Performance Assessment - R Squared

We have already discussed what linear regression does and how it is to be interpreted. Now we're ready to move to the more important topic which is **Performance Assessment**.

We've run regression and have built our estimates of the ϴ's. Maybe we have also learned the model and generated the prediction.

But we need to ask a few questions here:

- How much do we trust our predictions i.e how much do we trust the model that we have?
- What can we learn from the model that we have?
- How accurate are the parameters that we have estimated?

So, we will look through different aspects of this question.

R2 (R Squared):

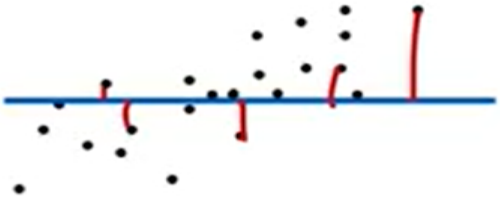
It is one of the metrics using which performance is assessed. This is the performance of a predictor through a single number that tells us how well this predictor is doing on our data set.

Say, we have our dataset. If we didn't do any regression at all, and somebody was asking us, 'can you predict Y for a new individual without knowing what the X value is?'.

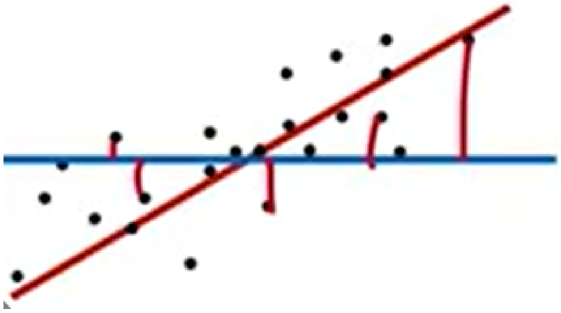
Now, if you were to make a prediction for a new person, then you can just look at the average of the Y values in your data set, and that's as good an estimate of a new Y as any other. So we're not looking at the X's at all. We're just making a prediction without looking at the X's so that prediction is constant and the reasonable prediction is to just take the average.

- So, prediction if no regression : $\overline{Y} = \frac{1}{n}Y_i$

Now, if we make that prediction, we're going to be making some errors. And we take all those errors and sum the squares of those and this is the **Total Sum of the Squares** and it tells us how much Y varies from one individual to another.



- Total Sum of Squares: TSS = $\sum_{i=1}^{n}(Yi - \bar{Y})^2$



But then, we run our prediction and our prediction gives us a predictor.

For example, the red line that is shown here.
Once we have that red line, then we should be looking at the sum of the squared residuals based on that line.

We look at those sums of squared errors and see what they are. So now we're using our predictors ($\Theta^T X_i$) and looking at the prediction errors using the best predictor that was given to us by linear regression. We call that the **Residual Sum of Squares**. It is the sum of squared errors after we've done our regression.

- Residual Sum of Squares: RSS = $\sum_{i=1}^{n}(Y_i - \hat{\theta}^T \mathbf{X}_i)^2$

Because we choose theta(Θ) in the best possible way so the residual sum of squares is going to be less than the initial sum of squares that we had. The red line could have been the same as the blue line. But when we tweak it, we do better. This residual sum of squares in the unexplained variation in Y, after we took into account the value of X.

Now the question is, how much have we reduced in the sum of squares i.e using the values of X's, how much smaller is the variation in the Y's?

This is captured by this numerical quantity, which is called **R Squared**. We look at how much variation was left out of the total, and when we take one minus, that tells us how much was removed out of the total.

- $R^2 = 1 - \frac{RSS}{TSS}$, fraction of variation in Y that has been explained

So, R Squared is an important performance metric that tells us about the quality of the fit. It indicates the variance explained in the dependent variable by the independent variables.

$$0 \le R^2 \le 1$$

R2 is always going to be non-negative and between 0 and 1. As such it tells us, out of the total variation, what fraction has been explained. The closer it is to 1, the better it is, as it defines that more variance has been explained in such a case. And so a **higher R squared is preferred.**

R^2 has also a mathematical interpretation in simple regression when X is one-dimensional. R^2 is just the square of the correlation coefficient between the X's and the Y's. So it has a mathematical interpretation. Although when we are in higher dimensions when X is high dimensional, that interpretation doesn't hold anymore.

Let us look at our example:

$$\widehat{\text{Sales}} = 2.94 + 0.046 \cdot (\text{TV}) + 0.19 \cdot (\text{Radio}) - 0.001 \cdot (\text{NewsP})$$

$$R^2 = 0.897$$

We ran our multiple regression model and we got the predictor for sales. And this predictor is pretty good. R^2 is about 0.9 here. That means that 90% of the variation in sales between different regions can be explained by just looking at the **advertising budget**. So the advertising budget explains the sales.

We can repeat the same exercise by looking at just one variable at a time. Let's look at newspaper advertising.

$$\widehat{\text{Sales}} = 12.35 + 0.055 \cdot (\text{NewsP})$$

$$R^2 = 0.05 \qquad \text{Newspaper budget explains little}$$

We run a simple regression where we try to predict sales by looking at the advertising budget of the Newspaper only.

In this case, R^2 turns out to be really, really small. And that means that the newspaper budget is not very useful for predicting sales. They do not explain the sales in any way.

Now,

- For TV alone: R^2 = 0. 61
- For Radio alone: R^2= 0. 33

We also repeat the same exercise and run simple linear regressions using just the TV budget and it does a decent job in predicting sales to some extent. Radio also by itself turns out to explain sales to some extent, although to a much smaller extent.

So what's happening here is that each individual variable TV or radio helps explain the sales to some extent. But when we put all three variables together, then they provide a much better explanation. So, together they predict the sales quite well.

This is typical in regression. When you use more variables, R^2 can only go up. The more variables i.e the more X's we throw in, the better our predictions are going to be.
On the other hand, we shouldn't overdo that. If we throw lots and lots of variables, we're going to make perfect predictions that can drive R^2 all the way up to 1.

But those perfect predictions might correspond to overfitting so we need to be careful about this.

$$- \text{ adjusted } R^2\text{:} \quad 1 - \frac{\text{RSS}/(n-m-1)}{\text{TSS}/(n-1)} \qquad 0.897 \rightarrow 0.896$$

They have some formulas that say when you increase the number of variables (‘m’ in the above equation), then you should adjust the R^2 that you report taking into account that you have one extra degree of freedom for fitting the data.

In our particular example, the adjusted R^2 is almost the same as the R^2 that we had before. So this formula doesn't really make much of a difference. And in general, whenever n is big and much bigger than m, then (n - m) is about the same as ‘n’ and so the adjusted R^2 is essentially the same as the R^2, and it's not really a difference that we should worry about.