Linear Regression Notes

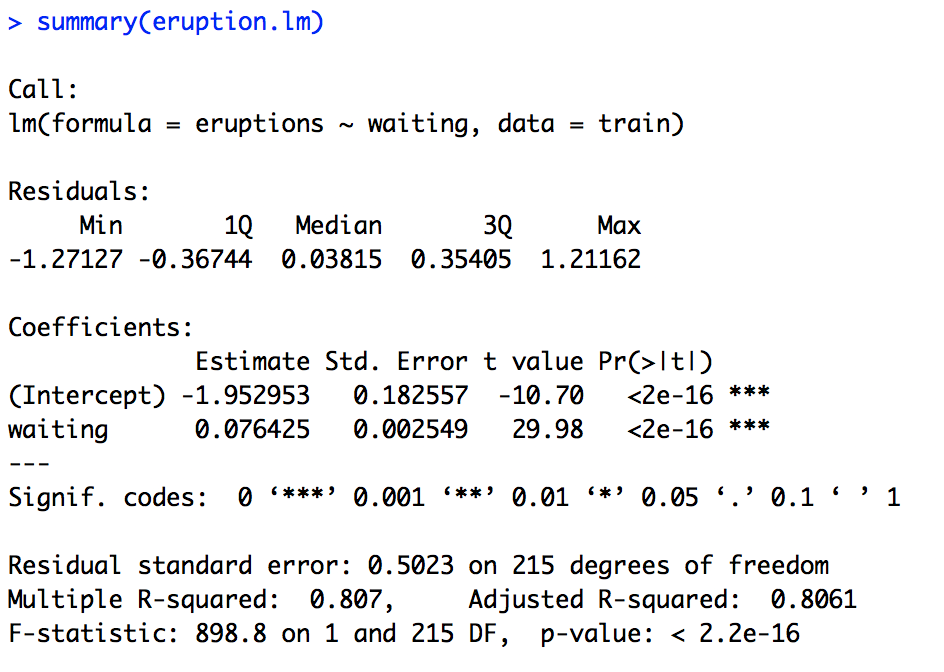
Univariate:

* (Predicted Value from Univariate Linear Regression Model)
* (R-Squared formula) where equals the model predicted value, equals the mean and equals the observed value
* The closer to 1 the R-Squared value is, the better fit the prediction line is
* Null Hypotheses is for linear regression (meaning that there is no correlation as the slope is a straight line.) We reject that hypothesis if the p-value <0.5 meaning we are 95% confident that there is a correlation between the x and y variables.
* How far off is the estimate to the actual

Difference between R-Squared and R-Squared Adjusted

* If predictors variables are added (multivariate) then R-Squared will always increase whereas R-Squared Adjusted will only increase if the added predictor does a better job of predicting the outcome. R-Squared Adjusted values can decrease as other predictors are added.
* Best practice to use R-Squared Adjusted measurement for multivariate linear regression

Output of Summary(model.lm) in R



Below we define and briefly explain each component of the model output:

**Formula Call**

The formula R used to fit the data. Predictor variable (waiting) and the target/response variable (eruptions). The data used to fit the model is the (train) dataset.

**Residuals**

Residuals are essentially the difference between the actual observed response values (eruption time in our case) and the response values that the model predicted. The Residuals section of the model output breaks it down into 5 summary points. When assessing how well the model fit the data, we are looking for a symmetrical distribution across these points on the mean value zero (0). In this example, we can see that the distribution of the residuals appears to be symmetrical as the median is close to 0 and Q1 and Q3 are fairly equidistant from 0. We could take this further consider plotting the residuals to see whether they are normally distributed.

**Coefficients**

In coefficients are the intercept and the slope for each of the predictor variables.

*Coefficient - Estimate*

The coefficient Estimate contains two rows; the intercept and the slope of the predictor variable. In out example, for every minute of waiting in between eruptions, we expect the next eruption to last **0.076** minutes longer

*Coefficient - Standard Error*

The coefficient Standard Error measures the average amount that the coefficient estimates vary from the actual average value of our response variable. We ideally want a lower number relative to its coefficients. In this example, we’ve previously determined that for every minute waiting increase in between eruptions, the eruption time increases by **0.076** minutes. The Standard Error can be used to compute an estimate of the expected difference in case we ran the model again and again. In other words, we can say that the eruption time can vary by **0.0025** minutes. The Standard Errors can also be used to compute confidence intervals and to statistically test the hypothesis of the existence of a relationship between waiting time and eruption time.

*Coefficient - t value*

The coefficient t-value is a measure of how many standard deviations our coefficient estimate is far away from 0. The further away it is from zero the more it indicates the null hypothesis can be rejected – meaning we could declare a relationship between waiting time and eruption time exists. In our example, the t-statistic values are relatively far away from zero and are large relative to the standard error, which could indicate a relationship exists. In general, t-values are also used to compute p-values.

*Coefficient - Pr(>t)*

The *Pr(>t)* acronym found in the model output relates to the probability of observing any value equal or larger than *t*. A small p-value indicates that it is unlikely we will observe a relationship between the predictor (waiting) and response (eruptions) variables due to chance. Typically, a p-value of 5% or less is a good cut-off point. In our model example, the p-values are very close to zero. Note the ‘signif. Codes’ associated to each estimate. Three stars (or asterisks) represent a highly significant p-value. Consequently, a small p-value for the intercept and the slope indicates that we can reject the null hypothesis which allows us to conclude that there is a relationship between waiting time and eruption time.

**Residual Standard Error**

Residual Standard Error is measure of the *quality* of a linear regression fit. Theoretically, every linear model is assumed to contain an error term *E*. Due to the presence of this error term, we are not capable of perfectly predicting our response variable (eruptions) from the predictor (waiting) one. The Residual Standard Error is the average amount that the response (eruptions) will deviate from the true regression line. In our example, the actual eruption time given a waiting time can deviate from the true regression line by approximately **0.523** minutes on average. It’s also worth noting that the Residual Standard Error was calculated with 215 degrees of freedom. Simplistically, degrees of freedom are the number of data points that went into the estimation of the parameters used after taking into account these parameters (restriction). In our case, we had 217 data points and two parameters (intercept and slope).

**Multiple R-squared, Adjusted R-squared**

The R-squared (R2) statistic provides a measure of how well the model is fitting the actual data. It takes the form of a proportion of variance. R2 is a measure of the linear relationship between our predictor variable (waiting time) and our response / target variable (eruption time). It always lies between 0 and 1 (i.e.: a number near 0 represents a regression that does not explain the variance in the response variable well and a number close to 1 does explain the observed variance in the response variable). In our example, the R2 we get is **0.807**. Or roughly **81%** of the variance found in the response variable (eruption time) can be explained by the predictor variable (waiting time).

A side note: In multiple regression settings, the R2 will always increase as more variables are included in the model. That’s why the adjusted R2 is the preferred measure as it adjusts for the number of variables considered.

**F-Statistic**

F-statistic is a good indicator of whether there is a relationship between our predictor and the response variables. The further the F-statistic is from 1 the better it is. However, how much larger the F-statistic needs to be depends on both the number of data points and the number of predictors. Generally, when the number of data points is large, an F-statistic that is only a little bit larger than 1 is already sufficient to reject the null hypothesis (H0 : There is no relationship between waiting time and eruption time). The reverse is true as if the number of data points is small, a large F-statistic is required to be able to ascertain that there may be a relationship between predictor and response variables. In our example the F-statistic is **898.8** which is relatively larger than 1 given the size of our data.

Difference between a Confidence Interval and a Prediction Interval

* A Confidence Interval gives a range of for the given
  + Given value we are 95% confident that what we estimate for will fall into this range
* A Prediction Interval give a range of for the given
  + Given value we are 95% confident that the actual/observed will fall into this range
* A Prediction Interval will always be wider than the Confidence Interval because we are better able to predict the estimated value of more precisely than itself. A Prediction Interval accounts for the variance of the data.
* **Confidence intervals** tell you about how well you have determined the mean. Assume that the data really are randomly sampled from a Gaussian distribution. If you do this many times and calculate a confidence interval of the mean from each sample, you'd expect about 95 % of those intervals to include the true value of the population mean. The key point is that the confidence interval tells you about the likely location of the true population parameter.
* **Prediction intervals** tell you where you can expect to see the next data point sampled. Assume that the data really are randomly sampled from a Gaussian distribution. Collect a sample of data and calculate a prediction interval. Then sample one more value from the population. If you do this many times, you'd expect that next value to lie within that prediction interval in 95% of the samples. The key point is that the prediction interval tells you about the distribution of values, not the uncertainty in determining the population mean.
* Prediction intervals must account for both the uncertainty in knowing the value of the population mean, plus data scatter. So, a prediction interval is always wider than a confidence interval.