# LECTURE - Regression

## Simple Linear Regression (SLR) Model -Part 2

How good is this model?

Now we have the equation to predict one future value as well as its interval estimation. But these prediction is as good as the model. If the model is not good then the predictions are also not expected to be good as predictions are based on the model estimation.

Models are judged by reduction of error or amount of unexplained error for the model. As seen before, the sum of squares of error is a measure of error. If the information regarding X is ignored then it is expected that predicted value of Y is $\bar{Y}$. As a result the sum of squares of error for that is

$$SST = \sum_{1}^{n} (y_i - \bar{y})^2 \quad \text{read as Total Sum of Squares}$$

On the other hand if we use the least square line to predict using the value of X, then sum of squares of error is

$$SSE = \sum_{1}^{n} (y_i - \hat{y}_i)^2 \quad \text{read as Sum of Squares due to Error}$$

As we know that least square line will have have the least amount of error among any lines, it is clear that $SST \geq SSE$.

The reduction in error is called Sum of Squares due to Regression, hence

$$SSR = SST - SSE = \sum_{1}^{n} (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1 . S_{xy}$$

So in short, the reduction in error due to model is SSR out of SST. Hence a natural measure of "how good is the model" is the proportion of error saved. That's why **Coefficient of Determination -** $R^2$ is defined as

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{b_1 \sum X_i Y_i + b_o \sum Y_i - n\bar{Y}^2}{\sum Y_i^2 - n\bar{Y}^2}$$

$R^2$ is also called the proportion of variation explained by the regressor X. Because $0 \leq SSR \leq SST$ , it follows that $0 \leq R^2 \leq 1$. Values of $R^2$ that are close to 1 imply that most of the variability in y is explained by the regression model. In practice, $R^2$ is multiplied by 100 and expressed as % so easy understanding.

Though $R^2$ is one of the most popular criterion people look at, it should be used with caution. It is always possible to make $R^2$ large by adding enough terms to the model (will be discussed in MLR model). For example, if there are no repeat points (more than one y value at the same x value), a polynomial of degree (n-1) will give a perfect fit ($R^2 = 1$) to n data points.

Also the magnitude of $R^2$ depends on the range of variability in the regressor variable. Generally $R^2$ will increase as the spread of the x's increases and decrease as the spread of the x's decreases provided the assumed model form is correct. It is possible to show that expected value of $R^2$ from a straight-line regression is approximately

$$E(R^2) \approx \frac{\beta_1^2 \, S_{xx}/n-1}{\beta_1^2 \, S_{xx}/(n-1) + \sigma^2}$$

Clearly the expected value of $R^2$ will increase (decrease) as $S_{xx}$ (a measure of the spread of the x's) increases (decreases). Thus, a large value of $R^2$ may result simply because x has been varied over an unrealistically large range. On the other hand, $R^2$ may be small because the range of x was too small to allow its relationship with y to be detected. (We will discuss it more in MLR Model.)

## An alternative Form

There is an alternate form of the simple linear regression model that is occasionally useful. Suppose that we redefine the regressor variable $X_i$ as the deviation from its own average say,$(x_i - \bar{x})$. The regression model then becomes

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \, X + e \\ &= \beta_0 + \beta_1 \, (X_i - \bar{X}) + \beta_1 \, \bar{X} + e \\ &= (\beta_0 + \beta_1 \, \bar{X}) + \beta_1 \, (X_i - \bar{X}) + e = \beta_0^{\star} + \beta_1 \, (X_i - \bar{X}) + e \end{aligned}$$

Note that redefining the regressor variable above has shifted the origin of the x's from $\bar{x}$ to zero. In order to keep the fitted values the same in both the original and transformed models, it is necessary to modify the original intercept. It is easy to show that the least-squares estimator of the transformed intercept is $\hat{\beta}^{\star} = \hat{\beta}_0 + \hat{\beta}_1 . \bar{x}$. The estimator of the slope is unaffected by the transformation. This alternate form of the model has some advantages. First, the least-squares estimators of the slope and the intercept are uncorrelated, that is, This will make some applications of the model easier, such as finding confidence intervals on the mean of y. Finally, the fitted model is

$$\hat{y} = \bar{y} + \beta_1 . (x - \bar{x})$$

# Correlation

The linear regression model that we have presented in this chapter assumes that the values of the regressor variable x are known constants. There are many situations in which assuming that the x's are fixed constants is inappropriate. It is more reasonable to assume that both y and x are random variables. Fortunately, under certain circumstances, all of our earlier results on parameter estimation, testing, and prediction are valid. We now discuss these situations.

One of the most common statistic people look at even before finding the least square line is estimated value of correlation coefficient r. This statistic is model free but it has lot of connections with SLR model. The population correlation is defined as

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X).Var(Y)}} = \frac{E(XY) - E(X).E(Y)}{\sqrt{Var(X).Var(Y)}}$$

Then the estimated value of $\rho$ is give by

$$\hat{\rho} = r = \frac{\sum XY - n.\bar{X}.\bar{Y}}{\sqrt{(\sum X^2 - n.\bar{X}^2) \times (\sum Y^2 - n.\bar{Y}^2)}} = \frac{S_{xy}}{\sqrt{S_{xx}.S_{yy}}}$$

It follows from the above formula that r is related to SLR models as

$$\hat{\rho}^2 = r^2 = \frac{S_{xy}^2}{S_{xx}.S_{yy}} = \hat{\beta_1}^2.\frac{S_{xx}}{S_{yy}} = \frac{SSR}{SST} = R^2 \quad \text{(Note that } S_{yy} = SST\text{)}$$

Remark: Even though $R^2$ has a completely different interpretation, it is numerically equal to $r^2$. (probably that's where it got its name from).

Now following our example, we see that

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}.SS_{yy}}} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum_{i=1}^n x_i^2 - n\bar{x}^2).(\sum_{i=1}^n y_i^2 - n\bar{y}^2)}}$$

$$= \frac{125.94 - 39 \times 1.73 \times 1.53}{\sqrt{(144.58 - 39 \times 1.73^2) \times (111.34 - 39 \times 1.53^2)}}$$

$$= \frac{22.69}{\sqrt{27.28 \times 20.04}} = 0.9602.$$

## Testing Table for $\rho$

| | | | |
|---|---|---|---|
| Hypothesis | $H_o : \rho \geq 0$<br>$H_1 : \rho < 0$ | $H_o : \rho \leq 0$<br>$H_1 : \rho > 0$ | $H_o : \rho = 0$<br>$H_1 : \rho \neq 0$ |
| Test Statistic | $t - stat = \frac{r}{\sqrt{(1-r^2)/(n-2)}}$ | | |
| Rejection Region | $t - stat < -t_{\alpha,n-2}$ | $t - stat > t_{\alpha,n-2}$ | $\|t - stat\| > t_{\alpha/2,n-2}$ |
| p-value | $P(t < t - stat)$ | $P(t > t - stat)$ | $2.P(t > \|t - stat\|)$ |

Testing against non-zero value and Confidence interval for population correlation $\rho$ will be discussed in Part-4.

# Inference about Regression Parameters

Once SLR model is assumed and least squares estimators are calculated, it is time to look at the inference for the $\beta$'s. First it is easy to see that the point estimators (the least square estimators) for the $\beta$'s are simply linear function (or weight average) of $Y_i$s.

$$\hat{\beta}_1 = \frac{Sxy}{Sxx} = \frac{\sum(x_i-\bar{x})(Y_i-\bar{Y})}{\sum(x_i-\bar{x})^2} = \sum\left(\frac{x_i-\bar{x}}{Sxx}\right)Y_i - \bar{Y}\sum\left(\frac{x_i-\bar{x}}{Sxx}\right) = \sum c_{1i}.Y_i$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1.\bar{x} = \sum\frac{Y_i}{n} - (\sum c_{1i}.Y_i)\bar{x} = \sum\left(\frac{1}{n} - c_{1i}.\bar{x}\right).Y_i = \sum c_{0i}.Y_i$$

It is to note that in the model, variable X is known and not a random variable but Y is a random variable. Since the errors $e_i$ are I.I.D $N(0, \sigma^2)$, the observations $y_i$'s are I.I.D $N(\beta_0 + \beta_1.x_i, \sigma^2)$. As $\hat{\beta}_i^{'}s$ are linear combination of the observations, they normally distributed with mean $\beta_i$ and variance $w_i.\sigma^2$. where $w_0 = \frac{\sum x_i^2}{n.Sxx} = \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\ and\ w_1 = \frac{1}{Sxx}$

The inferences are based on that facts. It can be easily shown that those point estimators are unbiased.

and 100(1-$\alpha$)% Confidence Interval for $\beta_i$ is $\hat{\beta}_i \pm t_{\alpha/2,n-2}.\hat{\sigma}.\sqrt{w_i}$

# Testing Table for $\beta_i$

|  |  |  |  |
|---|---|---|---|
| Hypothesis | $H_o : \beta_i \geq \beta_{io}$<br>$H_1 : \beta_i < \beta_{io}$ | $H_o : \beta_i \leq \beta_{io}$<br>$H_1 : \beta_i > \beta_{io}$ | $H_o : \beta_i = \beta_{io}$<br>$H_1 : \beta_i \neq \beta_{io}$ |
| Test Statistic | $t - stat = \frac{\hat{\beta}_i - \beta_{io}}{\hat{\sigma}\sqrt{w_i}}$ where $w_0 = \frac{\sum x_i^2}{n.S_{xx}}$ $and$ $w_1 = \frac{1}{S_{xx}}$ |  |  |
| Rejection Region | $t - stat < -t_{\alpha,n-2}$ | $t - stat > t_{\alpha,n-2}$ | $\lvert t - stat\rvert > t_{\alpha/2,n-2}$ |
| p-value | $P(t < t - stat)$ | $P(t > t - stat)$ | $2.P(t > \lvert t - stat\rvert)$ |

Inference regarding $\beta_i$ 's can be done using above table.

A very important special case of the hypotheses is when $\beta_1$ is tested against 0. These hypotheses relate to the significance of regression. Failing to reject $H_0 : \beta_1 = 0 \ vs \ H_1 : \beta_1 \neq 0$ implies that there is no linear relationship between X and Y as x will drop out of the equation. Note that this may imply either that x is of little value in explaining the variation in y and/or that the true relationship between x and y is not straight line. Therefore, failing to reject $H_0 : \beta_1 = 0$ is equivalent to saying that there is no linear (straight line) relationship between y and x.

This is also called overall test for regression. Hence it can be done from overall regression point of view as follows.

$H_0 : \beta_1 = 0 \ vs \ H_1 : \beta_1 \neq 0,$

test-statistic $= F - stat = \frac{SSR/(2-1)}{SSE/(n-2)}$

Reject$H_0 \ if \ F - stat > F_{\alpha,1,n-2}$

Remark: Both the tests for $\beta_1$ against 0 will result in a same decision all the time and p-values will be identical. Relationship between the statistic is

$(t - stat)^2 = F - stat \quad and \quad t^2_{\alpha/2,n-2} = F_{\alpha,1,n-2}$

Following our example, we find the confidence intervals as follows:

95% Confidence Interval for $\beta_1$ is

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot \frac{\hat{\sigma}}{\sqrt{S_{xx}}} = 0.8319 \pm 2.026 \times \frac{0.2078}{\sqrt{27.2784}} = (0.7513, 0.9125)$$

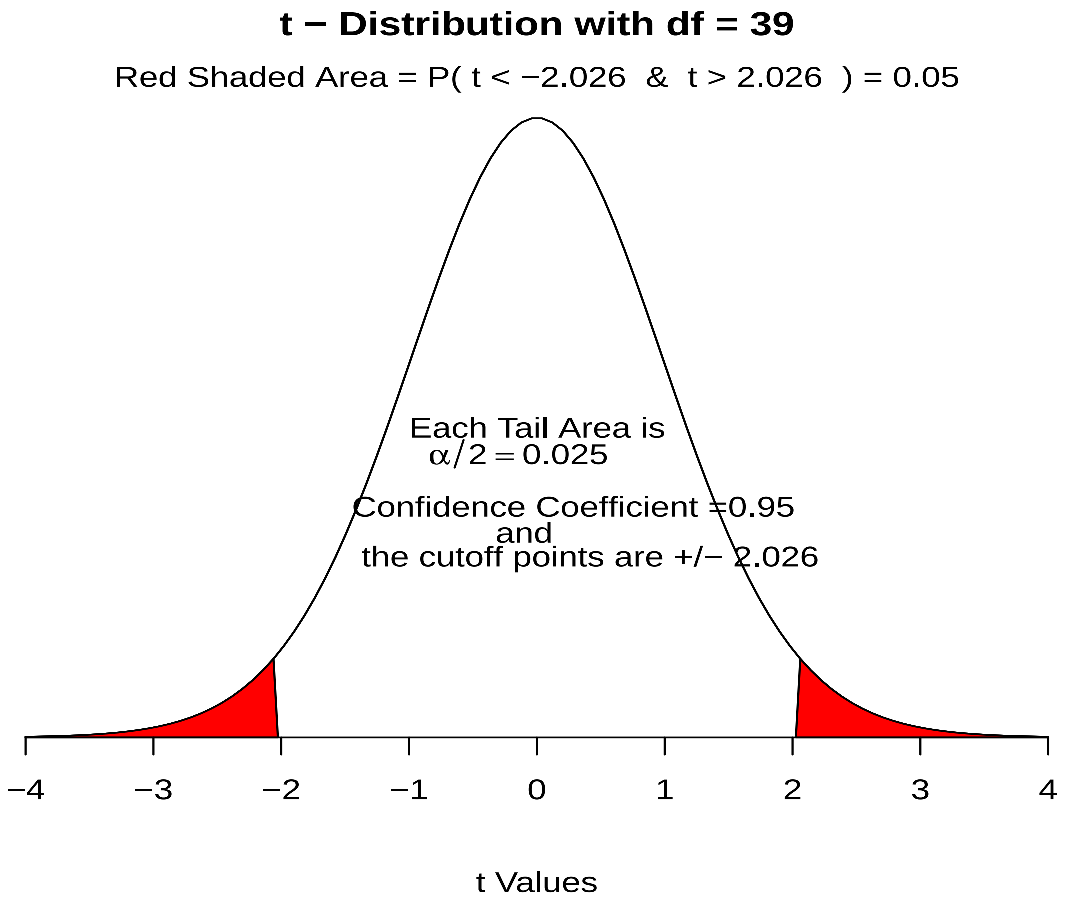

Figure 1: t Distribution with df=37 - Right Tail $\alpha = 0.05$

Similarly, 95% Confidence Interval for $\beta_0$ is

$$\hat{\beta}_0 \pm t_{\alpha/2,n-2}.\,\hat{\sigma} \times \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$

$$=0.0836 \pm 2.026 \times 0.2078 \times \sqrt{\left(\frac{1}{39} + \frac{1.7343^2}{27.2785}\right)} = (-0.0716, 0.2388)$$

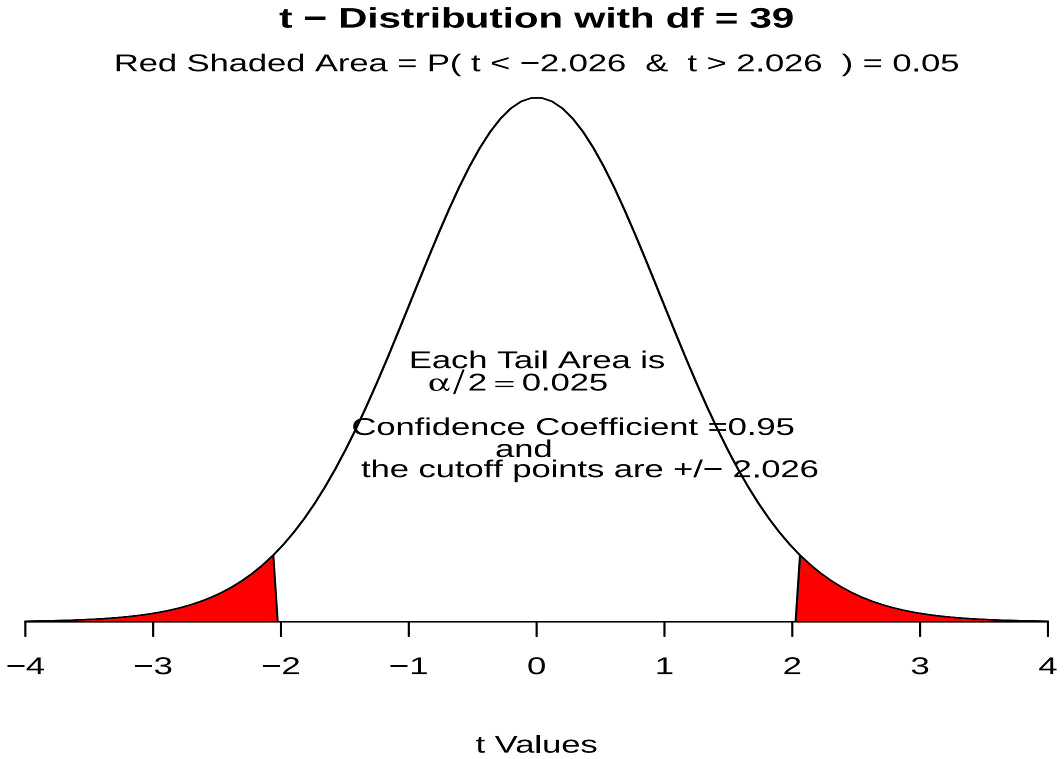

Figure 2: t Distribution with df=37 - Right Tail $\alpha = 0.05$

To perform testing on $\beta_1$ we set it up as follows:

$$H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$$

$$t - stat = \frac{\beta_1 \hat{-} \beta_{i0}}{\hat{\sigma}/S_{xx}} = \frac{0.8319 - 0}{0.2078/\sqrt{27.2784}} = 20.9140$$

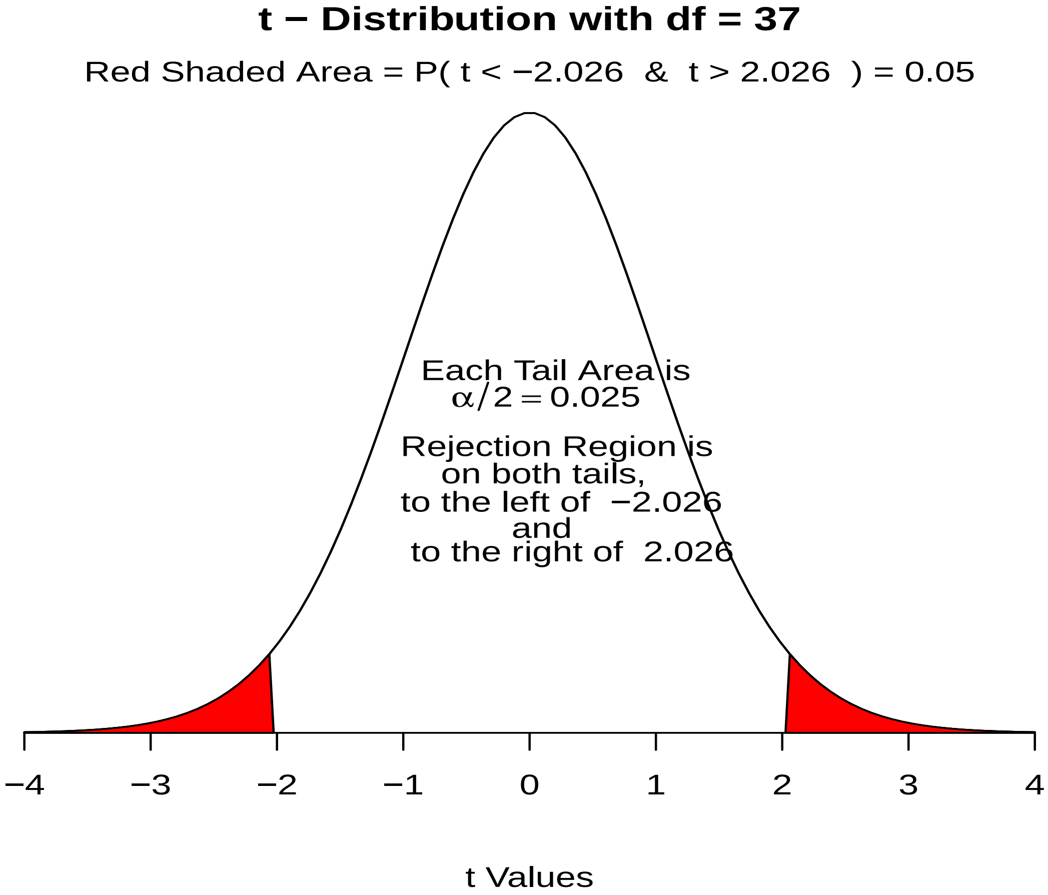

Figure 3: t Distribution with df=37 - Two Tailed Rejection Region $\alpha = 0.05$

All the softwares present the results of regression in two tables and anything can be calculated from these two. ANOVA table is a macro view of the model and Parameter estimates are micro view of the model.

## ANOVA TABLE FOR SLR MODEL

| Source | SS | df | MS | F-Stat | p-Value |
|---|---|---|---|---|---|
| Regression | $SSR$ | (2-1) | $MSR = \frac{SSR}{(2-1)}$ | $F - stat = \frac{MSR}{MSE}$ | $P(F > F - stat)$ |
| Error | $SSE$ | (n-2) | $MSE = \frac{SSE}{(n-2)}$ | | |
| Total | $SST$ | (n-1) | | | |

For our example the ANOVA table is as follows:

| Source | SS | df | MS | F-Stat | p-Value |
|---|---|---|---|---|---|
| Regression | 18.8793 | (2-1) | 18.8793 | 437.3962 | 0.00 |
| Error | 1.5970 | (39-2) | 0.0432 | | |
| Total | 20.4763 | (39-1) | | | |

The table below provides the information regrading the estimated regression parameters individually in each line, (Usually Confidence Interval Column is not included in the table. It is given here for completeness. Also you won't be able to calculate the p-value from the table but bounds can be provided).

## Parameter Estimates

| Parameter | Estimate | St. Deviation | t-stat | p-Value | Conf. Interval |
|---|---|---|---|---|---|
| $\beta_0$ | $b_0$ | $S_{b_0}$ | $t_0 = \frac{b_0}{S_{b_0}}$ | $2P(t > \|t_0\|)$ | $b_0 \pm t_{\alpha/2,n-k-1}.S_{b_0}$ |
| $\beta_1$ | $b_1$ | $S_{b_1}$ | $t_1 = \frac{b_1}{S_{b_1}}$ | $2P(t > \|t_1\|)$ | $b_1 \pm t_{\alpha/2,n-k-1}.S_{b_1}$ |

For our example, the table looks like

| Parameter | Estimate | St. Deviation | t-stat | p-Value | Conf. Interval |
|---|---|---|---|---|---|
| $\beta_0$ | 0.0836 | 0.0766 | 1.0916 | 0.2821 | (-0.0716, 0.2388) |
| $\beta_1$ | 0.8319 | 0.0398 | 20.9140 | 0.00 | (0.7513, 0.9125) |