Linear Regression - Analysis

We're going to pick up where we left off at the end of the exploration and define a linear model with two independent variables determining the dependent variable, Interest Rate.

Our investigation is now defined as:

Investigate FICO Score and Loan Amount as predictors of Interest Rate for the Lending Club sample of 2,500 loans.

We use Multivariate Linear Regression to model Interest Rate variance with FICO Score and Loan Amount using:

$\text{InterestRate} = a_0 + a_1 * FICOScore + a_2 * LoanAmount$

We're going to use modeling software to generate the model coefficients a_0, a_1 and a_2 and then some error estimates that we'll only touch upon lightly at this point.

```r
# Fit model
fit_linear_model <- lm(Interest.Rate ~ FICO.Score+Loan.Amount, data = df_loansf_1)
```

```r
#Summary:
summary(fit_linear_model)
```

```
##
## Call:
## lm(formula = Interest.Rate ~ FICO.Score + Loan.Amount, data = df_loansf_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0569 -1.7170 -0.2434  1.5324 10.1602
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.288e+01  9.885e-01   73.73   <2e-16 ***
## FICO.Score  -8.844e-02  1.403e-03  -63.02   <2e-16 ***
## Loan.Amount  2.107e-04  6.302e-06   33.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.449 on 2497 degrees of freedom
## Multiple R-squared:  0.6566, Adjusted R-squared:  0.6564
## F-statistic:  2388 on 2 and 2497 DF,  p-value: < 2.2e-16
```

```r
#Analysis of Variance Table:
anova(fit_linear_model)
```

```
## Analysis of Variance Table
##
## Response: Interest.Rate
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## FICO.Score     1 21937.1 21937.1  3656.7 < 2.2e-16 ***
## Loan.Amount    1  6709.5  6709.5  1118.4 < 2.2e-16 ***
## Residuals   2497 14979.9     6.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So we have a lot of numbers here and we're going to understand some of them.

Coefficients: contains a_1 and a_2 respectively. Intercept: is the a_0.

How good are these numbers, how reliable? We need to have some idea. After all we are estimating. We're going to learn a very simple pragmatic way to use a couple of these.

Let's look at the second two numbers. We are going to talk loosely here so as to give some flavor of why these are important. But this is by no means a formal explanation.

P-Values are probabilities. Informally, each number represents a probability that the respective coefficient we have is a really bad one. To be fairly confident we want this probability to be close to zero. The convention is it needs to be 0.05 or less. For now suffice it to say that if we have this true for each of our coefficients then we have good confidence in the model. If one or other of the coefficients is equal to or greater than 0.05 then we have less confidence in that particular dimension being useful in modeling and predicting.

R-squared or R^2 is a measure of how much of the variance in the data is captured by the model. What does this mean? For now let's understand this as a measure of how well the model captures the spread of the observed values not just the average trend.

R is a coefficient of correlation between the independent variables and the dependent variable - i.e. how much the Y depends on the separate X's. R lies between -1 and 1, so R^2 lies between 0 and 1.

A high R^2 would be close to 1.0 a low one close to 0. The value we have, 0.65, is a reasonably good one. It suggests an R with absolute value in the neighborhood of 0.8. The details of these error estimates deserve a separate discussion which we defer until another time.

In summary we have a linear multivariate regression model for Interest Rate based on FICO score and Loan Amount which is well described by the parameters above.