# Hypothesis Testing – Accuracy Check

**Null Hypothesis:** There is no relation ship between X and Y.

$$\theta_1 = 0 \quad \Rightarrow \quad Y = \theta_0$$

**Alternative Hypothesis:** There is some relation ship between X and Y.

$$\theta_1 \neq 0 \quad \Rightarrow \quad Y = \theta_0 + \theta_1 * X$$

→ To test null hypothesis, we need to determine whether $\Theta_1$ is sufficiently far from zero.
→ The $\Theta_1$ - 0 will give us the distance from zero, but it would be better if we can find probability of $\Theta_1$being close to zero. If this probability (p-value) is less than 5%, we can reject null hypothesis.
→ To find the probability, we have to calculate t-statistic. The formula is as below

$$t = \frac{\Theta_1 - 0}{\mathrm{SE}(\Theta_1)}$$

→ The probability of observing any value equal to |t| is 0 or larger, assuming $\Theta_1$ = 0 (null hypothesis) . We call this probability the p-value.

Mean of the target variable y (sample) → $\hat{\mu} = \bar{y}$, where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Standard Error - mean of the target variable (from CLT) → $\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}$

$$\mathrm{SE}(\theta_0)^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right], \quad \mathrm{SE}(\theta_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

95% Confidence Interval $\left[\theta_1 - 2 \cdot \mathrm{SE}(\theta_1),\ \theta_1 + 2 \cdot \mathrm{SE}(\theta_1)\right]$

# Assessing the Accuracy of the Model

**Residual Standard Error:** The RSE is an estimate of the standard of the error (epsilon).

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

The RSE is considered a measure of the lack of fit of the model to the data. If the predictions obtained using the model are very close to the true outcome values—that is, if $\hat{y}_i \approx y_i$ i for i = 1, . . . , n—then will be small, and we can conclude that the model fits the data very well. On the other hand, if $\hat{y}_i$ is very far from $y_i$ for one or more observations, then the RSE may be quite large, indicating that the model doesn't fit the data well.

$$\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$RSS = \sum_{i=1}^{n}(y_i - f(x_i))^2$$

To calculate $R^2$, we use the formula

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)},$$

→ Adjusted R - Squared

$$\mathrm{R}_a^2 = 1 - \left[ \left( \frac{\mathrm{n}-1}{\mathrm{n}-\mathrm{k}-1} \right) \times (1 - \mathrm{R}^2) \right]$$

where:
n = number of observations
k = number of independent variables
$\mathrm{R}_a^2$ = adjusted $\mathrm{R}^2$

→ To find probability we need to find standard error of $\Theta_1$

Residual Standard Error (RSE) $\sqrt{\sum_{i=1}^{n} (\hat{y} - y_i)^2 / (n-2)}$

→ Then calculate t-statistic (measure of standardization that $\Theta_1$ is away form zero)

t-statistic = $(\theta_1 - 0)/RSE$

→ Now we can find the probability of $\Theta_1$ being close to zero by looking into t-Distribution table

$$\mathrm{SE}(\hat{\theta}_0)^2 = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\theta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

## Regression Analysis – Multiple regression

Regression Analysis: The goal of regression analysis is to describe the relationship between one set of variables, called the dependent variables, and another set of variables, called independent or explanatory variables.

If the relationship between the dependent and explanatory variable is linear, that’s linear regression. For example, if the dependent variable is y and the explanatory variables are $x_1$, $x_2$, we would write the following linear regression model:

$$y = \theta_0 + \theta_1 * x_1 + \theta_2 * x_2 + \epsilon$$

$\theta_1$ is associated with $x_1$, $\theta_2$ is associated with $x_2$, $\epsilon$ is the error due to random variation or other factor.