Notes on Test Statistics

t-test

Introduction

In a linear regression model, we are interested in testing whether a particular coefficient is significantly different from zero. Consider the simple linear regression model:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where Y is the dependent variable, X is the independent variable, β_0 is the intercept, β_1 is the slope, and ϵ is the error term assumed to be normally distributed with mean 0 and variance σ^2.

The t-test helps us determine whether the estimated slope coefficient $\hat{\beta}_1$ is significantly different from zero, indicating whether the predictor X has a meaningful relationship with the response Y.

The null hypothesis for the t-test is:

$$H_0 : \beta_1 = 0$$

The alternative hypothesis is:

$$H_A : \beta_1 \neq 0$$

t-Test Statistic

The t-statistic is calculated as:

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

where $\hat{\beta}_1$ is the estimated slope coefficient, and $SE(\hat{\beta}_1)$ is the standard error of $\hat{\beta}_1$, given by:

$$SE(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum (X_i - \bar{X})^2}}$$

where $\hat{\sigma}^2$ is the estimate of the variance of the residuals, X_i are the individual values of X, and $\bar{X}$ is the mean of X.

Deriving the t-Test

In linear regression, the coefficient $\beta 1$ is estimated using the method of ordinary least squares (OLS). The OLS estimator for $\beta 1$ is:

$$\hat{\beta}_1 = \frac{\sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum \left(X_i - \bar{X}\right)^2}$$

Under the assumption that the errors ϵ are normally distributed with mean zero and variance σ^2, the estimator $\hat{\beta} 1$ is also normally distributed with:

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum(X_i - \bar{X})^2}\right)$$

This tells us that $\hat{\beta}_1$ follows a normal distribution centered around the true value β_1 with variance $\frac{\sigma^2}{\sum\left(X_i - \bar{X}\right)^2}$.

Since the true variance σ^2 is unknown, we estimate it using the residual sum of squares from the model:

$$\hat{\sigma}^2 = \frac{\sum \hat{\epsilon}_i^2}{n-2}$$

where n is the number of observations, and $\hat{\epsilon}_i$ are the residuals of the model. The standard error of $\hat{\beta}_1$ is thus:

$$SE\left(\hat{\beta}_1\right) = \sqrt{\frac{\hat{\sigma}^2}{\sum\left(X_i - \bar{X}\right)^2}}$$

Using the standard error, we standardize $\hat{\beta}_1$ to form the t-statistic:

$$t = \frac{\hat{\beta}_1}{SE\left(\hat{\beta}_1\right)}$$

Under the null hypothesis $H_0 : \beta_1 = 0$, this t-statistic follows a t-distribution with $n-2$ degrees of freedom.

F-Test

Introduction

The F-test in linear regression is used to evaluate the overall significance of the model or to compare different models. While the t-test assesses the significance of individual coefficients, the F-test checks whether the regression model as a whole explains a significant portion of the variation in the dependent variable.

F-Test Statistic

The F-test statistic is computed as:

$$F = \frac{\text{Explained Variance } /p}{\text{Unexplained Variance } /(n-p-1)}$$

where p is the number of predictors, n is the number of observations, the numerator reflects the variance explained by the regression model, and the denominator reflects the variance that is unexplained (the residuals). In this form, the F-statistic follows an F-distribution with p degrees of freedom in the numerator and $n-p-1$ degrees of freedom in the denominator.

We can rewrite the F-test as:

$$F = \frac{\text{Mean Sum of Squares for the Model (MSM)}}{\text{Mean Sum of Squares for the Residuals (MSE)}}$$

where:

– The Mean Sum of Squares for the Model (MSM) is the explained variance, defined as:

$$MSM = \frac{\sum \left(\hat{Y}_i - \bar{Y}\right)^2}{p} = \frac{SSR}{p}$$

where $\hat{Y}_i$ is the predicted value for Y_i and $\bar{Y}$ is the mean of the observed values of Y.

– The Mean Sum of Squares for the Residuals (MSE) is the unexplained variance, defined as:

$$MSE = \frac{\sum \left(Y_i - \hat{Y}_i\right)^2}{n - p - 1} = \frac{SSE}{n - p - 1}$$

where Y_i are the observed values and $\hat{Y}_i$ are the predicted values.

Under the null hypothesis this F-statistic follows an F-distribution with p and $n-p-1$ degrees of freedom.