# Data Science
# Linear regression, goodness of fit

Manuel Pita, Universidade Lusófona. CICANT October 29, 2023

This is version 0.1 of this booklet. if you find any errors, please send and email to
manuel.pita@ulusofona.pt

## Introduction

Up to this point we learnt to find the estimates for the coefficients in a simple linear model with the form $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Once we get these estimates we asked whether, given the evidence (the data points used for the regression), we are confident that there is a relationship between $x$ and $y$, which means that $\hat{\beta}_1 \neq 0$. Recall that in our lecture example we rejected the null hypothesis that $\hat{\beta}_1 = 0$. In other words, we came to the statistical conclusion that, given the data as evidence, it would be *extremely* unlikely that there was no relationship between $x$ and $y$.

Our next question assumes we solved the previous one, and that we have statistical confidence to believe $x$ is a predictor of $y$, and that question asks:

> To what extent the model fits the data?

The quality of a linear model is normally assessed using two interrelated quantities: the *residual standard error* (RSE) and the $R^2$ statistic.

## Residual standard error

The *residual standard error* (RSE) gives you an idea of how bad your model's predictions might be on average. If the residual standard error is large, it means your predictions might be quite a bit off. If it's small, you're probably predicting pretty accurately!

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

where RSS is the residual sum of squares we learnt about in previous lectures.

The Residual Standard Error (RSE) provides an estimate of the amount of variability in the response that's not explained by the predictors. Here's how to interpret the RSE value:

**Meaning**. The RSE gives you an average amount by which the observed responses deviate from the regression line. In other words, it is an average measure of how far off your predictions are likely to be.

**Scale**. It is essential to consider the scale of your dependent variable (the variable you're trying to predict) when interpreting the RSE. For instance, if you're predicting house prices and your RSE is €10,000, that means your predictions might typically be off by about €10,000.

**Comparison with the mean of the response**. Sometimes, it is helpful to compare the RSE to the mean of the response variable. If the RSE is very small in comparison to the mean, then the model's predictions are quite precise. If the RSE is large relative to the mean, the model might not be as accurate.

| Go and compute the RSE for the simple example we worked through in our last Zoom session. How would you interpret your results? Is the model good? bad? Are you unsure? Do not continue to the next section before completing this exercise. |
|---|

## $R^2$ statistic

As previously mentioned, RSE gives us a measure of how much our predictions typically deviate from the actual observed values. It is a measure of the model's prediction error. If RSE is large, our model might not be making very accurate predictions.

The $R^2$ statistic, also known as the coefficient of determination is a statistic statistical tells us the proportion of the variance in the dependent variable that is predictable from the independent variable(s). In simpler terms, $R^2$ gives us a measure of how well the variations in one variable explain or predict the variation in a second variable.

Unlike the RSE, $R^2$ can only take values between 0 and 1. An $R^2 = 0$ means none of the variance in the dependent variable is explained by the independent variable(s), whereas $R^2 = 1$ means all of the variance in the dependent variable is explained by the independent variable(s).

Unlike RSE, which directly measures prediction error in the units of the dependent variable, $R^2$ is a proportion. A higher value indicates a better fit of the model, but there's a catch. Adding more predictors to your model can artificially inflate $R^2$ (making the model look better than it actually is), even if those predictors aren't truly meaningful. This means a model with a higher $R^2$ isn't always better, especially if it's overfitting to the data. Practical Implications:

There is no universally agreed $R^2$ value for which we say that a given model is "good". This depends heavily on the domain of application. In some scientific fields, a low $R^2$ can still be of great practical significance.

Not an Absolute Measure: While $R^2$ can be a helpful measure of goodness-of-fit, and the first one we usually report, it shouldn't be solely relied upon. It's essential to consider other model diagnostics and understand the context of the problem being solved. You will learn a little more about this in the next lectures.

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

where TSS is total sum of squares, $\sum(y_i - \bar{y})^2$.

To better understand what $R^2$ measures, imagine you and your friends are trying to predict the time it takes to finish a specific video game based on the number of hours you train each day. Without any predictor (like training hours), your best guess for everyone's game finishing time might just be the average time taken by all players.

1. **Total Variability (TTS)**: This represents the difference between each person's actual time and the average time of all players. It's like saying, "If I always guessed that everyone would take the average time, how wrong would I be?"

2. **Residual Variability (RSS)**: Now, let's say you use your model that considers training hours. This value represents the difference between each person's actual time and what your model predicts. It's like saying, "Now that I'm using my prediction model, how wrong am I?"

3. $R^2$: This is essentially a fraction. The numerator (top part) is how wrong your model is, and the denominator (bottom part) is how wrong you'd be if you just guessed the average time for everyone. If your model is perfect, then the top part will be zero, and $R^2$ will be 1. If your model is as bad as just guessing the average time for everyone, $R^2$ will be 0.

In simpler terms, $R^2$ tells you what portion of the total variability in game finishing times is captured by considering training hours. If $R^2$ is close to 1, it means your model (based on training hours) does a great job in predicting game finishing times. If it's close to 0, it means the model isn't much better than just guessing the average finishing time for everyone.