Linear Regression Output

Anurag Nagar

Some good resources are below:

- https://www.geeksforgeeks.org/interpreting-the-results-of-linear-regression-using-ols-summary/
- https://medium.com/swlh/interpreting-linear-regression-through-statsmodels-summary-4796d359035a
- https://datascience.stackexchange.com/questions/41034/interpreting-multi-linear-regression-results
- https://machinelearningmastery.com/probabilistic-model-selection-measures/

Let's try to figure out what each of the above value means:

1. **Call** - gives the formula of the model.
2. **Residuals** - are defined as $\hat{y} - y$ where $\hat{y}$ is the predicted value and y is the actual value. The distribution of residuals is shown here. A large range implies the errors are high, and in all directions.
3. **Coefficients (β)** - gives the predicted coefficients for each variable and the constant (intercept) term.
4. **Standard Error ($se(\beta)$)** - measure of the variability in the estimate for the coefficient. A lower value, as compared with the coefficient, is a better indicator. The 95% confidence interval is defined as the region $\beta \pm 2se(\beta)$
5. **t-Value** - defined as the estimated coefficient divided by its standard error $\beta/se(\beta)$. A large value indicates a good precision, i.e. we are sure of the coefficient and its standard error is relatively small as compared to coefficient's value. It is used to test the hypothesis that the true value of the coefficient is non-zero, in order to confirm that the independent variable really belongs in the model.
6. **p-value** - indicates the probability of getting the obtained t-value if the null hypotesis were true. A smaller probability gives you more evidence that you can reject the null hypothesis and claim that the variable plays a significant role. The stars next to the row indicate how significant is the variable. 3 stars mean a very low p-value and we can safely reject the null hypothesis.
7. **Residual Standard Error(RSS)** - The Residual Std Error is just the standard deviation of your residuals. You'd like this number to be proportional to the quantiles of the residuals in part 2. For a normal distribution, the 1st and 3rd quantiles should be 1.5 +/- the std error.
8. **Degrees of Freedom** - The Degrees of Freedom is the difference between the number of observations included in your training sample and the number of variables used in your model (intercept counts as a variable). It's used in conjunction with the RSS estimate above.
9. **R Squared and Adjusted R Squared** - R squared isa statistical measure of how close the data are to the fitted regression line. It is also equal to the percent of the total variation in the dependent variable (y) that is explained by the independent variables (X), i.e., the model's overall "goodness of fit". It is possible to get higher R squared by simply adding more independent variables. Adjusted R Squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. Suppose you have a 1-predictor and 5-predictor models. Does the five predictor model have a higher R-squared because it's better? Or is the R-squared higher because it has more predictors? Adjusted R squared provides the answer for this.

10 **F-statistic and its p-value** Performs an F-test on the model. This takes the parameters of our model and compares it to a model where all the regression coefficients are 0.