Linear Regression

Linear regression is probably the simplest approach for statistical learning. It is a good starting point for more advanced approaches, and in fact, many fancy statistical learning techniques can be seen as an extension of linear regression. Therefore, understanding this simple model will build a good base before moving on to more complex approaches.

Linear regression is very good to answer the following questions:

- Is there a relationship between 2 variables?
- How strong is the relationship?
- Which variable contributes the most?
- How accurately can we estimate the effect of each variable?
- How accurately can we predict the target?
- Is the relationship linear? (duh)
- Is there an interaction effect?

Estimating the coefficients

Let's assume we only have one variable and one target. Then, linear regression is expressed as:

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

Equation for a linear model with 1 variable and 1 target

In the equation above, the *betas* are the coefficients. These coefficients are what we need in order to make predictions with our model.

So how do we find these parameters?

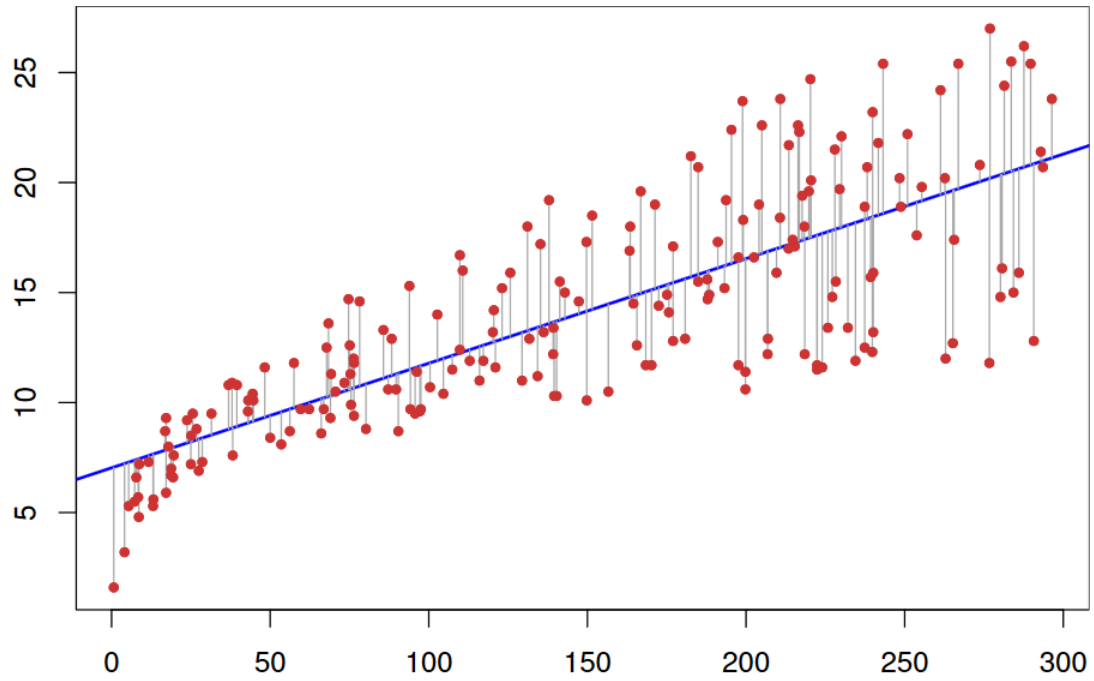
To find the parameters, we need to minimize the **least squares** or the **sum of squared errors**. Of course, the linear model is not perfect and it will not predict all the data accurately, meaning that there is a difference between the actual value and the prediction. The error is easily calculated with:

$$e_i = y_i - \hat{y}_i$$

Subtract the prediction from the true value

But why are the errors squared?

We square the error, because the prediction can be either above or below the true value, resulting in a negative or positive difference respectively. If we did not square the errors, the sum of errors could decrease because of negative differences and not because the model is a good fit. Also, squaring the errors penalizes large differences, and so the minimizing the squared errors "guarantees" a better model. Let's take a look at a graph to better understand.



Linear fit to a data set

In the graph above, the red dots are the true data and the blue line is linear model. The grey lines illustrate the errors between the predicted and the true values. The blue line is thus the one that minimizes the sum of the squared length of the grey lines.

After some math that is too heavy for a blog post, you can finally estimate the coefficients with the following equations:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Where *x bar* and *y bar* represent the mean.

Nice!

Estimate the relevancy of the coefficients

Now that you have coefficients, how can you tell if they are relevant to predict your target?

The best way is to find the *p-value.* The *p-value* is used to quantify statistical significance; it allows to tell whether the null hypothesis is to be rejected or not.

The null hypothesis?

For any modelling task, the hypothesis is that **there is some correlation** between the features and the target. The null hypothesis is therefore the opposite: **there is no correlation** between the features and the target.

So, finding the *p-value* for each coefficient will tell if the variable is statistically significant to predict the target. As a general rule of thumb, if the *p-value* is **less than 0.05**: there is a strong relationship between the variable and the target.

Assess the accuracy of the model

You found out that your variable was statistically significant by finding its *p-value*. Great!

Now, how do you know if your linear model is any good?

To assess that, we usually use the RSE (residual standard error) and the R^2 statistic.

R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. Whereas correlation explains the strength of the relationship between an independent and dependent variable, R-squared explains to what extent the variance of one variable explains the variance of the second variable.

The F value is the ratio of the mean regression sum of squares divided by the mean error sum of squares. Its value will range from zero to an arbitrarily large number. The value of Prob(F) is the probability that the null hypothesis for the full model is true (i.e., that all of the regression coefficients are zero).

Here, the F-statistic is calculated for the overall model, whereas the *p-value* is specific to each predictor. If there is a strong relationship, then F will be much larger than 1. Otherwise, it will be approximately equal to 1.

The first error metric is simple to understand: the lower the residual errors, the better the model fits the data (in this case, the closer the data is to a linear relationship).

As for the R^2 metric, it measures the **proportion of variability in the target that can be explained using a feature X**. Therefore, assuming a linear relationship, if feature X can explain (predict) the target, then the proportion is high and the R^2 value will be close to 1. If the opposite is true, the R^2 value is then closer to 0.

Multiple Linear Regression

In real life situations, there will never be a single feature to predict a target. So, do we perform linear regression on one feature at a time? Of course not. We simply perform multiple linear regression.

The equation is very similar to simple linear regression; simply add the number of predictors and their corresponding coefficients:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Multiple linear regression equation. **p** is the number of predictors

Assess the relevancy of a predictor

Previously, in simple linear regression, we assess the relevancy of a feature by finding its *p-value*.

In the case of multiple linear regression, we use another metric: the F-statistic.

How *larger than 1* is large enough?

This is hard to answer. Usually, if there is a large number of data points, F could be slightly larger than 1 and suggest a strong relationship. For small data sets, then the F value must be way larger than 1 to suggest a strong relationship.

Why can't we use the *p-value* in this case?

Since we are fitting many predictors, we need to consider a case where there are a lot of features (*p* is large). With a very large amount of predictors, there will always be about 5% of them that will have, by chance, a very small *p-value* ***even though they are not statistically significant.*** Therefore, we use the F-statistic to avoid considering unimportant predictors as significant predictors.

Assess the accuracy of the model

Just like in simple linear regression, the R^2 can be used for multiple linear regression. However, know that adding more predictors will always increase the R^2 value, because the model will necessarily better fit the training data.

Yet, this does not mean it will perform well on test data (making predictions for unknown data points).