MA4710 PROJECT 1

DANIEL HENDERSON

Introduction

In rock climbing, the motivation for many is too climb routes that are at the limit of their ability. To do this, we must grade each established route, which gives the climber a way of ranking his routes that he has climbed. The V-grading system is used to rate the difficulty of boulder problems, which are routes that are established on boulders. The V-Grading system is ordinal in nature, taking on values ranging from V0 to V17. But the difficulty difference between two different grades equally spaced apart is assumed to be the same. That is, the jump in difficulty from V0 to V2 is the same as the jump in difficulty from V2 to V4. With this assumption we can convert V0 to 0, V1 to 1, and so on and then use the transformed data as a numeric response variable in a linear regression model.

Initially, the first acentionist who opens a boulder proposes a V-grade after successfully climbing the route. Over time as more people climb the boulder, the route gets its grade adjusted to reflect the opinion of the masses. This is most commonly done through an app called mountain project, which is a database of climbing routes across the world. After a climber completes a route, he can go into mountain project and tick the route. In doing so, the climber proposes a new V-grade based upon his experience with the boulder problem. Mountain Project will then report the grade of the route as the Median of all reported grades they have received from the climbers who have ticked the route. Thus, overtime the V-grade assigned to a route should accurately reflect its difficulty.

Data and Data Cleaning

My data set is from a survey that collected basic information pertaining to an individuals climbing ability. The survey was posted to the /climbharder subreddit, which is Reddit's rock climbing training community. The survey had 551 responses each having 31 variables associated with it.

One of the variables is the climbers hardest V grade they have ever successfully climbed, and as mentioned earlier, this is our response variable in the linear regression model. However, the V-Grading system is not as consistent when applied to indoor climbs. It is the general consensus that the V-Grade applied to gym routes is in general an overestimate of its difficulty if the moves where to take place on rook. Therefore, the first step I did was to remove all observations from my data where participants reported that they only climb indoors.

My first predictor is another categorical variable in the survey was the amount of time that the participant has been climbing. The participant appears to have been asked to select how long they have been climbing from options ranging from $0-0.5$ years, $0.5-1$ years, $1-1.5$ years and so on up until $14.5-15$ years. Then there was a final option for the user to select that they have been climbing for 15+ years. I wanted to use this variable as a numeric predictor in my model, so I transformed the data in the following manner:

$$0-0.5 \text{ years} \to 0 \text{ years}, \ 0.5-1 \text{ years} \to 0.5 \text{ years}, \ \ldots \text{ and } 15+ \text{ years} \to 15 \text{ years}.$$

I used lower bound on each interval because the participant has atleast been climbing as long as the lower bound of the selected interval. As with anything in life, the more experience climbing you have the better you should be at climbing. Therefore, I suspect that this should be a good predictor in my model

Another numeric variable in the survey that the participant was asked to report was their maximum number of pull up reps that they can complete. I believe that this will be a strong predictor in my model. This belief comes from my experience with climbing, the more times you can pull your body weight up with your hands certainly helps. However, since this is a technical question many participants responded with answers such as, "idk" or "20 - 30?". So my next step in cleaning was to remove all observations that didn't provide a specific integral value. This approximately reduced the number of observations in my data set to half its size.

Data and Data Cleaning Continuted
The three other predictors that where used where the participants Height, Weight, and Ape Index (arm span). These values where reported in the metric system and I have choosen to leave them as such out of convience. However, there where numerous observations that reported answers to these questions in the US Customary system so I converted these to the metric system appropriately. I selected these predictors fairly niavely, but suspect that they will contribute in explaining the observed variance in our response. After cleaning the data, I was left with 284 observations

Analysis
The analysis of our data was performed in R and can be found in the script draft.R. Our linear model is,

$$Y \cong \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

where,

Y: Hardest V-Grade.
X_1: Height (cm)
X_2: Weight (kg)
X_3: ApeIndex (cm)
X_4: Years spent climbing
X_5: Max pull-up reps
β_i: Coefficient corresponding the X_i predictor

From, now on when we refer to the V-Grade of an individual we are talking about the Hardest V-Grade the the individual climber has sent. Then using linear algebra I obtained the least-squared solution for our model to be,

$$\boldsymbol{\beta} = \begin{bmatrix} 6.049368 \\ -0.02512294 \\ -0.01207272 \\ 0.02027981 \\ 0.1994149 \\ 0.1108341 \end{bmatrix} \implies Y \cong 6.049368 - 0.02512294 X_1 - 0.01207272 X_2 + 0.02027981 X_3 + 0.1994149 X_4 + 0.1108341 X_5$$

Below is a table of the t-statistics, standard errors, and p-values in each of the tests $H_0 : \beta_j = 0$ against $H_a : \beta_j \neq 0$

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.049368   1.890526   3.200  0.00153 **
Height        -0.025123   0.010801  -2.326  0.02073 *
Weight        -0.012073   0.006283  -1.921  0.05570 .
ApeIndex       0.020280   0.006653   3.048  0.00252 **
YearsClimbing  0.199415   0.023770   8.389 2.41e-15 ***
MaxPullUpReps  0.110834   0.011207   9.889  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.771 on 281 degrees of freedom
Multiple R-squared:  0.4228,    Adjusted R-squared:  0.4125
F-statistic: 41.16 on 5 and 281 DF,  p-value: < 2.2e-16
```

Analysis Continued

The next step is to interpret the results shown in the table on the previous page. The coefficient for the Weight predictor is the only one that does not have a p-value less than 0.05 therefore we will disregard further discusion of the β_3. We interpret the other coefficeints as follows:

β_1(units V-Grade/cm): For every one cm increase in Height, we predict the mean V-Grade will decrease by 0.02512294 while holding all other factors constant.

β_2(units V-Grade/Kg): For every one Kg increase in Weight, we predict the mean V-Grade will decrease by 0.01207272 while holding all other factors constant.

β_4(units V-Grade/years): For every year spent climbing, we predict the climbers mean V-Grade will increase by 0.1994149 while holding all other factors constant.

β_5(units V-Grade/pull-up reps): For every increase by one in max-pull up reps, we predict the climbers mean V-Grade will increase by 0.1108341 while holding all other factors constant.

The results from performing an overall regression test of our linear model are shown in the table on the previous page. The results align with what I determined by hand, where my F-Statistic was 41.16272 and having an associated p-value of $1 \cdot 10^{-31}$. Note that this p-value is not the same as the one that was reported from the lm function in R is $2e-16$. So we know our observed significance level of this test is very small, but what is the observed significance level? I hypothesize that the difference in the two answers are likely due to my explicit use of the normal equations to determine $\boldsymbol{\beta}$. The normal equations produce a relative error proportional to the square of the condition number of X. That is, the normal equations have a much worse conditioning than the origonal LS-problem which implies that using them may produce inaccurate results. I imagine the lm function, determines $\boldsymbol{\beta}$ using a more stable algorithm such as the QR factorization algorithm using householder transformations. Thus, I have more faith in $2e-16$ being more reflective of the actual p-value associated with the F-statistic.

Our model has a coefficient of determination of $R^2 = 0.4228$ which implies that 42.28% of the observed variability in our response is explained by our linear model. We determined the variance of our residual vector to be $\hat{\sigma}^2 = 3.135234$, which is our point estimate of σ^2. We use $\hat{\sigma}^2$ to obtain $\sigma = 1.770659$ V-Grade, which allows us to create a confidence interval around our predictions when using the model.

When looking at our coefficients, it appears that max pull-up reps (X_5) and years spent climbing (X_4) play the most significant roles in our model. Therefore, we performed the test $H_0 := \beta_1 = \beta_2 = \beta_3 = 0$ against H_a : not H_0. That is, we are testing the full model against $Y = \beta_0 + \beta_4 X_4 + \beta_5 X_5$. The p-value of this test is 0.02073. The coefficient of determination for the reduced model is 0.4125061, so we say that 1.03% of the variability in the response is explained by the predictors X_1, X_2, X_3.

Next, we hypothesize that $\beta_3 = 0.5$ and test it agianst $H_a : \beta_3 \neq 0.5$. That is, we think that a 2cm increase in ape index should correspond to an increase in 1 grade of the mean V-grade. This test had a p-value of $2.2e-16$ which means we reject the null and conclude that we are 95 percent sure that $\beta_3 \neq 0.5$.

And Lastly, we hypothesize that $\beta_1 = \beta_2$ and test it against its negation. This test says that a change in height of 1 cm will have the same impact on the mean V-grade as a 1 kg change in height. This test had a p-value of 0.02073 which means we reject the null and conclude that we are 95 percent sure that $\beta_1 \neq \beta_2$.

Running the script draft.R will output all of the information referenced in this report.