \leq 1$.

The limiting values of r^2 occur as follows:

1. If all observations fall on the fitted regression line, $SSE = 0$ and $r^2 = 1$.
2. If the slope of the fitted regression line is $\beta_1 = 0$ so that $\hat{Y}_i = \bar{Y}$, $SSE = SSTO$ and $r^2 = 0$.

Source of Image: Applied linear regression models by John Neter

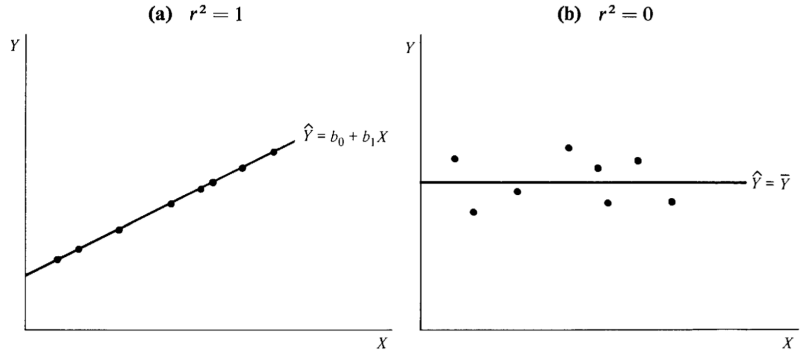

Figure 1. Scatter Plot when $r^2 = 0$ and $r^2 = 1$

Coefficient of Correlation

The square root of r^2:

$$r = \pm\sqrt{r^2} \tag{3.28}$$

is called the *coefficient of correlation*. A plus or minus sign is attached to this measure according to whether the slope of the fitted regression line is positive or negative. Thus, the range of r is:

$$-1 \leq r \leq 1.$$

Take Note!!!

The r mentioned here is the same r mentioned in the previous topic (Pearson Product Moment Correlation Coefficient).

There is a relation between β_1 and r that is worth nothing.

$$\beta_1 = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}\ (r) = \left(\frac{\hat{\sigma}_Y}{\hat{\sigma}_X}\right) r \tag{3.29}$$

If $r = 0$, then $\beta_1 = 0$, and vice versa. Therefore, if the value of r is negative, we would expect that the value of β_1 is also negative, implying an inverse correlation between X and Y, since the sample standard deviation is always a non-negative value.

Exercise:

Using formula (3.29), compute β_1 and observe whether the direction of the regression coefficient is the same as the coefficient of correlation.

Inference About the Individual Regression Parameter

Inference about β_0 and β_1:

Table 1: Test of Hypotheses for β_0 and β_1

Parameter	Null Hypothesis, H_o	Alternative Hypothesis, H_a	Test Statistic	Region of Rejection
β_0	$\beta_0 = 0$	$\beta_0 \neq 0$	$t = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}(\hat{\beta}_0)}$	$\pm t_{(1-\alpha/2, n-2)}$
β_1	$\beta_1 = 0$	$\beta_1 \neq 0$	$t = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}(\hat{\beta}_1)}$	$\pm t_{(1-\alpha/2, n-2)}$

What does the coefficient mean?

The sign of a linear regression coefficient tells you whether there is a *positive* or *negative correlation* between each independent variable and the dependent variable.

- A positive coefficient indicates that as the value of the independent variable increases, the mean of the dependent variable also tends to increase.
- A negative coefficient suggests that as the independent variable increases, the dependent variable tends to decrease.

A p-value below 0.05 indicates 95% confidence that the slope of the regression line is not zero and hence there is a significant linear relationship between the dependent and independent variables.

Take note!!!

The ***coefficient value*** (β_1) signifies how much the mean of the dependent variable ***changes given a one-unit shift*** in the independent variable while holding other variables in the model constant.

Sample Interpretation:

The dependent variable is child's height and the independent variable is parents' height, and the value of the slope regression coefficient is 0.6463. The results suggest that:

"For every 1 inch increase in the parents' height, we estimate a 0.6463 inch increase in the child's height".

Inference about the model:

Table 2: Test of Hypothesis of Regression Model

Parameter	Null Hypothesis, H_o	Alternative Hypothesis, H_a	Test Statistic	Region of Rejection
$E(Y) = \mu$	$E(Y) = \beta_0 + \beta_1 X_i + \ldots + \beta_i X_i$	$E(Y) \neq \beta_0 + \beta_1 X_i + \ldots + \beta_i X_i$	$F = \frac{MSR}{MSE}$	$F_{(\alpha, df_1=1, df_2=n-2)}$

For a given α level, the F test of $\beta_1 = 0$ versus $\beta_1 \neq 0$ is **equivalent algebraically to the two-tailed *t*-test**.

$$F = \frac{MSR}{MSE} = \left[\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}(\hat{\beta}_1)}\right]^2 = (t)^2$$

Exercise:

Show that $F = \frac{MSR}{MSE} = \left[\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}(\hat{\beta}_1)}\right]^2 = (t)^2$.

Confidence Interval about β_0 and β_1:

The confidence-interval approach is the concept of **interval estimation.** An interval estimator is an interval or range constructed in such a manner that it has a specified probability of including within its limits the true value of the unknown parameter. The interval thus constructed is known as a **confidence interval,** which is often stated in percent form, such as 90% or 95%. The confidence interval provides a set of plausible hypotheses about the value of the unknown parameter. If the null hypothesized value lies in the confidence interval, the hypothesis is not rejected, whereas if it lies outside this interval, the null hypothesis can be rejected.

$100(1-\alpha)\%$ **Confidence Interval for β_0:** $\hat{\beta}_0 \pm t_{(1-\alpha/2, n-2)}\hat{\sigma}(\hat{\beta}_0)$

$100(1-\alpha)\%$ **Confidence Interval for β_1:** $\hat{\beta}_1 \pm t_{(1-\alpha/2, n-2)}\hat{\sigma}(\hat{\beta}_1)$

Standardized Coefficient of β:

$$\hat{\beta}_{z,1} = \left(\frac{\hat{\sigma}_x}{\hat{\sigma}_y}\right)\hat{\beta}_1$$

$$\hat{\beta}_{z,2} = \left(\frac{\hat{\sigma}_x}{\hat{\sigma}_y}\right)\hat{\beta}_2$$

$$\hat{\beta}_{z,i} = \left(\frac{\hat{\sigma}_x}{\hat{\sigma}_y}\right)\hat{\beta}_i$$

Where:

$$\hat{\sigma}_x = \sqrt{\frac{\sum (X - \bar{X})}{n-1}} \text{ and } \hat{\sigma}_y = \sqrt{\frac{\sum (Y - \bar{Y})}{n-1}}$$

Exercise:

Using the data in the previous Exercises, compute the following:

1. Test statistic of β_0.

2. Test statistic of β_1.

3. Confidence interval of β_0.

4. Confidence interval of β_1.

5. Adjuster r-squared.

6. Determine if the independent variable in the previous data, is significant predictor of the dependent variable.

Prepared by:

KATRINA D. ELIZON
Department of Mathematics and Statistics
College of Science