Module 3: Linear Regression

Dan the Quant

June 14, 2025

Abstract

Many of those who work in finance and economics might argue that one of humanity's greatest inventions is linear regression and the Ordinary Least Squares algorithm. A large part of modern economic and social science is built upon the simplicity and elegance of the linear regression method. It holds within itself a magnificent power to provide sustainability and substance to many of the great economic and financial problems that have been solved in recent decades.

Contents

1 Simple Linear Regression Model

Simple linear regression is a statistical model that helps us identify the relationship between two variables, a dependent one that we call y, and an independent one that we call x. The primary objective of linear regression is to explain the relationship between the two variables, as well as to predict the values for given values of x. This model assumes a strictly linear relationship between both variables (hence its name) and is represented by the following form:

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1}$$

The coefficients β determine the relationships between the dependent variable and the explanatory variable. First, β_1 represents the absolute rate of change that links x to y: if x changes by one unit, y changes by β_1 units. On the other hand, β_0 represents the average value of the dependent variable when the independent variable is zero. In other words, it accounts for values that cannot be explained by the model. Geometrically, this value is the intercept of the line on the y-axis (axis of ordinates).

In addition, there is an error term (ε), which represents the discrepancies between our linear regression model and reality. While statistical models are useful for understanding reality, they do not always perfectly capture natural observations. Therefore, it is important to account for these differences to analyze them appropriately.

1.1 Ordinary Least Squares (OLS)

The algorithm used to find this combination of coefficients β is called **Ordinary Least Squares (OLS)**. This method is widely used in economics and the social sciences due to its efficiency in identifying relationships between two or more variables. To derive OLS, we begin with the Linear Sample Regression equation (which is the expected value of the Population Linear Regression, previously shown in Equation 1).

$$E[y] = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{2}$$

What differentiates this model from the one seen above (in addition to the accents on the coefficients) is that we are not able to capture the entire population to fully understand a phenomenon; therefore, we will never be able to know the true β coefficients, so we seek to get some estimators $\hat{\beta}$. On the other hand, $\hat{\varepsilon}$ will be the residuals of the regression, if we clear for this variable:

$$\hat{\varepsilon} = y - \hat{y} \tag{3}$$

$$\hat{\varepsilon} = y - \left(\hat{\beta}_0 + \hat{\beta}_1 x\right) \tag{4}$$

One of the main assumptions of the OLS estimator is that the expected value of the residuals is zero; hence, their sum will also be zero. Therefore, we can square the equation and then apply the summation operator.

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 \tag{5}$$

The left-hand part of the equation is equivalent to the sum of squares of the error we talked about earlier. Now, what we want is to find what is the combination of coefficients $\hat{\beta}$ that minimize this variable. Therefore, what we will do is derive with respect to both coefficients, remembering that the derivative of a sum of a function is equal to the sum of the derivatives of it:

$$\frac{\partial \sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\partial \hat{\beta}_0} = \sum_{i=1}^{n} 2\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)(-1) = 0 \tag{6}$$

$$\frac{\partial \sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\partial \hat{\beta}_1} = \sum_{i=1}^{n} 2\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)(-x_i) = 0 \tag{7}$$

Since this is a minimization problem, we must set the derivatives equal to zero. By doing so, we obtain a system of two equations with two unknowns; we can also make some mathematical tricks to simplify the expressions. Remember, we can factor out all constants from the summation and simplify by dividing by two, which eliminates the coefficients of two. Ultimately, we derive two equations that involve only the independent and dependent variables, along with the coefficients to be estimated.

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \tag{8}$$

$$\sum_{i=1}^{n} \left(y_i x_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2\right) = 0 \tag{9}$$

To obtain $\hat{\beta}_0$, we start from the first equation and distribute the summation sign across all values within the parentheses. This step is mathematically valid because the summation operator preserves the property of commutativity.

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 x_i = 0 \tag{10}$$

Let us remember that $\hat{\beta}_0$ and $\hat{\beta}_1$ are both constants, as they can only take specific values. Therefore, we can factor them out of their respective summations. Consequently, we multiply the entire equation by $\frac{1}{n}$ to simplify further.

$$\frac{1}{n}\left[\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 x_i\right] = 0 \tag{11}$$

$$\frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\hat{\beta}_0 \sum_{i=1}^{n} 1 - \frac{1}{n}\hat{\beta}_1 \sum_{i=1}^{n} x_i = 0 \tag{12}$$

Additionally, we can recall the formula for the arithmetic mean, given by $\bar{X} = \frac{1}{n}\sum X$. Recalling that $\sum 1 = n$ and distributing the $\frac{1}{n}$:

$$\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} = 0 \tag{13}$$

And finally simplifying and clearing for $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{14}$$

To obtain $\hat{\beta}_1$, we start with the second equation derived from minimizing the Sum of Squares of the Error. Applying the same rules previously described, we simplify arriving at the following mathematical expression.

$$\sum_{i=1}^{n} y_i x_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \tag{15}$$

Now, if we substitute $\hat{\beta}_0$ with the expression we derived earlier, we obtain a new equation. Simplifying by solving the terms within the parentheses, we get:

$$\sum_{i=1}^{n} y_i x_i - \left[\bar{y} - \hat{\beta}_1 \bar{x}\right] \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \tag{16}$$

$$\sum_{i=1}^{n} y_i x_i - \bar{y} \sum_{i=1}^{n} x_i + \hat{\beta}_1 \bar{x} \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \tag{17}$$

By isolating the components multiplied by $\hat{\beta}_1$ on the left side of the equation, we can then factor out $\hat{\beta}_1$ as a common term:

$$\hat{\beta}_1 \bar{x} \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \bar{y} \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i x_i \tag{18}$$

$$\hat{\beta}_1 \left[\bar{x} \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i^2\right] = \bar{y} \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i x_i \tag{19}$$

Multiplying both sides of the equation by $\frac{1}{n}$, we obtain the following:

$$\hat{\beta}_1 \left[\bar{x} \frac{1}{n} \sum_{i=1}^{n} x_i - \frac{1}{n} \sum_{i=1}^{n} x_i^2\right] = \bar{y} \frac{1}{n} \sum_{i=1}^{n} x_i - \frac{1}{n} \sum_{i=1}^{n} y_i x_i \tag{20}$$

$$\hat{\beta}_1 \left[\bar{x}^2 - \frac{1}{n} \sum_{i=1}^{n} x_i^2\right] = \bar{y}\bar{x} - \frac{1}{n} \sum_{i=1}^{n} y_i x_i \tag{21}$$

Simplifying all arithmetic means and isolating $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\frac{1}{n} \sum_{i=1}^{n} y_i x_i - \bar{y}\bar{x}}{\frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2} \tag{22}$$

Given the definitions of variance and covariance for a random variable:

$$Var(X) = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \tag{23}$$

$$Cov(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right) \left(Y_i - \bar{Y}\right) \tag{24}$$

$$\tag{25}$$

The, the expression for the slope coefficient of our linear regression model can be easily derived as the ratio between the covariance of the two variables and the variance of the independent variable:

$$\hat{\beta}_1 = \frac{Cov(x, y)}{Var(x)} \tag{26}$$

1.2 The R-Squared Coefficient

Since the Ordinary Least Squares algorithm is based on minimizing the Residuals Sum of Squares, much of the interpretation of the statistical model rests on the reasoning that the Sums of Squares mean us. The **Residuals Sum of Squares (RSS)** can be understood as the variation of the real phenomenon, which is captured only by the error term, in other words, everything that affects our variable of interest and that cannot be explained by our model. Therefore, what is captured by our model can be quantified and we will call it the **Estimation Sum of Squares (ESS)**, which is defined as follows:

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{27}$$

It is interpreted as the sum of the squares of the differences between the values of our estimates (made by the model) and the arithmetic mean of the dependent variable. The ESS helps us to understand how much of the variation in the observed phenomenon is explained by our model. It is easy to guess that if we add the RSS to the ESS, we will obtain the **Total Sum of Squares (TSS)**, which captures the total variation of the observed phenomenon.

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{28}$$

So how might we know how much of our phenomenon to be understood is explained by our simple linear regression model? To do this, we could use the R-Squared (R^2), which is an adjustment measure used in econometrics for the diagnosis and analysis of the results of a model.

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{29}$$

If we recall the definition of the variance of a random variable:

$$R^2 = \frac{Var(\hat{y})}{Var(y)} \tag{30}$$

This coefficient is defined as the ratio of the Explained Sum of Squares (ESS) to the Total Sum of Squares (TSS). It can only take values between zero and one, representing the percentage of the variation in the phenomenon that our model can explain. The closer R^2 is to zero, the less explanatory power the model has; conversely, the closer it is to one, the more explanatory power the regression model possesses. Since R^2 will always be less than or equal to one, we can conclude that the variance of the fitted values ($\hat{y}$) will always be less than or equal to the variance of the original variable.

$$Var(\hat{y}) < Var(y) \tag{31}$$

If we remember that the adjusted values ($\hat{y}$) are extracted from our linear regression model (without the residuals) we can obtain:

$$R^2 = \frac{\sum_{i=1}^{n} \left(\left[\hat{\beta}_0 + \hat{\beta}_1 x_i \right] - \bar{y} \right)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{32}$$

Now we can substitute the arithmetic mean of the dependent variable ($\bar{y}$) with the equation of the linear regression model in means (as we obtained it to derive $\hat{\beta}_0$):

$$R^2 = \frac{\sum_{i=1}^{n} \left(\left[\hat{\beta}_0 + \hat{\beta}_1 x_i \right] - \left[\hat{\beta}_0 + \hat{\beta}_1 \bar{x} \right] \right)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{33}$$

Simplifying the numerator of the equation and taking out a common factor:

$$R^2 = \frac{\sum_{i=1}^{n} \left(\hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x} \right)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{34}$$

$$R^2 = \frac{\sum_{i=1}^{n} \left(\hat{\beta}_1 [x_i - \bar{x}] \right)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{35}$$

$$R^2 = \hat{\beta}_1^2 \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{36}$$

If we recall the alternative method for obtaining $\hat{\beta}_1$, which is the ratio between the covariance of the dependent and independent variables and the variance of the latter, we note that to calculate the variances of x and y, we divide by n. Substituting the numerator with the variance of x and the denominator with the variance of y, we obtain:

$$R^2 = \left(\frac{Cov(x,y)}{Var(x)} \right)^2 \left(\frac{Var(x)}{Var(y)} \right) \tag{37}$$

Finally, simplifying the expression, we obtain a new form of the R^2 coefficient.

$$R^2 = \frac{Cov(x,y)^2}{Var(x)Var(y)} = \frac{\gamma_{xy}^2}{\sigma_x^2 \sigma_y^2} \tag{38}$$

From this equation and after taking squeared-root, we can recognize a well-known statistic: Pearson's Correlation Coefficient.

$$R = \frac{\gamma_{xy}}{\sigma_x \sigma_y} = \rho \tag{39}$$

In Module 1, we discussed that Pearson's correlation coefficient measures both the strength and direction of the linear relationship between two random variables, with a range from -1 to 1. A value close to 1 indicates a strong positive correlation, while a value close to -1 suggests a strong negative correlation. In the context of a simple linear regression model, this coefficient is essential for understanding the preliminary relationship between the variables and for formulating a grounded hypothesis. This hypothesis can then be validated and explored further through the regression model.

2 Multiple Regression Model

Not always will be the case that a linear regression model is only formulated with one independent variable, but many of them. We can propose an equation with multiple independent variables in the following form:

$$y_i = \beta_0 + \sum_{k=1}^{n} \beta_k x_{ki} + \varepsilon_i \tag{40}$$

Where k is the number of dependent variables we want to include in our model specification. Technically, we use matrices to explain and calculate the multiple regression model, so we can rewrite the equation as the following form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{41}$$

Or in the classical matrix form:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The vector $\mathbf{Y}$ has a dimension of $n \times 1$ and contains all the observations of the dependent variable. These are the values we want to predict with our regression model. The matrix $\mathbf{X}$ contains all the observations for all the independent variables we are including in our model, this is why it has a dimension of $n \times k$, where n is the sample size and k the number of regressors. The vector β contains all the coefficients we are trying to estimate, and it has a dimension of $k \times 1$. The product $\mathbf{X}\beta$ must have a $n \times 1$ dimension to be added to the residuals vector (ε), which has the same dimension.

2.1 Coefficients Estimation

Now, to estimate the coefficients we must clear up the residuals vector:

$$\varepsilon = \mathbf{Y} - \mathbf{X}\beta \tag{42}$$

To square a vector, we usually multiply by its transposed form, this means that:

$$\varepsilon^\top \varepsilon = (\mathbf{Y} - \mathbf{X}\beta)^\top (\mathbf{Y} - \mathbf{X}\beta) \tag{43}$$

The number $\varepsilon^\top \varepsilon$ might be interpreted as the Residuals Sum of Squares, so we are going to minimize it to obtain the coefficients. First, we must expand the parenthesis, a squared binomial in linear algebra will have the next form:

$$\varepsilon^\top \varepsilon = \mathbf{Y}^\top \mathbf{Y} - 2\mathbf{X}^\top \mathbf{Y}\beta + \beta^\top \mathbf{X}^\top \mathbf{X}\beta \tag{44}$$

Then, just like the simple regression case, we have to derive the RSS, and then equalize to zero. We can cancel both twos and rearrange the equation obtaining:

$$\frac{\partial \left(\varepsilon^\top \varepsilon\right)}{\partial \beta} = 2\mathbf{X}^\top \mathbf{X}\beta - 2\mathbf{X}^\top \mathbf{Y} = 0 \tag{45}$$

To clean up for β we have to multiply both sides by the inverse matrix of $\mathbf{X}^\top \mathbf{X}$:

$$\beta = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \left(\mathbf{X}^\top \mathbf{Y}\right) \tag{46}$$

In econometric literature, the $\mathbf{X}$ will be known as the **Design Matrix** (or the Information Matrix) which contains all the data needed to predict the phenomena of study. Then $\mathbf{X}^\top \mathbf{X}$ shall be interpreted as the square of the information matrix, containing all the variability of the regressors. By some equivalence, we might conclude that $\mathbf{X}^\top \mathbf{X}$ approximates the denominator of the coefficient (β_1) of a simple regression

model. The Design Matrix should always contain a column of 1 to capture the effect of the constant on the linear equation ($beta_0$). Let's analyze the properties of $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{X}^\top\mathbf{Y}$:

$$X^\top X = \begin{bmatrix} n & \sum X_{1i} & \cdots & \sum X_{ki} \\ \sum X_{1i} & \sum X_{1i}^2 & \cdots & \sum X_{1i}X_{ki} \\ \vdots & \vdots & \ddots & \vdots \\ \sum X_{ki} & \sum X_{1i}X_{k1} & \cdots & \sum X_{k1}^2 \end{bmatrix}$$

The matrix $\mathbf{X}^\top\mathbf{X}$ is symmetric with a dimension of $k \times k$, meaning that the elements above the diagonal are identical to those below it. The diagonal entries contain the sums of the squares of the regressors, which provide an approximation of the variability of each independent variable (recalling the form of the variance). We can interpret the interaction terms as the shared variability between X_m and X_k. Ideally, we would like the regressors to be uncorrelated, though this only happens in the case of perfect independence.

$$X^\top Y = \begin{bmatrix} \sum Y_i \\ \sum X_{i1}Y_i \\ \vdots \\ \sum X_{ik}Y_i \end{bmatrix}$$

Now, the vector $\mathbf{X}^\top\mathbf{Y}$, with dimensions $k \times 1$, contains the shared relationship between the dependent variable and the regressors in the linear equation. This vector plays a key role in determining the numerator of the formula for β_1 in a simple linear regression model. The scalars in the vector reflect the degree of covariation between the dependent variable Y and any regressor X_k offering insight into how changes in the independent variables are associated with changes in the dependent variable. This explains why it has a direct relationship with the estimated coefficients.

2.2 Fitted Values and the Hat Matrix

The predicted values of the dependent variable ($\hat{\mathbf{Y}}$), will then have the next form:

$$\hat{\mathbf{Y}} = \mathbf{X}\beta \tag{47}$$

Where $\hat{\mathbf{Y}}$ is a vector with dimensions $n \times 1$ containing all the fitted values of our regression model. But by remembering the form of β:

$$\hat{\mathbf{Y}} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y} \tag{48}$$

Then we can note that there is a "transformer" of $\mathbf{Y}$ that turns it into the predicted values ($\hat{\mathbf{Y}}$). Let us call this transformer the **Hat Matrix**:

$$\hat{\mathbf{H}} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top \tag{49}$$

Then, using the Hat Matrix we can also calculate the residuals of the model by transforming the real values of $\mathbf{Y}$:

$$\varepsilon = \mathbf{Y} - \hat{\mathbf{Y}} \tag{50}$$

$$\varepsilon = \mathbf{Y} - \hat{\mathbf{H}}\mathbf{Y} \tag{51}$$

$$\varepsilon = \left(\mathbf{1} - \hat{\mathbf{H}}\right)\mathbf{Y} \tag{52}$$

The Hat Matrix is a powerful tool in econometrics, as it helps us to 'separate' $\mathbf{Y}$ into its deterministic part $\hat{\mathbf{y}}$ and its stochastic component ε. The Hat Matrix has dimensions $n \times n$ and exhibits several interesting characteristics that aid in the understanding of the linear regression model. The Hat Matrix is symmetric and idempotent, meaning that multiplying this matrix by itself does not alter it. But one of the most important properties is the orthogonality condition:

$$\hat{\mathbf{H}}\varepsilon = 0 \tag{53}$$

This condition confirms that the fitted values of the dependent variable are orthogonal to or independent of the error term (as proven in the previous section). This means that the error term or the residuals do not contain any additional information relevant to explaining the phenomenon under study. In other words, there are no specification errors, no bias in the estimated coefficients, and the other OLS assumptions are satisfied.

2.3 The Adjusted R-Squared

The R-Squared (R^2) coefficient can also be estimated in the context of multiple linear regression models, the calculation of the Sums of Squared are practically the same, but in the matrix notation we can use the next formulas:

$$RSS = \varepsilon^\top \varepsilon = \mathbf{Y}^\top \mathbf{Y} - \beta \mathbf{X}^\top \mathbf{Y} \tag{54}$$

$$ESS = \hat{\mathbf{Y}}^\top \hat{\mathbf{Y}} - \frac{1}{n}\left(\mathbf{N}^\top \mathbf{Y}\right)^2 \tag{55}$$

$$TSS = \mathbf{Y}^\top \mathbf{Y} - \frac{1}{n}\left(\mathbf{N}^\top \mathbf{Y}\right)^2 \tag{56}$$

Being $\mathbf{N}$ a vector of ones $n \times 1$, the R-square will then have the same classic form since all the sums of squares are scalars (only one value). Nevertheless, in the context of a multivariate regression model, just adding independent variables will inflate the R^2 giving us a wrong impression of better specifications, this is why we have to 'punish' the R-Squared if some useless variables are added. This is why we introduce the Adjusted R-Squared (R^2_{Adj}):

$$R^2_{Adj} = 1 - \left(\frac{RSS}{TSS} * \frac{n-1}{n-k+1}\right) \tag{57}$$

By adding more regressors, we penalize R^2_{Adj}, due to its negative relationship with k, the number of estimated coefficients. Furthermore, if these variables are not relevant in explaining $\mathbf{Y}$, RSS will also increase, causing R^2_{Adj} to fall as well. Therefore, it is preferable to analyze the adjusted version of R^2 rather than the original form, as it provides a more accurate measure of the quality of the model. However, it's important to note that in a simple regression model with just 2 coefficients, both versions of R^2_{Adj} will be identical.

2.4 Testing Significance

When deriving the β coefficients with the Ordinary Least Squares Algorithm it is truly relevant to diagnose them for statistical significance. This means that we want to assess the coefficients to be statistically different from zero, concluding that their respective independent variables are relevant to explain the phenomena we are studying. Then, **hypothesis testing** is a fundamental tool at evaluating the quality of an econometric model. Technically speaking, we use robust statistical evidence to assess and prove the hypothesis we are proposing, those being:

$$H_0 \colon \hat{\beta}_j = 0$$
$$H_1 \colon \hat{\beta}_j \neq 0$$

We understand H_0 as the null hypothesis, the base case in the hypothesis testing; this is going to be the hypothesis we want to prove as false since it implies that the coefficients derived from the OLS algorithm are not zero, which means that the independent variables are irrelevant to comprehend the phenomenon of study. Therefore, H_1 is the alternative hypothesis, which is the case we want to prove, and help us to conclude that the linear model we are proposing is the best one to predict the variable $\mathbf{Y}$. To do so, we must calculate the t-Student Scores with the form:

$$t_j = \frac{\hat{\beta}_j}{\text{SSE}(\hat{\beta}_j)} \tag{58}$$

To do this, we must obtain the covariance matrix — an array that contains the variances of the coefficients and the covariances among them. The covariance matrix takes the following form:

$$\mathbf{C} = \frac{RSS}{n-k}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \tag{59}$$

The first part of the formula, $RSS/(n-k)$, is the variance of the errors (σ^2), while the second part is the inverse of the square of the design matrix.

$$\mathbf{C} = \begin{bmatrix} \text{Var}(\hat{\beta}_0) & \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \cdots & \text{Cov}(\hat{\beta}_0, \hat{\beta}_k) \\ \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \text{Var}(\hat{\beta}_1) & \cdots & \text{Cov}(\hat{\beta}_1, \hat{\beta}_k) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(\hat{\beta}_0, \hat{\beta}_k) & \text{Cov}(\hat{\beta}_1, \hat{\beta}_k) & \cdots & \text{Var}(\hat{\beta}_k) \end{bmatrix}$$

Note the equivalence of the previous equation with the formula of the variance of the $\hat{\beta}_1$ in the previous chapter. Not that the diagonal of the Covariance Matrix contains all the variances of the coefficients. Now let us build a matrix where its diagonal will be the square root of the diagonal of C (the standard errors), and the rest of values will be 0:

$$\mathbf{D} = \begin{bmatrix} \text{SSE}(\hat{\beta}_0) & 0 & \cdots & 0 \\ 0 & \text{SSE}(\hat{\beta}_1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \text{SSE}(\hat{\beta}_k) \end{bmatrix}$$

Then to obtain the t-Student Scores for the hypothesis testing, we can multiply the inverse of D and the vector β:

$$\mathbf{T} = \mathbf{D^{-1}}\beta \tag{60}$$

Since $\mathbf{D}$ has a dimension of $k \times k$ and the vector β has a dimension of $k \times 1$, the vector $\mathbf{T}$ will be $k \times 1$ too.

$$\mathbf{T} = \begin{bmatrix} t(\hat{\beta}_0) \\ t(\hat{\beta}_1) \\ \vdots \\ t(\hat{\beta}_k) \end{bmatrix}$$

The t-student distribution is a good approximation of the normal distribution, being remarkably close to it when the degrees of freedom are sufficiently large. Then we can build the confidence intervals for the hypothesis testing:

$$\Pr\left[-t_{\alpha/2} \leq \frac{\hat{\beta}_j - \beta_j}{\text{SSE}(\hat{\beta}_j)} \leq t_{\alpha/2}\right] = 1 - \alpha \tag{61}$$

By rearranging the equation, we can find:

$$\Pr\left[\hat{\beta}_j - t_{\alpha/2} * \text{SSE}(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2} * \text{SSE}(\hat{\beta}_j)\right] = 1 - \alpha \tag{62}$$

We can interpret this interval as the probability that the true coefficient β lies within the bounds $\hat{\beta}_j \pm t_{\alpha/2} \cdot \text{SSE}(\hat{\beta}_j)$ is $1-\alpha$, where α is the significance level and $1-\alpha$ is the confidence level. This means that, in a large number of samples, we expect the confidence interval to contain the true value of the coefficient β in $(1-\alpha)\%$ of those cases. Furthermore, if zero is not included in this interval, we can conclude that there is sufficient evidence to assert that $\hat{\beta}_j$ is a statistically significant approximation of the true coefficient β. This highlights the robustness of our conclusion regarding the relationship between the independent variable and the dependent variable in the context of the regression model.

The values $t_{\alpha/2}$ are interpreted as the critical values of the distribution, so the probability of being outside the boundaries is α. However, since we are using a two-tailed test, the significance level is divided into two. When the degrees of freedom are extremely large, these critical values approximate to ± 1.96. Also, traditionally, statisticians use a level of significance of 5% as a consensus, but in econometric literature we will find hypothesis testing with different α values (such as 10% or 1%). Then we can rewrite the previous equation as:

$$\Pr\left[\hat{\beta}_j - 1.96 * \text{SSE}(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + 1.96 * \text{SSE}(\hat{\beta}_j)\right] = 0.95 \tag{63}$$

Being the case, for each value of $|t| > t_{\alpha/2}$ (previously defined), we reject the null hypothesis (H_0), concluding that the $\hat{\beta}_j$ we are testing is statistically significantly different from zero with a confidence of 95%, ensuring that the variable X_j is relevant to understanding $\mathbf{Y}$. We can also obtain the *p-value* (probability) for a certain value of t. For a two-tailed test, we can use:

$$\text{p-value} = 2 \cdot P\left(T > |t_j|\right) = 2 \int_{|t_j|}^{\infty} f_T(x)\, dx \tag{64}$$

If the *p-value* is smaller than the significance level (α), we can reject the null hypothesis. This implies that there is sufficient statistical evidence to conclude that the observed data is unlikely under the null hypothesis. For instance, with a significance level of 0.05, we expect to observe a *p-value* < 0.05. The function $f_T(x)$ is the probability density function of the Student's t-distribution. By integrating this function, we obtain the probability that the t-statistic falls within the rejection area of the null hypothesis. This integration ensures that our conclusions are robust and grounded in the appropriate distribution of our data.

2.5 Joint Significance

For the case of multiple lineal regression (when there is more than one regressor), we also want to assess the joint significance of the model. This means that we want to prove that, jointly our model is useful to predict the variable $\mathbf{Y}$. Then we want to prove the alternative hypothesis that establishes at least one of the $\hat{\beta}_j$ coefficients are statistically different from zero, contrasting with the null hypothesis that sets all the $\hat{\beta}_j$ are zero:

$$H_0 \colon \hat{\beta}_0 = \hat{\beta}_1 = \hat{\beta}_2 = \cdots = \hat{\beta}_k = 0$$
$$H_1 \colon \text{At least one } \hat{\beta}_j \neq 0$$

So, to prove the alternative hypothesis we are going to build the F-Score with the formula:

$$F = \frac{\frac{ESS}{k-1}}{\frac{RSS}{n-k}} \tag{65}$$

With $k-1$ degrees of freedom on the numerator and $n-k$ degrees of freedom on the denominator. The F-Score is distributed with the F-Distribution; we also evaluate the probability of our F-Score to be greater than the critical values, given its own distribution, or our p-value to be smaller than the significance level of 5%. We can rearrange the previous formula as:

$$F = \frac{ESS}{RSS}\frac{n-k}{k-1} \tag{66}$$

The greater ESS and the lower the RSS (meaning a higher R- Squared) and a big enough sample size (n) with a reasonable number of regressors (k), we are going to obtain an F-score sufficiently large to reject the hypothesis null with certain confidence. Nevertheless, there is nothing wrong with obtaining the p-value to be sure.

3 Regression Diagnosis

Mathematicians, economists, and data scientists have developed numerous tests and mathematical proofs to help us diagnose the asymptotic conditions of our regression model. To ensure that our linear regression model is of the highest possible quality, we must test whether all the assumptions of linear regression hold, including homoscedasticity, no serial correlation, and normality of the errors. Now, we will study some of the most well-known tests for heteroskedasticity, multicollinearity, autocorrelation, normality and linearity.

3.1 Heteroskedasticity

The fourth fundamental assumption for the development of linear regression by ordinary least squares is that of homoskedasticity in errors. By this we mean that the conditional variance of the errors (given a value of our independent variable) will be constant or independent of the same values of x. Given the fulfillment of the homoskedasticity condition, we will always call the variance of errors as: σ^2.

3.1.1 Breusch-Pagan Test

The Breusch-Pagan test seeks to determine whether the variance of the errors is a function of the independent variables in our linear regression. In mathematical terms, we aim to test if this holds true:

$$\sigma_i^2 = \sigma^2 h(x_i) \tag{67}$$

We can assume that the function $h(x_i)$ is a lineal combination of parameters to be estimated and the independent variables involved in the original econometric model. In this way, we can redefine the form of the heteroskedastic variance as:

$$\sigma_i^2 = \sigma^2 \left[\alpha_0 + \sum_{j=1}^{J} \alpha_j X_{ji} \right] \tag{68}$$

There will be heteroskedasticity if at least one of the α_j coefficients are different from zero. So, to develop the test, first of all we have to remember our original regression model, it can be a bi-variate model or a multivariate model:

$$y_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j X_{ji} + \hat{\varepsilon}_i \tag{69}$$

From this equation, we can obtain the residuals $\hat{\varepsilon}$ and square them to propose a new auxiliary regression model in the form:

$$\hat{\varepsilon}_i^2 = \hat{\gamma}_0 + \sum_{j=1}^{k} \hat{\gamma}_j X_{ji} + \hat{\theta}_i \tag{70}$$

Then, what we want to evaluate is whether this model is relevant for explaining the squared errors of the original regression. If this is the case, we can conclude that the variance of the errors is not constant, or there is evidence of heteroskedasticity. If this happens, the variance of the estimated coefficients in our model will not be the minimum possible, and the hypothesis testing for statistical significance will not be as reliable as we would like. The hypothesis we want to test is as follows:

$$H_0 \colon \sigma^2 = h(x)$$
$$H_1 \colon \sigma^2 \neq h(x)$$

How can we test these hypotheses? There is a large body of literature on the topic, but no consensus on the best approach, as the goal is to assess whether the model helps explain the squared residuals. The following table presents some common methods for calculating the test statistic.

Table 1: Breusch-Pagan Test

Statistic	**Form**	**Distribution**	**Degrees of Freedom (df)**
F-Statistic (Wooldridge)	$F = \dfrac{\text{ESS}}{\text{RSS}} \cdot \dfrac{n-k}{k-1}$	$F \sim F_{k-1,\ n-k}$	$df_1 = k-1$, $df_2 = n-k$
LM-Statistic (Greene)	$LM = nR^2$	$LM \sim \chi^2_k$	$df = k$

The objective is not to reject the null hypothesis, or to conclude that the variance is homoscedastic. Therefore, we want these statistics to be as close to zero as possible. If the auxiliary model is not relevant for explaining the variance of the errors, then the Estimated Sum of Squares (ESS) will be smaller and closer to zero. As a result, the three statistics will also be smaller. Since these three tests are based on the ESS, we can be indifferent (to some extent) in choosing among them. However, a good scientist would not be satisfied until they have gathered all the possible evidence.

3.1.2 White Test

Similar to the Breusch-Pagan test, the White test attempts to determine whether the variance of the errors is a function of all the independent variables used in the original model. However, unlike the previous test, the White test is more flexible regarding the form of the function for the variance. The BP test suggests a linear equation for the auxiliary model, whereas the White test allows us to introduce various monotonic transformations of the explanatory variables, such as quadratic or logarithmic transformations. Let us define the auxiliar model with the following form of a quadratic equation (which is just an example):

$$\hat{\varepsilon}_i^2 = \hat{\gamma}_0 + \sum_{j=1}^{k} \hat{\gamma}_j X_{ji} + \sum_{j=1}^{k} \hat{\delta}_j X_{ji}^2 + \hat{\theta}_i \tag{71}$$

The hypothesis we are assessing will be similar, but not exactly the same, since we are now proposing a non-linear function instead. However, the reasoning behind the White test is very similar. Why are we using transformations of X? Because they can help us identify correlations that a simple linear model might miss. This model would be better specified, and therefore, it has a higher probability of uncovering hidden relationships between the variance of the errors and the covariates.

$$H_0 \colon \sigma^2 = h(x, x^2, \ldots, x^p)$$
$$H_1 \colon \sigma^2 \neq h(x, x^2, \ldots, x^p)$$

Naturally, since this is not a linear equation, we would want to use appropriate statistics. Econometricians suggest that for non-linear equations, we can use the Lagrange Multiplier approximations, as these tests are based on the maximization of the Log-Likelihood function subject to the natural restrictions of our model (in this case, the heteroskedasticity condition). Therefore, we will use the LM statistic, like the Breusch-Pagan test.

Table 2: White Test

Statistic	**Form**	**Distribution**	**Degrees of Freedom (df)**
LM-Statistic (Greene)	$LM = nR^2$	$LM \sim \chi_k^2$	$df = k$

Note that if the auxiliary regression is not robust enough to explain the variance, the R^2 will be very close to zero, reducing the size of the LM statistic and providing evidence of low or no heteroskedasticity in our model. The disadvantage is that for models with few variables and very large samples, even when the R^2 is very low, we could obtain a sufficiently large LM statistic to reject the null hypothesis, making models with large samples more sensitive to conclusions about non-constant variance.

3.1.3 Goldfeld-Quandt Test

The Goldfeld-Quandt test is somewhat more intuitive than the previous two tests. The purpose is to determine whether the variance of two different samples of the residuals is statistically different, and thereby conclude whether the model suffers from heteroskedasticity. The first step is to calculate the original model, just as in the previous two tests, and then obtain the residuals. Then we must select two random subsamples and obtain the variance of the errors:

$$\hat{\sigma}_1^2 = \frac{1}{n_1 - k} \sum_{i=1}^{n_1} \hat{\varepsilon}_i^2 \quad \Big| \quad \hat{\sigma}_2^2 = \frac{1}{n_2 - k} \sum_{i=1}^{n_2} \hat{\varepsilon}_i^2 \tag{72}$$

Not necessarily must both samples be the same size; the objective is to determine whether the model shows evidence of non-constant variance, and this can be assessed using so many different subsamples from the same model; but indeed it is necessary that both n_1 and n_2 to be greater than the number of estimated coefficients k. The problem is framed to assess the following null and alternative hypotheses:

$$H_0 \colon \sigma_1^2 = \sigma_2^2$$
$$H_1 \colon \sigma_1^2 \neq \sigma_2^2$$

To decide whether to reject the null hypothesis, we must calculate the F-statistic, which follows the F-distribution. Next, we calculate the critical F-value for a significance level of $\frac{\alpha}{2}$ (since it is a two-tailed proof), with $n_1 - k$ degrees of freedom in the numerator and $n_2 - k$ degrees of freedom in the denominator, rejecting the null hypothesis if our F-statistic is greater than the critical value.

Table 3: Goldfeld-Quandt Test

Statistic	**Form**	**Distribution**	**Degrees of Freedom (df)**
F-Statistic	$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$	$F \sim F_{n_1-k, n_2-k}$	$df_1 = n_1 - k$, $df_2 = n_2 - k$

We use a two-tailed test because the hypothesis determines whether the variances are different. If $\hat{\sigma}_1^2 > \hat{\sigma}_2^2$, the F-statistic will be greater than zero and approach values larger than the critical value. Conversely, if $\hat{\sigma}_1^2 < \hat{\sigma}_2^2$, the F-statistic will be small, closer to zero. In both cases, we reject the null hypothesis and conclude the presence of heteroskedasticity.

One of the main disadvantages of the Goldfeld-Quandt test is that some samples might share the same variance, even in the presence of general heteroskedasticity. A more appropriate approach would be to repeat the test several times and assess how many of those samples exhibit different variances. For this reason, other methods are often preferred to assess non-constant variances.

3.2 Multicollinearity

One of the most important assumptions for a Linear Regression Model is the absence of perfect multicollinearity among the covariates or independent variables of the model. Multicollinearity refers to a situation in linear regression where there is a linear relationship among the explanatory variables. In mathematical terms, this means that one variable can be expressed as a linear combination of the others:

$$X_k = f\left(X_1, X_2, ..., X_{k-1}\right) \tag{73}$$

3.2.1 Variance Inflation Factor

Since multicollinearity has negative effects on our linear regression model, making our estimators inefficient and even rendering the model incalculable due to perfect collinearity. Then, how can we assess the extent of inefficiency in our estimators? To do this, we use the Variance Inflation Factor (VIF), a coefficient that quantifies how much the variances of our estimators are 'inflated' due to multicollinearity. First, recall the form of the variance of a coefficient:

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\text{Var}(X_j)} \tag{74}$$

This form is only an approximation, as this expression applies specifically to a simple linear regression model, as we saw in the previous section. However, it helps us to understand the main idea behind this diagnostic. The Variance Inflation Factor (VIF) can then be understood as the inverse of the complement of the R-Squared (R^2). The mathematical expression is as follows:

$$\text{VIF}(\hat{\beta}_j) = \frac{1}{1 - R_J^2} \tag{75}$$

But what does R_J^2 represent? This is the coefficient of determination from a regression where X_j is the dependent variable and the other covariates are the explanatory variables. This regression will take the following form:

$$X_j = \alpha_0 + \sum_{\substack{i=1 \\ j \neq i}}^{k} \alpha_i X_i + v \tag{76}$$

Then, the adjusted form of the variance of an estimated coefficient in the presence of multicollinearity will be:

$$\text{Var}(\hat{\beta}_j)_{VIF} = \frac{\sigma^2}{\text{Var}(X_j)} * \frac{1}{1 - R_J^2} \tag{77}$$

Since the R-Squared will be always less than one and greater than zero $0 < R_J^2 < 1$ (unless we are on the case of perfect correlation), the variance of the coefficients adjusted by the VIF will be always greater than the original form. Concluding that multicollinearity will always make our estimators inefficient, and the conclusions of the hypothesis tests will not be as robust as the case of non-collinearity.

3.3 Autocorrelation

In time series econometrics, autocorrelation (also known as serial correlation) is an important issue that must be addressed. The presence of autocorrelation means that a variable is correlated with its own past values. This can lead to serious problems in the estimation of coefficients, specifically making them less reliable because their variances are no longer minimized. Therefore, it is crucial to detect and correct for autocorrelation.

3.3.1 Durbin-Watson Test

Now, let us discuss other tests that can be used to detect autocorrelation. The Durbin-Watson test is a statistical method specifically designed to diagnose serial correlation in the error terms of a linear regression model. However, it can also be applied to time series data after detrending or demeaning the series. The Durbin-Watson test provides a more limited perspective on autocorrelation, as it evaluates correlation only at the first lag.

$$DW = \frac{\sum_{t=1}^{T} (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{T} \hat{\varepsilon}_t^2} \tag{78}$$

If we simplify the parentheses in the numerator, we can derive a new expression involving the variances of the errors. First, we must recall that if the errors have a zero mean, and since the variable and its lags share the same variance, we can then obtain:

$$DW = \frac{2\sigma^2 - 2\gamma_1}{\sigma^2} \tag{79}$$

Then, after simplifying the equation and recalling the form of the correlation coefficient, we can arrive at an approximation of the DW-Score:

$$DW = 2(1 - \rho_1) \tag{80}$$

The Durbin-Watson test does not follow a formal hypothesis-testing framework. Instead, we intuitively assess the presence of autocorrelation based on the value of the test statistics. For instance, a value close to 2 suggests no evidence of first-order autocorrelation in the residuals. Conversely, a value near 0 indicates evidence of positive autocorrelation, while a value close to 4 suggests the potential existence of negative autocorrelation.

3.3.2 Breusch-Godfrey Test

Another notable test for detecting autocorrelation in the residuals of a linear regression model is the Breusch-Godfrey test. This test diagnoses serial correlation of one or more orders by evaluating the statistical significance of the coefficients in an auxiliary regression, where the residuals are regressed on their own lags. Mathematically, for a linear model with k regressors, the test proposes an alternative model of the following form:

$$\hat{\varepsilon}_t = \alpha_0 + \sum_{k=1}^{K} \beta_j X jt + \sum_{h=1}^{H} \rho_l \hat{\varepsilon}_{t-l} + e_t \tag{81}$$

The ρ_h coefficients represent the correlation coefficients between the errors and their h^{th} lag. Naturally, we aim to evaluate the statistical significance of these coefficients both individually and jointly. The hypotheses to be tested are as follows:

$$H_0 \colon \rho_0 = \rho_1 = \cdots = \rho_H = 0$$
$$H_1 \colon \text{At least one } \rho_h \neq 0$$

To test the null hypothesis, we use the Lagrange Multiplier (LM) score. This score follows a χ^2 distribution with H degrees of freedom. Recall that we reject the null hypothesis if the LM score exceeds the critical value from the χ^2 distribution:

Table 4: Breusch-Godfrey Test

Statistic	**Form**	**Distribution**	**Degrees of Freedom (df)**
LM-Statistic (Greene)	$LM = nR^2$	$LM \sim \chi^2_H$	$df = H$

Remember that the R^2 coefficient from the auxiliary regression model is used to compute the LM score. The limitation of this approach is that it requires the residuals from a regression model to test for autocorrelation. Therefore, this type of test can be applied to time series only after detrending them using OLS.

3.4 Normality

Another essential test for the linear regression model is assessing whether the residuals are normally distributed. Although the assumptions of zero mean and constant variance of the errors support the approximation of normality, additional conditions must be met to ensure that the residuals are truly normally distributed. Mathematically, this can be expressed as:

$$\hat{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2) \tag{82}$$

3.4.1 Jarque-Bera Test

This test is based on constructing a statistic that considers the shape of the distribution, specifically focusing on its skewness and kurtosis. Since it is straightforward to verify whether the arithmetic mean of the residuals is zero, and other tests can assess constant variance, the Jarque-Bera test specifically examines the higher moments of the distribution. Mathematically, the Jarque-Bera score is:

$$JB = \frac{n}{6}\left(\xi^2 + \frac{(\kappa - 3)^2}{4}\right) \tag{83}$$

Evidently, if the residuals are perfectly normally distributed, the skewness will be zero ($\xi = 0$) and the kurtosis will be three ($\kappa = 3$), as we demonstrated in the previous section. In this case, the Jarque-Bera score will be exactly zero. The score follows a χ^2 distribution with two degrees of freedom, and the hypothesis being tested is:

$$H_0 \colon \hat{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2)$$
$$H_1 \colon \hat{\varepsilon}_i \not\sim \mathcal{N}(0, \sigma^2)$$

Table 5: Jarque-Bera Test

Statistic	**Form**	**Distribution**	**Degrees of Freedom (df)**
JB-Statistic	$JB = \frac{n}{6}\left(\xi^2 + \frac{(\kappa-3)^2}{4}\right)$	$JB \sim \chi^2_2$	$df = 2$

For distributions with high skewness (either negative or positive) or with kurtosis values different from three (leptokurtic or platykurtic distributions), there is a higher likelihood of rejecting the null hypothesis at a given α significance level. This is because the Jarque-Bera score will be greater than zero and closer to the critical value $\chi^2_{0.05} = 5.991$. However, this test is not always used exclusively for diagnosing linear regression models. Some econometricians and statisticians also apply the Jarque-Bera test to determine if the dependent variable and the explanatory variables are normally distributed. This is done to prevent spurious correlations that can arise when building a model with non-normally distributed variables.

3.5 Linearity

When testing for linearity, we are assessing whether the specification of our regression model follows a linear pattern. In other words, the relationship between the dependent variable and the explanatory variables should be strictly linear—not quadratic, cubic, or of any other form. Nevertheless, it is important to note that we can also estimate parabolic (or other nonlinear) relationships using a linear regression model by including quadratic (or other functional) transformations of a variable as additional regressors.

3.5.1 Ramsey's Regression Specification Error Test

The Ramsey test is commonly used to determine whether a regression model is correctly specified using only linear terms, or whether additional transformations of the fitted values should be included. In other words, the test involves estimating a regression in which transformations of $\hat{y}$ (the fitted values) are included as additional regressors.

$$y = \beta_0 + \sum_{k=1}^{K} \beta_k X_k + \sum_{p=2}^{P} \gamma_p \hat{y}^p + u \tag{84}$$

Given the definition of the fitted values $\hat{y} = \mathbf{X}\beta$, the Ramsey test not only evaluates the linear relationship between the independent variables and the dependent variable, but also implicitly tests the adequacy of a linear specification in terms of the coefficients β. The natural hypothesis, then, is to test the significance of the additional coefficients (γ) in the augmented model:

$$H_0 \colon \gamma_0 = \gamma_1 = \cdots = \gamma_P = 0$$
$$H_1 \colon \text{At least one } \gamma_p \neq 0$$

The test is based on the calculation of the F-statistic, which uses the R-squared values from both the original regression (also called the restricted regression, R^2_r) and the augmented regression that includes

the nonlinear terms (R_a^2), with k_a coefficients to estimate. In this context, the number of restrictions (q) corresponds to the number of nonlinear transformations being tested. For example, if we are testing only for a quadratic term (i.e., a parabolic specification), then the number of restrictions is one.:

$$F = \frac{R_a^2 - R_r^2}{1 - R_a^2} * \frac{n - k_a}{q} \tag{85}$$

Naturally, the F-statistic follows an F-distribution—just like in the joint significance test—with q degrees of freedom in the numerator and $n - k_a$ degrees of freedom in the denominator. As you can observe, since polynomial functions can approximate a wide range of functional forms (as in a Taylor series expansion), increasing the number of restrictions q typically leads to a smaller F-statistic and a lower probability of rejecting the null hypothesis. This is because the denominator grows faster relative to the gain in explanatory power, making it harder to detect model misspecification when too many nonlinear terms are included.

Table 6: Ramsey's Regression Specification Error Test

Statistic	**Form**	**Distribution**	**Degrees of Freedom (df)**
F-Statistic	$F = \frac{R_a^2 - R_r^2}{1 - R_a^2} * \frac{n - k_a}{q}$	$F \sim \mathcal{F}(q, n - k_a)$	$df_1 = q, df_2 = n - k_a$