**Unit-1 Chapter-1 Lecture-1.5**

**Regression**

A regression model expresses a ‘dependent’ variable as a function of one or more ‘independent’ variables, generally in the form:

7.png

 Regression with a single dependent variable y whose value is dependent upon the independent variable x is expressed as

8.png

where **α** is a constant, so is **β**. **x**is the independent variable and **ϵ** is the error term. Given a set of data points, it is fairly easy to calculate alpha and beta. If done manually, beta is calculated as:

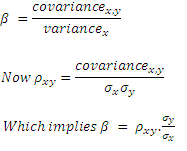
**β = covariance of the two variables/variance of the independent variable**

Once the beta is known, alpha can be calculated as

**α = mean of the dependent variable (ie y) - β \* mean of the independent variable (ie x)**

**Beta and correlation**

At this point, it is important to point out the relationship between beta and correlation.



**Predicted versus the observed value**

Now let us go back to the initial equation:

9.png

Now that we have seen how to calculate α and β, it is probably possible to say that we can ‘predict’ y if we know the value of x. The ‘predicted’ value of y is provided to us by the regression equation. This is unlikely to be exactly equal to the actual observed value of y.

The difference between the two is explained by the error term - ϵ. This is a random ‘error’ – error not in the sense of being a mistake – but in the sense that the value predicted by the regression equation is not equal to the actual observed value. This error is ‘random’ and not biased, which means that if you sum up ϵ across all data points, you get a total of zero. Some observations are farther away from the predicted value than others, but the sum of all the differences will add up to zero. (If it weren't zero, the model would be biased in the sense it was likely to either overstate or understate the value of y.)

Intuitively, the smaller the individual observed values of ϵ, even though adding up to zero, the better is our regression model. How do we measure how small the values of ϵ are? One obvious way would be to add them up and divide by the number of observations to get an ‘average’ value per data point – but that would just be zero as just explained. So what we do is the next best thing: take a sum of the squares of ϵ and divide by the number of observations. For a variable whose mean is zero, this is nothing but its variance.

This number is called the**standard error of the regression**

This error variable ϵ is considered normally distributed with a mean of zero, and a variance equal to σ2.

The standard error can be used to calculate confidence intervals around an estimate provided by our regression model because using this we can calculate the number of standard deviations either side of the predicted value and use the normal distribution to compute a confidence interval. We may need to use at-distribution if our sample size is small.

**Interpreting the standard error of the regression**

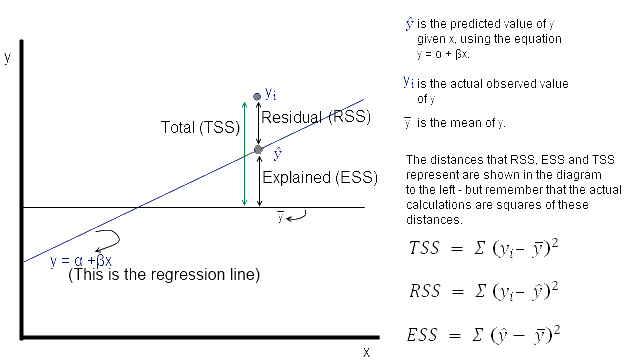
The standard error of the regression is a measure of how good our regression model is – or its ‘goodness of fit’. The problem though is that the standard error is in units of the dependent variable, and on its own is difficult to interpret as being big or small. The fact that it is expressed in the squares of the units makes it a bit more difficult to comprehend.

(RMS error: We can also then take a square root of this variance to get to the standard deviation equivalent, called the RMS error. RMS stands for Root Mean Square – which is exactly what we did, we squared the errors, took their mean, and then the square root of the resultant. This takes care of the problem that the standard error is expressed in square units.)

Coming back to the standard error - what do we compare the standard error to in order to determine how good our regression is? How big is big? This takes us to the next step – understanding the sums of squares – TSS, RSS, and ESS.

**TSS, RSS, and ESS (Total Sum of Squares, Residual Sum of Squares and Explained Sum of Squares)**

Consider the diagram below. Yi is the actual observed value of the dependent variable, y-hat is the value of the dependent variable according to the regression line, as predicted by our regression model. What we want to get is a feel for is the variability of actual y around the regression line, ie, the volatility of ϵ. This is given by the distance yi minus y-hat. Represented in the figure below as RSS. The figure below also shows TSS and ESS – spend a few minutes looking at what TSS, RSS, and ESS represent.



**Fig 5.1-Residual**

    Now **ϵ = observed – the expected value of y**

Thus, ϵ = Yi – y^. The sum of ϵ is expected to be zero. So we look at the sum of squares: The value of interest to us is = Σ (yi – y^ )2. Since this value will change as the number of observations change, we divide by ‘n’ to get a ‘per observation’ number. (Since this is a square, we take the root to get a more intuitive number, ie the RMS error explained a little while earlier. Effectively, RMS gives us the standard deviation of the variation of the actual values of y when compared to the observed values.)

If s is the standard error of the regression, then

**s = sqrt(RSS/(n – 2))**

(where n is the number of observations, and we subtract 2 from this to take away 2 degrees of freedom\*.)

Now

11 (1).png

12 (1).png

13 (1).png

**How good is the regression?**

Intuitively, the regression line given by α + βx will be a more accurate prediction of y if the correlation between x and y is high. We don’t any math to say that if the correlation between the variables is low, then the quality of the regression model will be lower because the regression model is merely trying to fit a straight line on the scatter plot in the best possible way.

Generally, R2, called the coefficient of determination, is used to evaluate how good the ‘fit’ of the regression model is. R2 is calculated as ESS/TSS, ie the ratio of the explained variation to the total variation.

**R2 = ESS/TSS**

R2 is also the same thing as the square of the correlation (stated without proof, but you can verify it in Excel). This means that our initial intuition that the quality of our regression model depends upon the correlation of the variables was correct. (Note that in the ratio ESS/TSS, both the numerator and denominator are squares of some sort – which means this ratio explains how much of the ‘variance’ is explained, not standard deviation. Variance is always in terms of the square of the units, which makes it slightly difficult to interpret intuitively, which is why we have a standard deviation.).

**Other Reading and Video Material**

* Understanding Machine Learning: From Theory to Algorithms by Shai Shalev-Shwartz and Shai Ben-David-Cambridge University Press 2014 [Download](https://www.cse.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf) Buy at Amazon
* Introduction to Machine Learning – the Wikipedia guide [Download](http://datascienceassn.org/sites/default/files/Introduction%20to%20Machine%20Learning.pdf)
* [Weblink (towardsdatascience)](https://towardsdatascience.com/data-science-simplified-simple-linear-regression-models-3a97811a6a3d)
* [Online Reading Material (IIT-Kanpur)](http://home.iitk.ac.in/~shalab/regression/Chapter2-Regression-SimpleLinearRegressionAnalysis.pdf)
* [NPTEL Video](https://www.youtube.com/watch?v=MXTsSXIa4i0)