Lecture 11

Problem Formulation of Linear Regression

Linear regression is probably one of the most important and widely used regression techniques. It's among the simplest regression methods. One of its main advantages is the ease of interpreting results.

Problem Formulation

When implementing linear regression of some dependent variable y on the set of independent variables $\mathbf{x} = (x_1, \ldots, x_r)$, where r is the number of predictors, you assume a linear relationship between y and $\mathbf{x}$: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + \varepsilon$. This equation is the **regression equation**. β_0, β_1, …, β_r are the **regression coefficients**, and ε is the **random error**.

Linear regression calculates the **estimators** of the regression coefficients or simply the **predicted weights**, denoted with b_0, b_1, …, b_r. They define the **estimated regression function** $(\mathbf{x}) = b_0 + b_1 x_1 + \cdots + b_r x_r$. This function should capture the dependencies between the inputs and output sufficiently well.

The **estimated** or **predicted response**, $(\mathbf{x}_i)$, for each observation $i = 1, \ldots, n$, should be as close as possible to the corresponding **actual response** y_i. The differences $y_i - (\mathbf{x}_i)$ for all observations $i = 1, \ldots, n$, are called the **residuals**. Regression is about determining the **best predicted weights**, that is the weights corresponding to the smallest residuals.

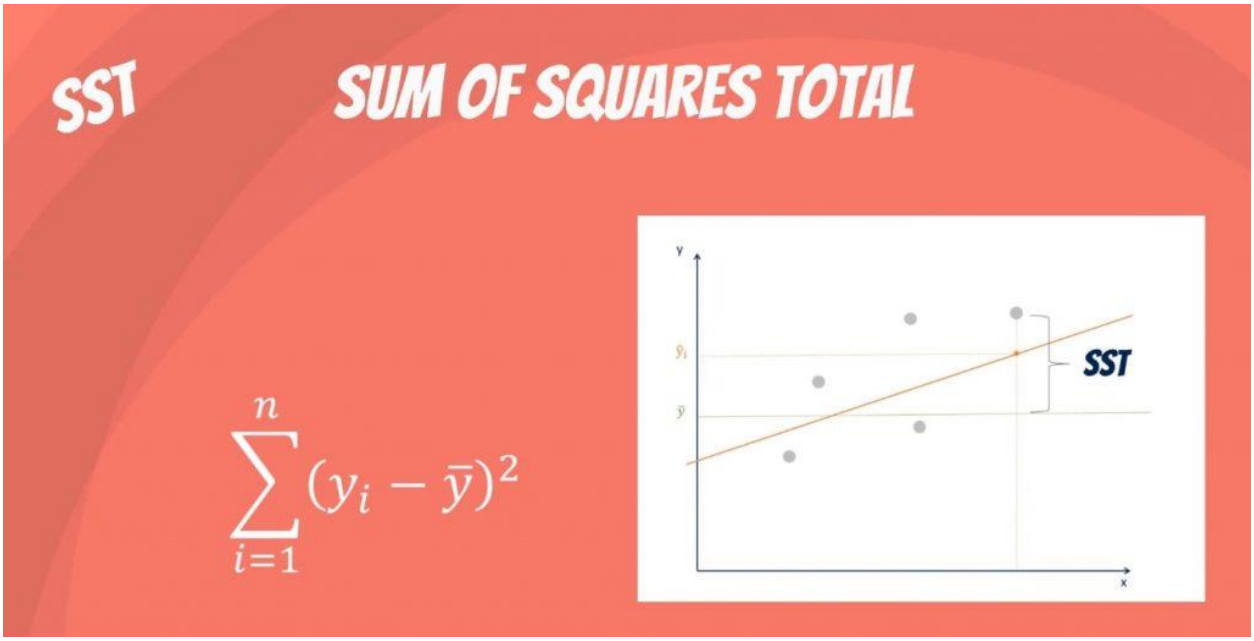
To get the best weights, you usually **minimize the residuals sum of squared** (RSS) for all observations $i = 1, \ldots, n$: $\text{RSS} = \Sigma_i (y_i - f(\mathbf{x}_i))^2$. This approach is called the **method of ordinary least squares**.

SST, SSR, SSE: Definition and Formulas

There are three terms we must define. The **sum of squares total**, the **sum of squares regression**, and the **sum of squares error**.

What is the SST?

The **sum of squares total**, denoted **SST**, is the squared differences between the observed *dependent variable* and its **mean**. You can think of this as the dispersion of the observed variables around the **mean** – much like the **variance** in descriptive statistics.

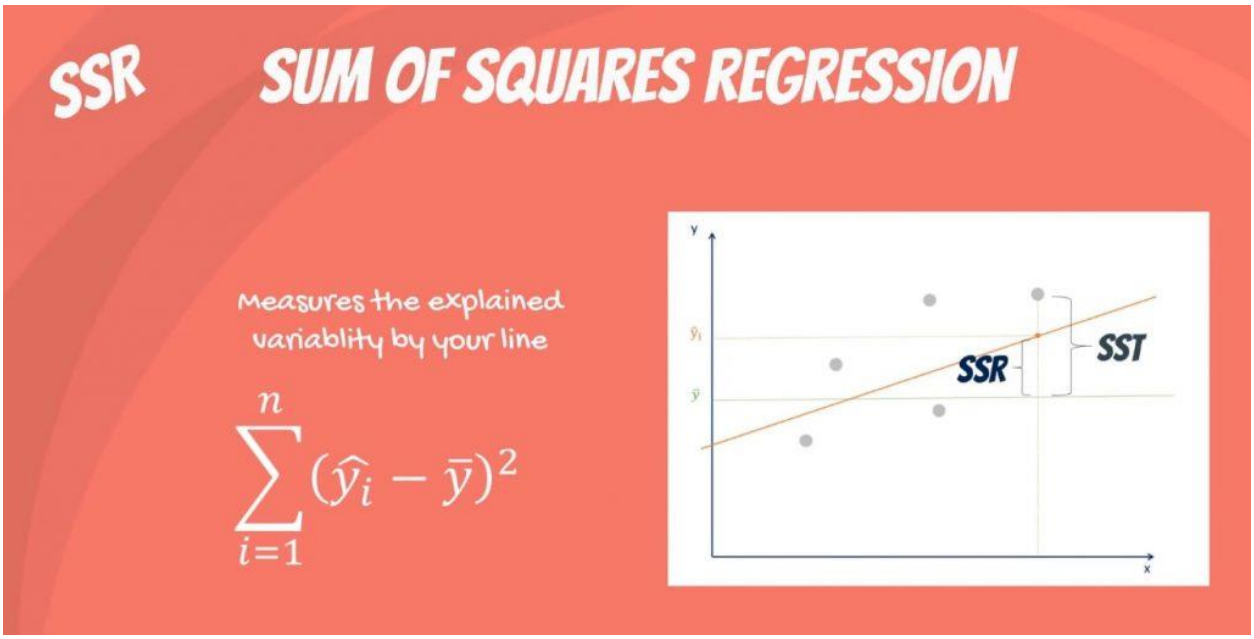


It is a measure of the total variability of the dataset.

Side note: There is another notation for the **SST**. It is **TSS** or **total sum of squares**.

What is the SSR?

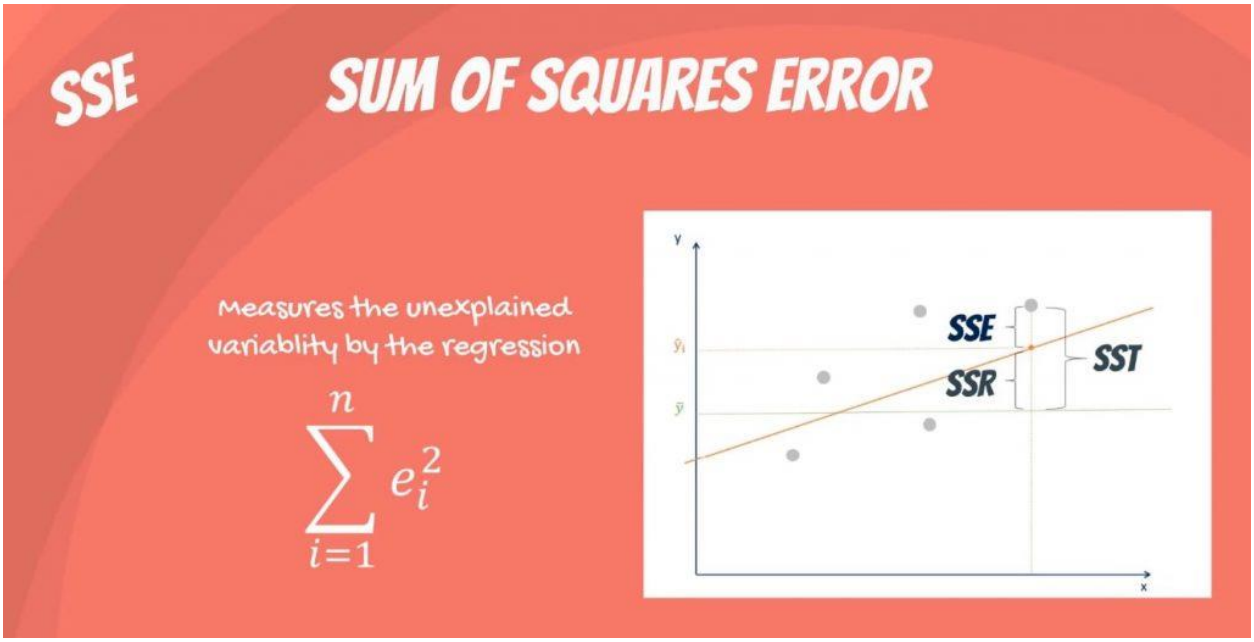
The second term is the **sum of squares due to regression**, or **SSR**. It is the sum of the differences between the *predicted* value and the **mean** of the *dependent variable*. Think of it as a measure that describes how well our line fits the data.



If this value of **SSR** is equal to the **sum of squares total**, it means our **regression model** captures all the observed variability and is perfect. Once again, we have to mention that another common notation is **ESS** or **explained sum of squares**.

What is the SSE?

The last term is the **sum of squares error**, or **SSE**. The error is the difference between the *observed* value and the *predicted* value.

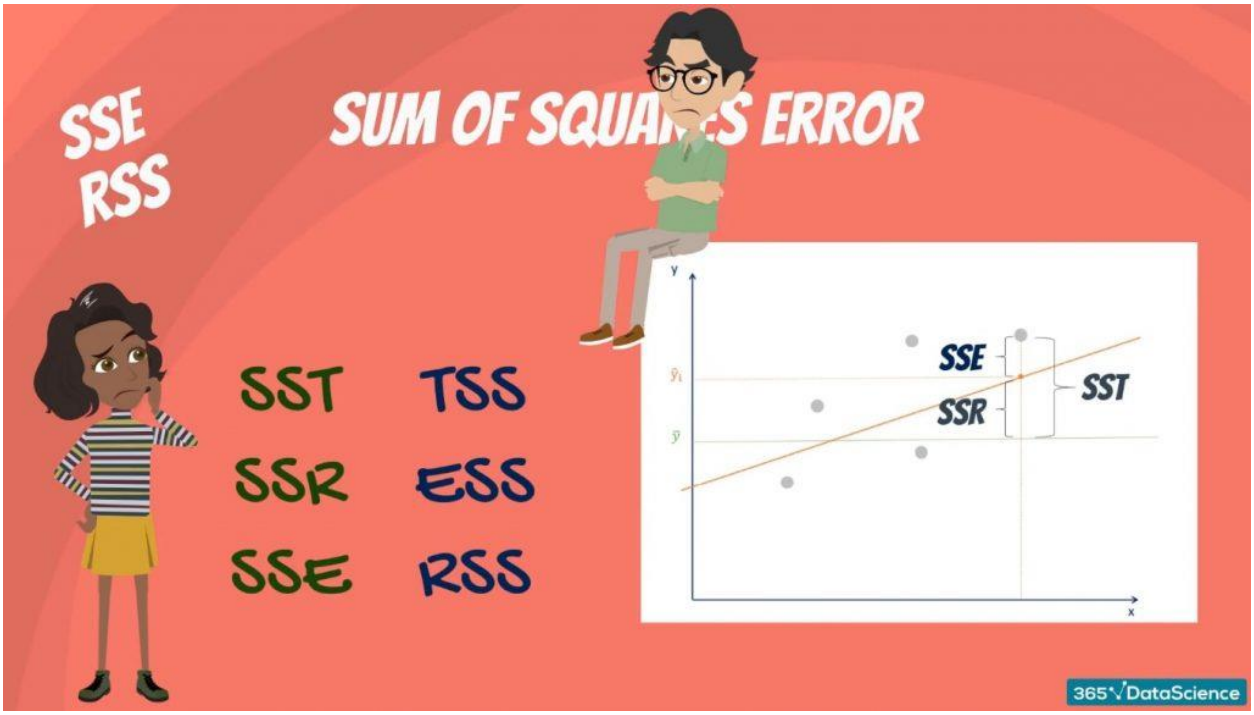


We usually want to minimize the error. The smaller the error, the better the estimation power of the **regression**. Finally, I should add that it is also known as **RSS** or **residual sum of squares**. Residual as in: remaining or unexplained.

The Confusion between the Different Abbreviations

It becomes really confusing because some people denote it as **SSR**. This makes it unclear whether we are talking about the **sum of squares due to regression** or **sum of squared residuals**.

Simply remember that the two notations are **SST**, **SSR**, **SSE**, or **TSS**, **ESS**, **RSS**.



How Are They Related?

Mathematically, **SST** = **SSR** + **SSE**.

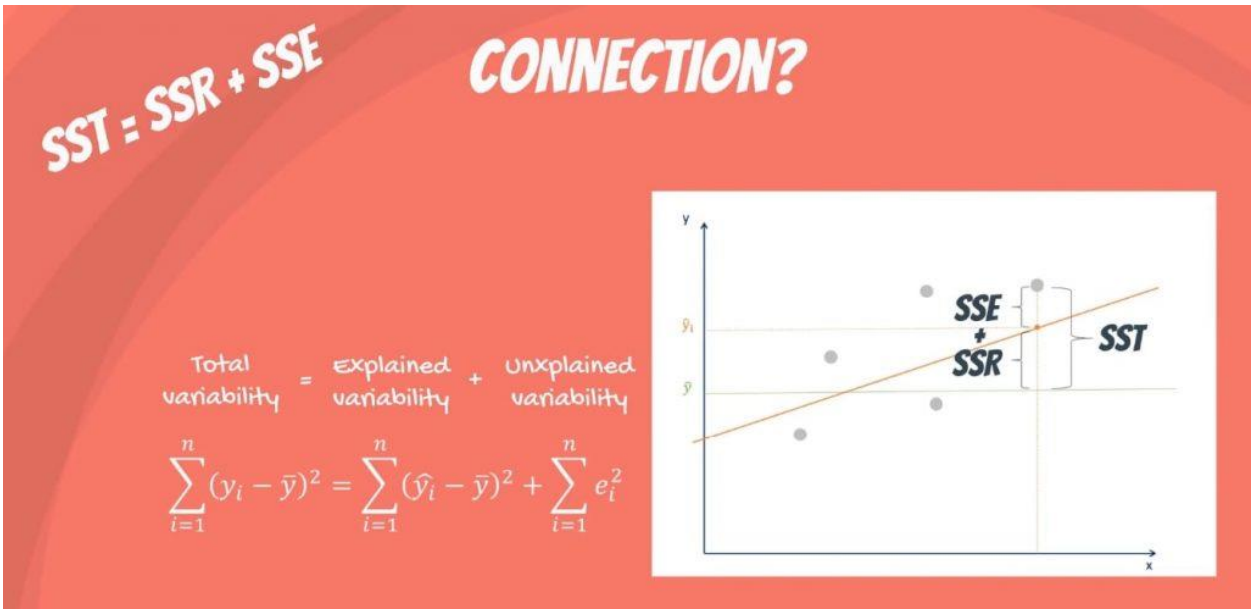
$\Sigma (y_i - \bar{y})^2 = \Sigma (\hat{y}_i - \bar{y})^2 + \Sigma e_i^2$

Total sum of squares = Regression sum of squares + Error sum of squares

$SST = \Sigma (y_i - \bar{y})^2$

$SSR = \Sigma (\hat{y}_i - \bar{y})^2$

$SSE = \Sigma e_i^2 = \Sigma (y_i - \hat{y}_i)^2$



https://365datascience.com/tutorials/statistics-tutorials/sum-squares/

Regression Performance

The variation of actual responses y_i, $i = 1, \ldots, n$, occurs partly due to the dependence on the predictors $\mathbf{x}_i$. However, there is also an additional inherent variance of the output.

The **coefficient of determination**, denoted as R^2, tells you which amount of variation in y can be explained by the dependence on **x** using the particular regression model. Larger R^2 indicates a better fit and means that the model can better explain the variation of the output with different inputs.

Now the proportion of the sample variability of the dependent variable explained by its linear relationship with the independent variable is given by the coefficient of determination, $R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$ where $0 \leq R^2 \leq 1$.

The higher is R^2, the higher is the explanatory power of the regression.

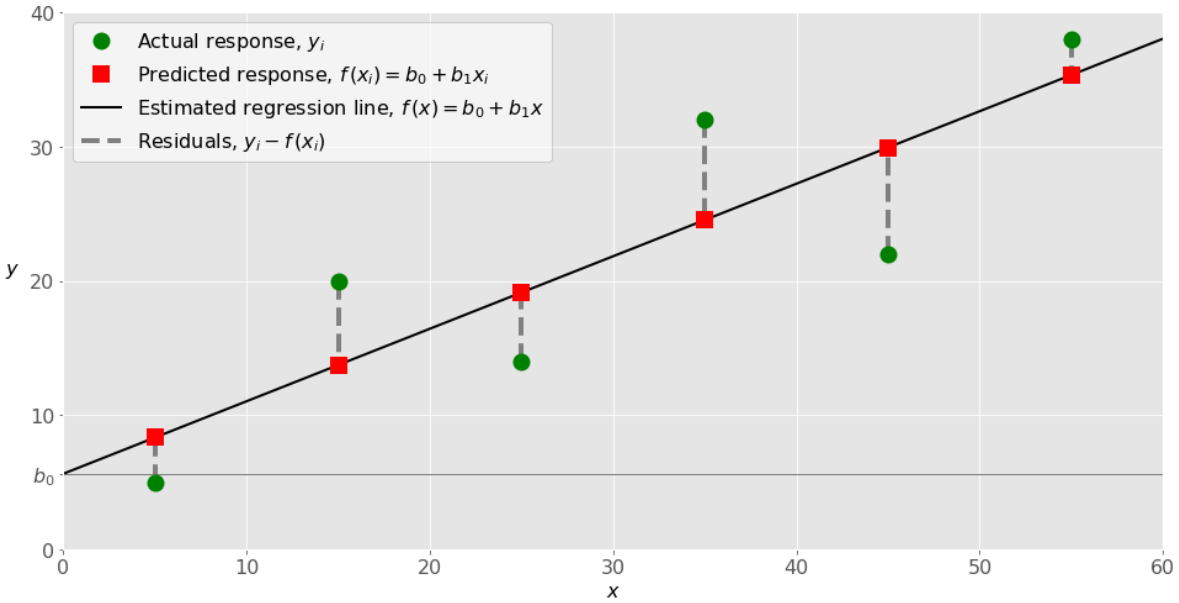
If $R^2 = 1$ then SSE = 0 (every observation falls on the regression line), that is to the **perfect fit** since the values of predicted and actual responses fit completely to each other.

If $R^2 = 0$ then SSR = 0, SSE = SST

Simple Linear Regression

Simple or single-variate linear regression is the simplest case of linear regression with a single independent variable, $\mathbf{x} = x$.

The following figure illustrates simple linear regression:



Example of simple linear regression

When implementing simple linear regression, you typically start with a given set of input-output (x-y) pairs (green circles). These pairs are your observations. For example, the leftmost observation (green circle) has the input $x = 5$ and the actual output (response) $y = 5$. The next one has $x = 15$ and $y = 20$, and so on.

The estimated regression function (black line) has the equation $(x) = b_0 + b_1 x$. Your goal is to calculate the optimal values of the predicted weights b_0 and b_1 that minimize SSR and determine the estimated regression function. The value of b_0, also called the **intercept**, shows the point where the estimated regression line crosses the y axis. It is the value of the estimated response (x) for $x = 0$. The value of b_1 determines the **slope** of the estimated regression line.

The predicted responses (red squares) are the points on the regression line that correspond to the input values. For example, for the input $x = 5$, the predicted response is (5) = 8.33 (represented with the leftmost red square).

The residuals (vertical dashed gray lines) can be calculated as $y_i - (\mathbf{x}_i) = y_i - b_0 - b_1 x_i$ for $i = 1, \ldots, n$. They are the distances between the green circles and red squares. When you implement linear regression, you are actually trying to minimize these distances and make the red squares as close to the predefined green circles as possible.

Multiple Linear Regression

Multiple or multivariate linear regression is a case of linear regression with two or more independent variables.

If there are just two independent variables, the estimated regression function is (x_1, x_2) , y= $b_0 + b_1x_1 + b_2x_2$. It represents a regression plane in a three-dimensional space. The goal of regression is to determine the values of the weights b_0, b_1, and b_2 such that this plane is as close as possible to the actual responses and yield the minimal SSR.

The case of more than two independent variables is similar, but more general. The estimated regression function is $(x_1, \ldots, x_r)$, $y = b_0 + b_1x_1 + \cdots + b_rx_r$, and there are $r + 1$ weights to be determined when the number of inputs is r.

$Y=b_0+b_1x_1+b_2x_2$

Polynomial Regression

You can regard polynomial regression as a generalized case of linear regression. You assume the polynomial dependence between the output and inputs and, consequently, the polynomial estimated regression function.

In other words, in addition to linear terms like b_1x_1, your regression function f can include non-linear terms such as $b_2x_1^2$, $b_3x_1^3$, or even $b_4x_1x_2$, $b_5x_1^2x_2$, and so on.

The simplest example of polynomial regression has a single independent variable, and the estimated regression function is a polynomial of degree 2: $(x) = b_0 + b_1x + b_2x^2$.

Now, remember that you want to calculate b_0, b_1, and b_2, which minimize SSR. These are your unknowns!

Keeping this in mind, compare the previous regression function with the function $(x_1, x_2) = b_0 + b_1x_1 + b_2x_2$ used for linear regression. They look very similar and are both linear functions of the unknowns b_0, b_1, and b_2. This is why you can **solve the polynomial regression problem as a linear problem** with the term x^2 regarded as an input variable.

In the case of two variables and the polynomial of degree 2, the regression function has this form: $(x_1, x_2) = b_0 + b_1x_1 + b_2x_2 + b_3x_1^2 + b_4x_1x_2 + b_5x_2^2$.

Underfitting and Overfitting

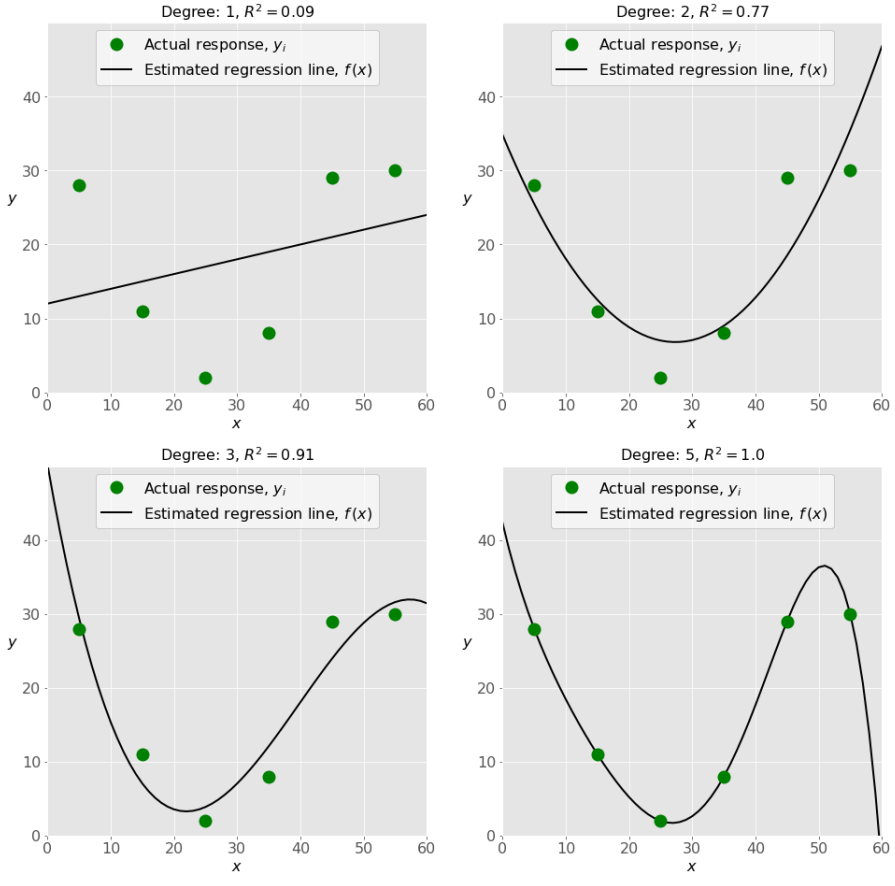
One very important question that might arise when you're implementing polynomial regression is related to **the choice of the optimal degree of the polynomial regression function**.

There is no straightforward rule for doing this. It depends on the case. You should, however, be aware of two problems that might follow the choice of the degree: **underfitting** and **overfitting**.

Underfitting occurs when a model can't accurately capture the dependencies among data, usually as a consequence of its own simplicity. It often yields a low R^2 with known data and bad generalization capabilities when applied with new data.

Overfitting happens when a model learns both dependencies among data and random fluctuations. In other words, a model learns the existing data too well. Complex models, which have many features or terms, are often prone to overfitting. When applied to known data, such models usually yield high R^2. However, they often don't generalize well and have significantly lower R^2 when used with new data.

The next figure illustrates the underfitted, well-fitted, and overfitted models:



Example of underfitted, well-fitted and overfitted models

The top left plot shows a linear regression line that has a low R^2. It might also be important that a straight line can't take into account the fact that the actual response increases as x moves away from 25 towards zero. This is likely an example of underfitting.

The top right plot illustrates polynomial regression with the degree equal to 2. In this instance, this might be the optimal degree for modeling this data. The model has a value of R^2 that is satisfactory in many cases and shows trends nicely.

The bottom left plot presents polynomial regression with the degree equal to 3. The value of R^2 is higher than in the preceding cases. This model behaves better with known data than the previous ones. However, it shows some signs of overfitting, especially for the input values close to 60 where the line starts decreasing, although actual data don't show that.

Finally, on the bottom right plot, you can see the perfect fit: six points and the polynomial line of the degree 5 (or higher) yield $R^2 = 1$. Each actual response equals its corresponding prediction.

In some situations, this might be exactly what you're looking for. In many cases, however, this is an overfitted model. It is likely to have poor behavior with unseen data, especially with the inputs larger than 50.

For example, it assumes, without any evidence, that there is a significant drop in responses for $x > 50$ and that y reaches zero for x near 60. Such behavior is the consequence of excessive effort to learn and fit the existing data.