

|  |  |  |
| --- | --- | --- |
| **#** | **Name** | **Description** |
| **1** | **Residuals** | The residuals are the difference between the actual values of the variable you're predicting and predicted values from your regression--y - ŷ. For most regressions you want your residuals to look like a normal distribution when plotted. If our residuals are normally distributed, this indicates the mean of the difference between our predictions and the actual values is close to 0 (good) and that when we miss, we're missing both short and long of the actual value, and the likelihood of a miss being far from the actual value gets smaller as the distance from the actual value gets larger.  Think of it like a dartboard. A good model is going to hit the bullseye some of the time (but not everytime). When it doesn't hit the bullseye, it's missing in all of the other buckets evenly (i.e. not just missing in the 16 bin) and it also misses closer to the bullseye as opposed to on the outer edges of the dartboard. |
| **2** | **Significance Stars** | The stars are shorthand for significance levels, with the number of asterisks displayed according to the p-value computed. \*\*\* for high significance and \* for low significance. In this case, \*\*\* indicates that it's unlikely that no relationship exists b/w heights of parents and heights of their children. |
| **3** | **Estimated Coeffecient** | The estimated coefficient is the value of slope calculated by the regression. It might seem a little confusing that the Intercept also has a value, but just think of it as a slope that is always multiplied by 1. This number will obviously vary based on the magnitude of the variable you're inputting into the regression, but it's always good to spot check this number to make sure it seems reasonable. |
| **4** | **Standard Error of the Coefficient Estimate** | Measure of the variability in the estimate for the coefficient. Lower means better but this number is relative to the value of the coefficient. As a rule of thumb, you'd like this value to be at least an order of magnitude less than the coefficient estimate.  In our example, the std error or the parent variable is 0.04 which is 16x less than the estimate of the coefficient (or 1.6 orders of magnitude greater). |
| **5** | **t-value of the Coefficient Estimate** | Score that measures whether or not the coefficient for this variable is meaningful for the model. You probably won't use this value itself, but know that it is used to calculate the p-value and the significance levels. |
| **6** | **Variable p-value** | Probability the variable is *NOT* relevant. You want this number to be as small as possible. If the number is *really* small, R will display it in scientific notation. In or example 2e-16 means that the odds that parent is meaningless is about 1⁄5000000000000000 |
| **7** | **Significance Legend** | The more punctuation there is next to your variables, the better.  Blank=bad, Dots=pretty good, Stars=good, More Stars=very good |
| **8** | **Residual Std Error / Degrees of Freedom** | 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 #1. For a normal distribution, the 1st and 3rd quantiles should be 1.5 +/- the std error.   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). |
| **9** | **R-squared** | Metric for evaluating the goodness of fit of your model. Higher is better with 1 being the best. Corresponds with the amount of variability in what you're predicting that is explained by the model. In this instance, ~21% of the cause for a child's height is due to the height their parent.  WARNING: While a high R-squared indicates good correlation, [correlation does *not* always imply causation](http://en.wikipedia.org/wiki/Correlation_does_not_imply_causation). |
| **10** | **F-statistic & resulting p-value** | Performs an [F-test](http://en.wikipedia.org/wiki/F-test) on the model. This takes the parameters of our model (in our case we only have 1) and compares it to a model that has fewer parameters. In theory the model with more parameters should fit better. If the model with more parameters (your model) doesn't perform better than the model with fewer parameters, the F-test will have a high p-value (probability *NOT* significant boost). If the model with more parameters is better than the model with fewer parameters, you will have a lower p-value.  The DF, or degrees of freedom, pertains to how many variables are in the model. In our case there is one variable so there is one degree of freedom. |