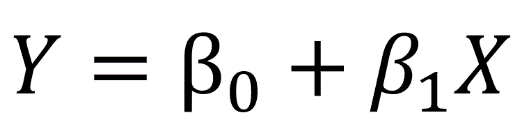
**Linear Regression**

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?
* Is there an interaction effect?

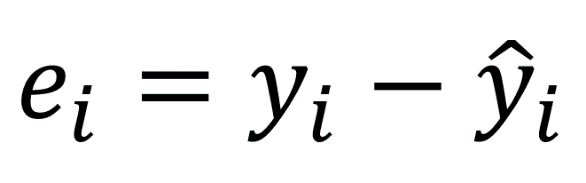
**Estimating the coefficients**

* one variable and one target. Then, linear regression is expressed as:



To find the parameters, we need to minimize the **least squares** or the **sum of squared errors**.

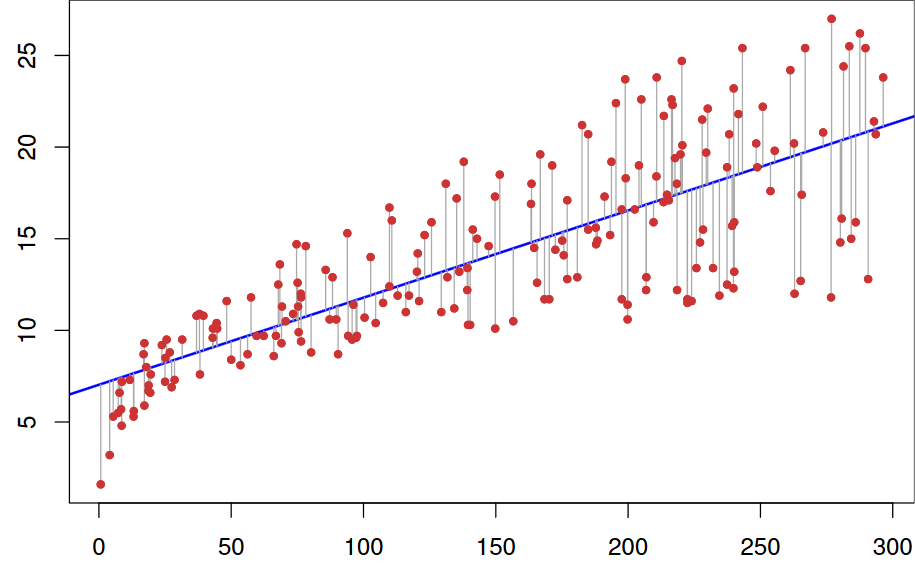
The error is easily calculated with:



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.

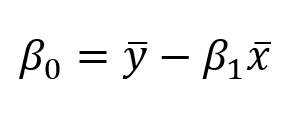
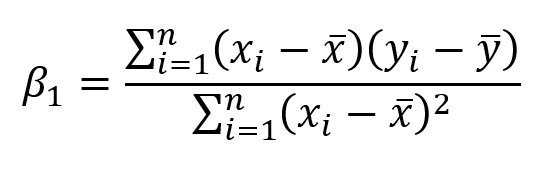
Also, squaring the errors penalizes large differences, and so the minimizing the squared errors “guarantees” a better model.



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.



**Estimate the relevancy of the coefficients**

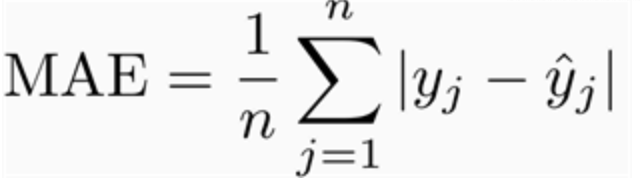
The best way is to find the *p-value.* The *p-value* is used to quantify statistical significance; it allows to tell whether the null hypothesis is to be rejected or not.

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.

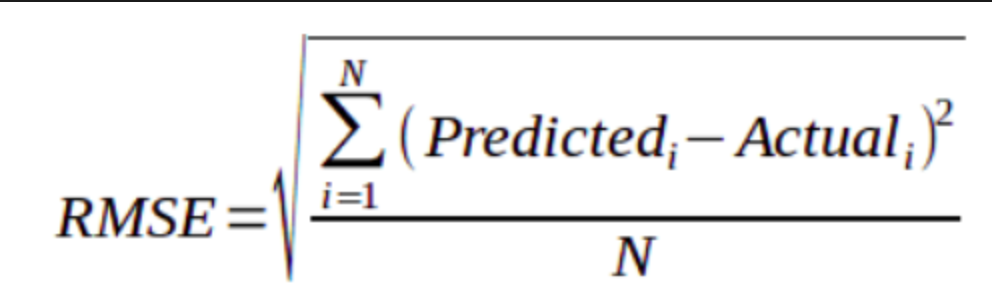
So, finding the *p-value* for each coefficient will tell if the variable is statistically significant to predict the target. As a rule of thumb, if the *p-value* is **less than 0.05**: there is a strong relationship between the variable and the target.

### Mean Absolute Error (MAE)

The Mean Absolute Error measures the average of the absolute difference between each ground truth and the predictions



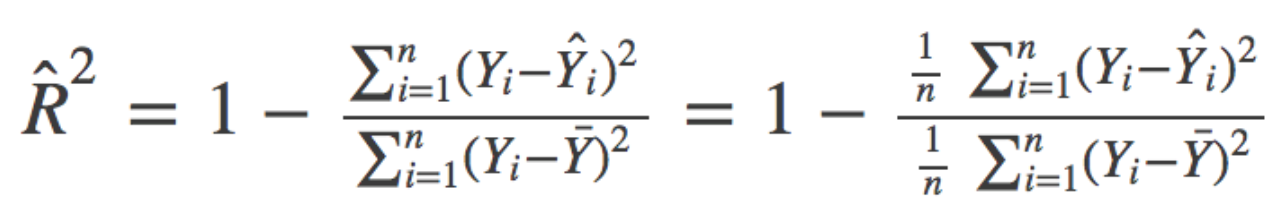
Root Mean Squared Error (RMSE)



The Root Mean Squared Error measures the square root of the average of the squared difference between the predictions and the ground truth.

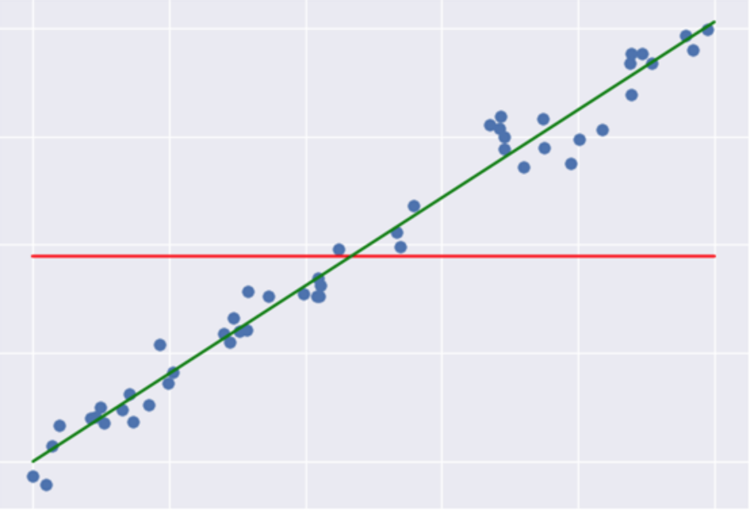
Since the RMSE is squaring the difference between the predictions and the ground truth, any significant difference is made more substantial when it is being squared. **RMSE is more sensitive to outliers**.

**R-Squared**



As for the R² 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 R² value will be close to 1. If the opposite is true, the R² value is then closer to 0.



**SSR: Regression Sum of Squares** quantifies how far the estimated sloped regression line (green line) is from the horizontal (red line). The red line is the average of the ground truth.

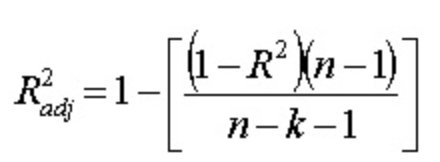
**SSE: Error sum of squares** quantifies how much the data points (blue dots) is from the prediction (green line).

**SSTO: Total Sum of Squares** quantifies how much the data points (blue dot) is from the horizontal (red line).

**SSTO = SSE + SSR**

**R-Squared = SSR / SSTO**

**Adjusted R-squared**



*The n represents the total number of observations.*

*The k represents the total number of variables in your model.*

Focus on the equation above. As you increase the number of variables (*keeping everything else constant*), you are decreasing the denominator.

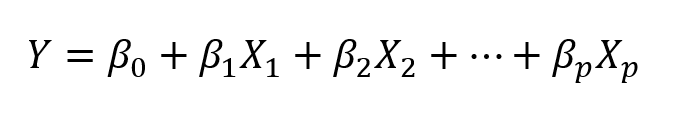
Hence, you are increasing the left side of the equation (inside the brackets). Since we are subtracting by 1, we are lowering the Adjusted R-Squared. Hence, the improvement in the model must be stronger than the penalty for the Adjusted R-Squared to improve as we increase the number of variables.

#### Notes:

* For the R-Squared and the Adjusted R-Squared, the closer the value to 1, the better performer our model!
* RMSE/MAE is used to evaluate the variance in the errors. Additionally, the value within itself doesn’t tell you much. You must compare different models to reap from the RMSE/MAE.
* However, R-Square/Adjusted R-Squared doesn’t need to be compared between different models. If the R-Squared/Adjusted R-Squared is .10, we can acknowledge that the model is not doing a great job is modeling.

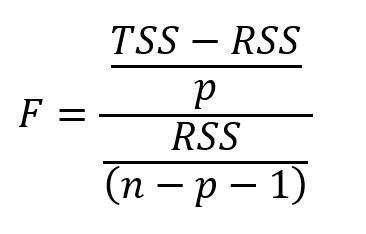
**Multiple Linear Regression**

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



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

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



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.

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.

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.