CS 555 Final Project

Predicting Max Climbing Grade

Tatiana Ediger, Jiankun Dong, Grigoriy Gressel

Dec 3, 2023

Abstract

We are creating a linear regression model for predicting climbing grade based on age, sex, height, weight, and years spent climbing. We want to see which variables/characteristics are either good or poor predictors of the climbing grade. Initially looking into the data we can see that all the predictive variables are pretty normally distributed which is needed for a linear regression model. We ran a Global F-test to determine whether we should reject or not reject the null hypothesis that "there is no linear relationship between the response and explanatory variables". Based on the Global F-test, we concluded that we can reject the null hypothesis and claim the overall model is significant. We then dug deeper into the data and found that we can confirm that each individual explanatory variable is significant based on the p-values at the $\alpha = 0.05$ level. Based on the model each variable is a significant predictor. However, even though the model and each variable is significant we can only explain 22.69% of the variation in the climbing score.

Research Scenario + Statistical Methods

Climbing as a sport has blown up in the past few years, with lots of people working towards higher grades. There are a few grading conventions, and this dataset uses the French grading convention, translated into a number from 0 to 84. Our goal is to see whether we can predict a user's maximum climbing grade based on their age, sex, height, weight, and years spent climbing, and whether each of them are significant predictors.

In order to answer this question - we want to first test the overall model with a Global F-test:

- H_0: $\beta_1 = \beta_2 = \dots = \beta_k = 0$ there is no linear relationship between the response and explanatory variables.
- H_0: $\beta_i \neq 0$ for at least one i, or that at least one of the slope coefficients is different from zero.

Then, if the overall model is significant, we will perform testing on each parameter to verify that each independent variable is a significant predictor by running pairwise t-tests. We will test all of this at the $\alpha = 0.05$ level.

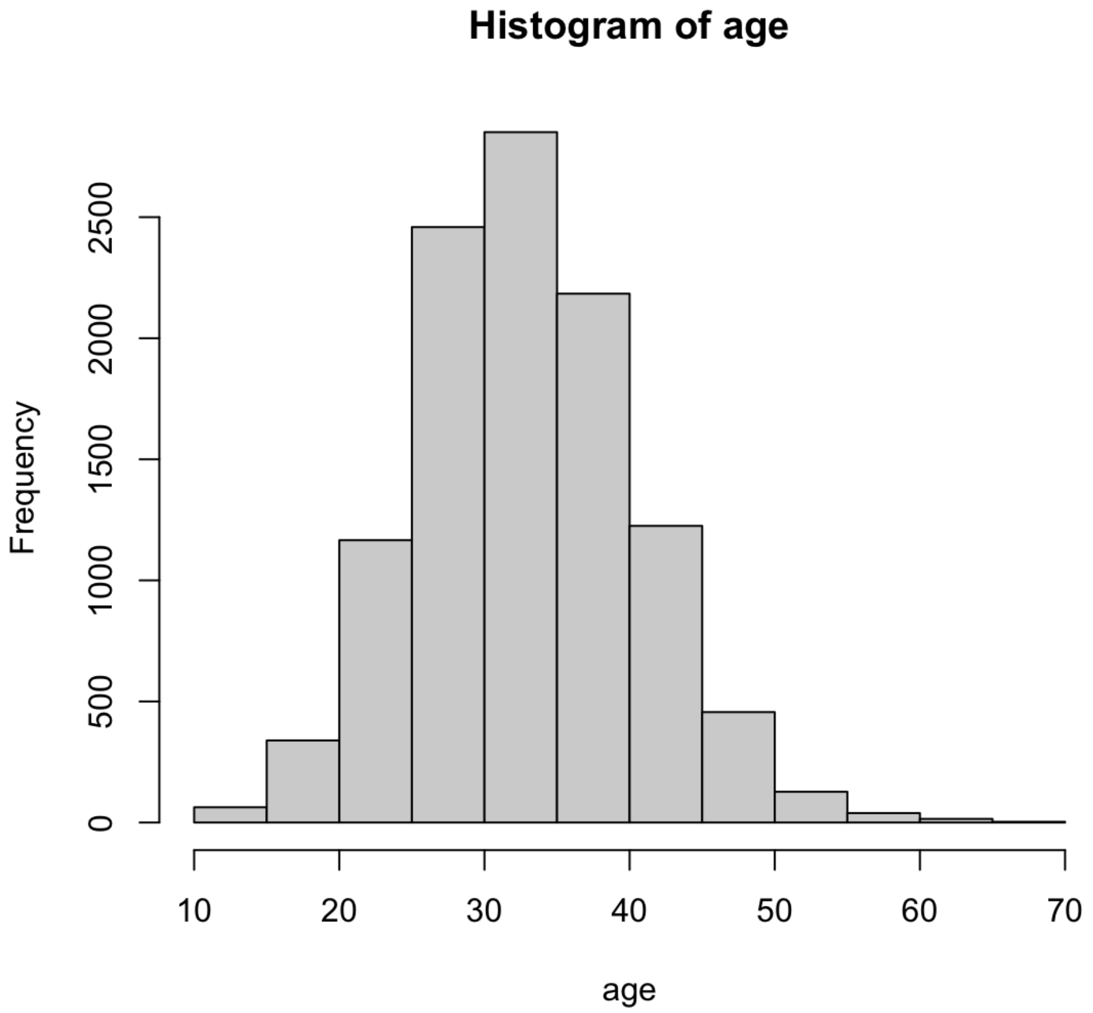
The Dataset

Dataset link: https://www.kaggle.com/datasets/jordizar/climb-dataset/?select=climber_df.csv

We are using a dataset with 10927 rows. The columns we are using are Age, Sex, Height, Weight, Years Climbed, and Max Grade.

Here are the distributions of all of these columns:

Age



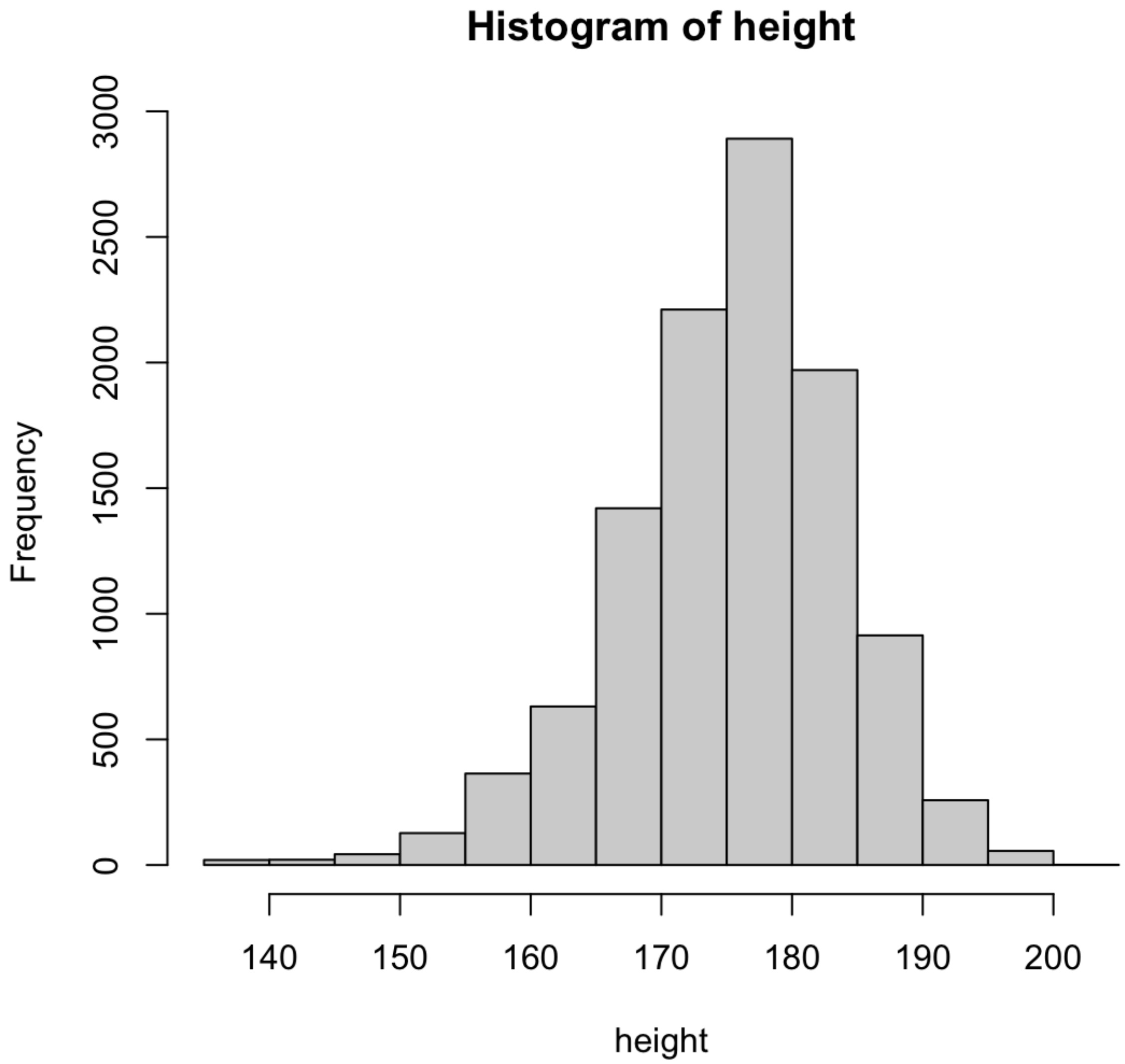
```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.00   28.00   33.00   33.33   38.00   69.00
```

Age is normally distributed around 33 years old.

Sex

87% of the dataset is Male (0), and 13% of the dataset is Female (1).

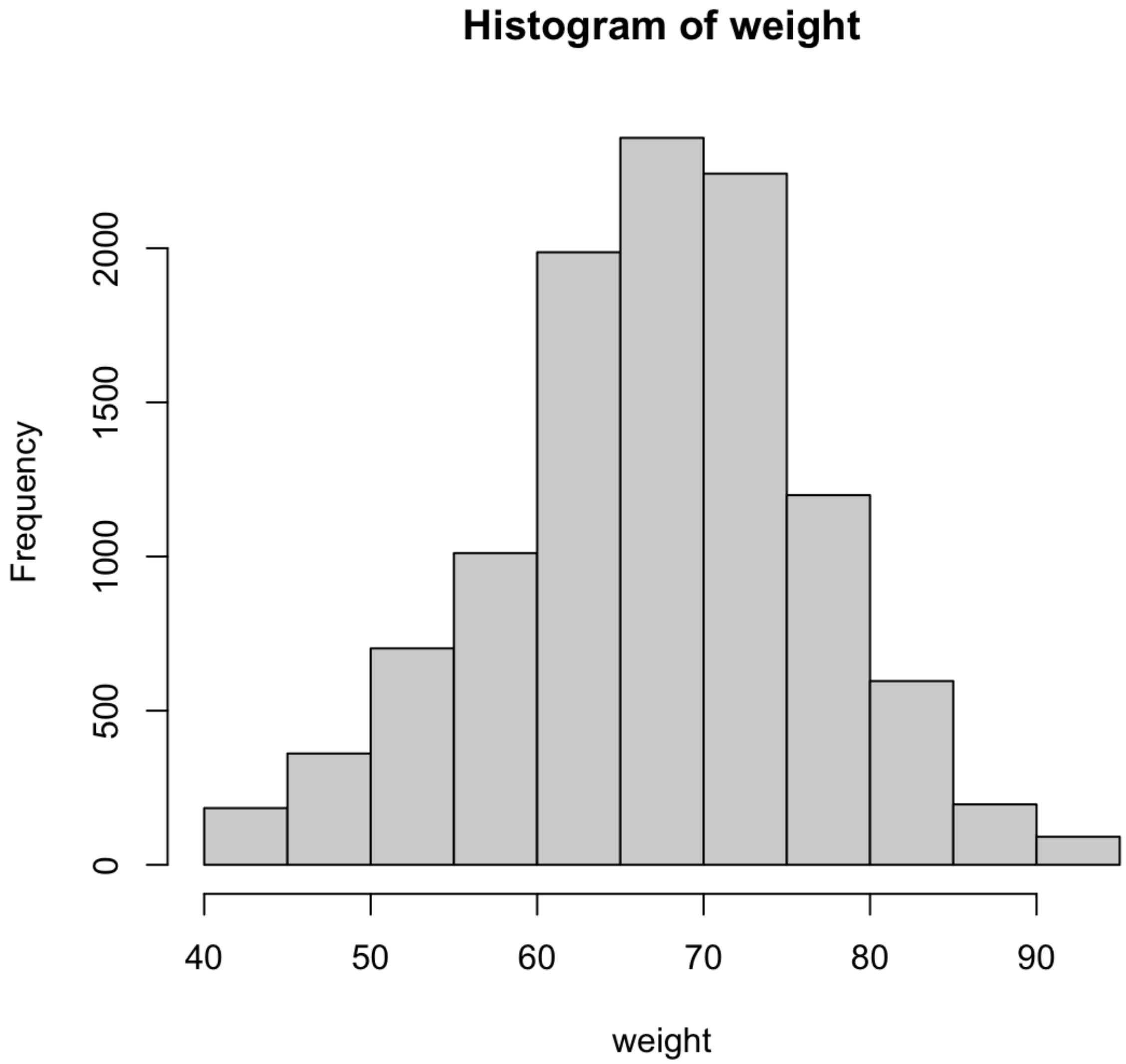
Height



```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 137.0   171.0   177.0   176.2   182.0   202.0 
```

Height is normally distributed around 176 cm (~5’9”) with a slight left skew.

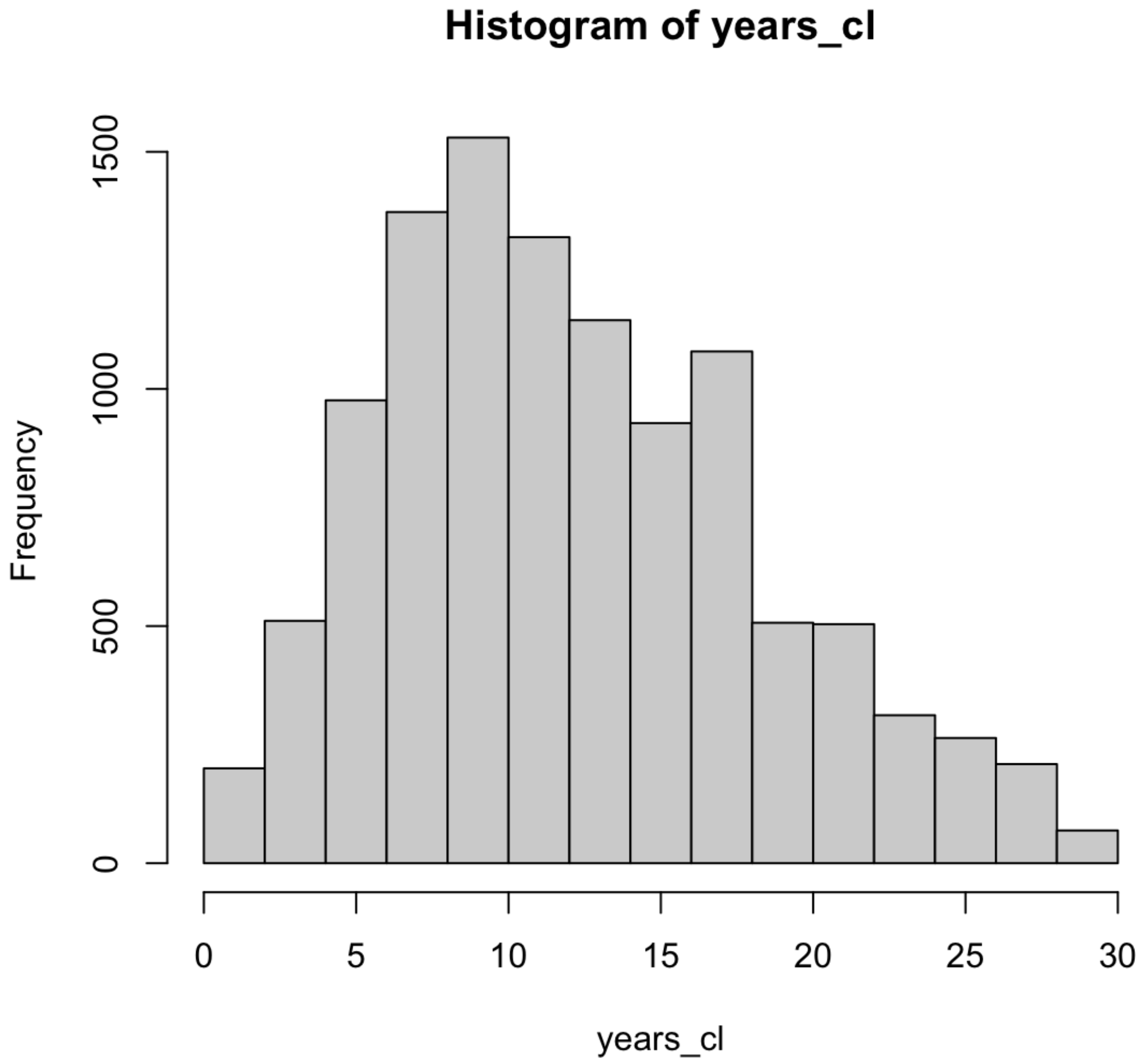
Weight



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  40.00   63.00   68.00   67.61   73.00   93.00
```

Weight is normally distributed around 67 kg (147 lbs).

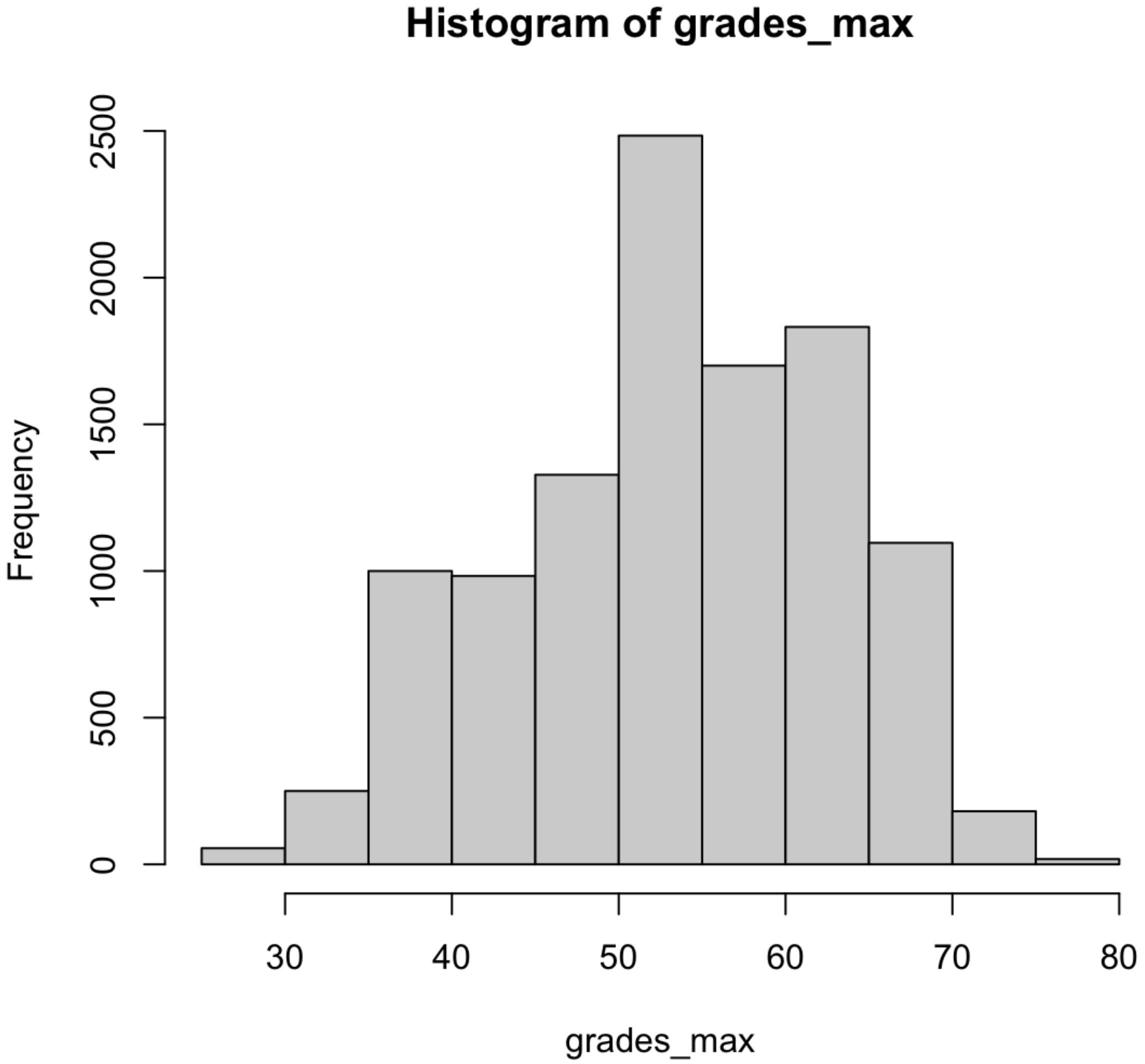
Years Climbing



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.00    8.00   12.00   12.67   17.00   29.00 
```

Years climbing is roughly normally distributed around 12.67 years, with a slight right skew.

Max Climbing Grade



```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 29.00   46.00   55.00   53.76   62.00   77.00 
```

Max climbing grade is roughly normally distributed around 53.76, with a slight left skew.

Data Preparation + Model Building

In order to prepare the data, we first check for NAs. Once we have confirmed there are no NAs in the dataset, we can proceed to build our model.

Our model is:

$$Max\ Grade\ =\ (-\ 0.357\ *\ AGE)\ +\ (-\ 5.511\ *\ SEX)\ +$$
$$(\ 0.091\ *\ HEIGHT)\ +\ (-\ 0.252\ *\ WEIGHT)\ +\ (0.749\ *\ YEARS\ CLIMBING)$$

Interpreting our model:

- Every one year increase in age, leads to a -0.357 decrease in max climbing grade.
- Females (1) tend to have a max grade 5.511 lower than men.
- Every one cm increase in height leads to an 0.091 increase in max climbing grade.
- Every one kg increase in weight leads to a -0.252 decrease in max climbing grade.
- Every one year increase in years climbing, leads to a 0.749 increase in max climbing grade.

Formal Testing Results

Global F-test:

(1) Set up hypotheses + alpha level:
   (a) H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$
   (b) H_1: $\beta_i \neq 0$ for at least one i
   (c) α = 0.05

(2) Select appropriate test statistic:
   (a) $F = \frac{MS_{reg}}{MS_{res}}$ with 5 and 10921 degrees of freedom.

(3) State the decision rule:
   (a) F-distribution with 5, 10921 degrees of freedom and associated with α = 0.05
   (b) $F_{5,10921,,0.05}$= 2.215
   (c) Decision Rule: Reject H_0 if F ≥ 2.215, else do not reject H_0

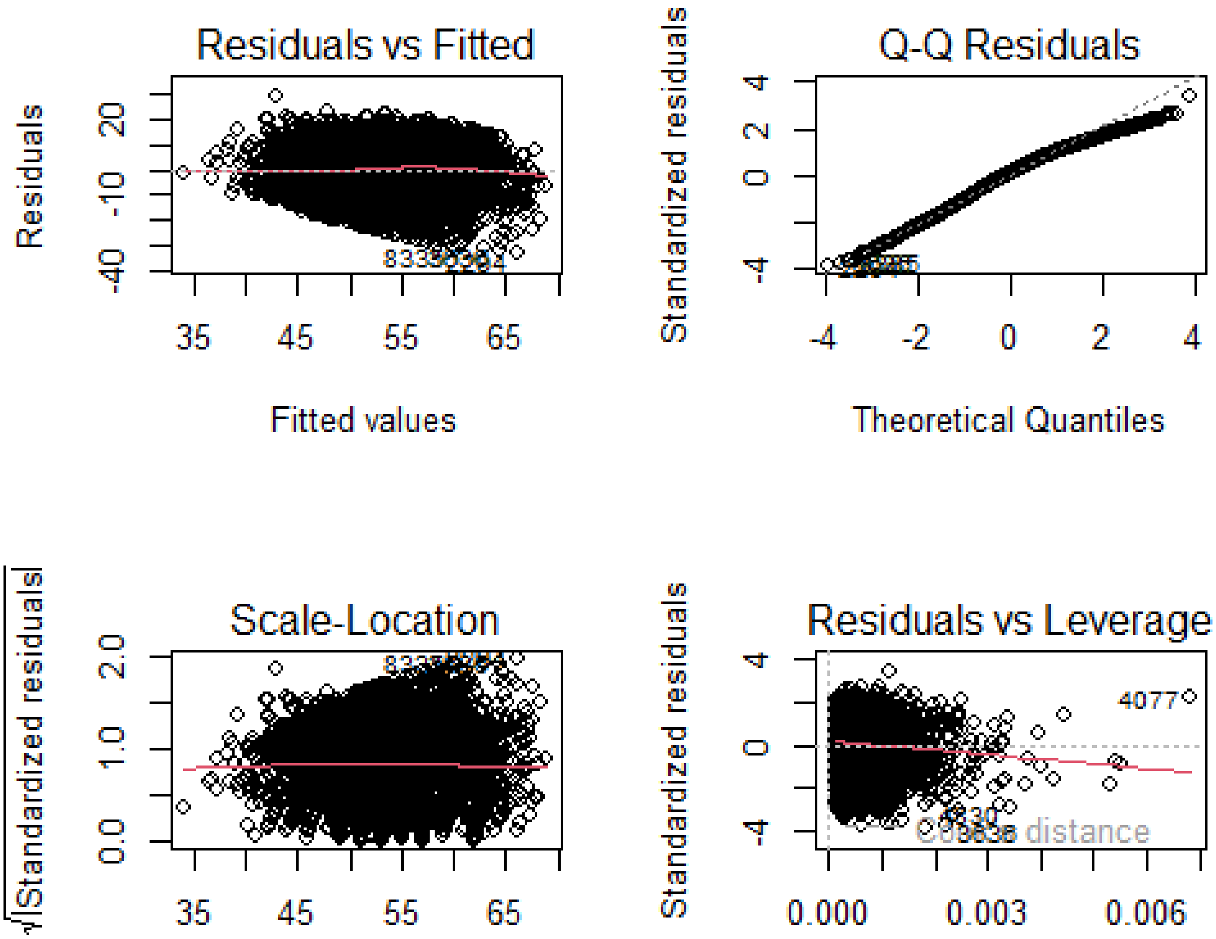
(4) Compute the test statistic:
   (a) F-statistic = 642.4

(5) State conclusions:
   (a) Reject H_0 since F=642.4 ≥ 2.215. We have significant evidence at the α = 0.05 level that $\beta_i \neq 0$ for at least one i.

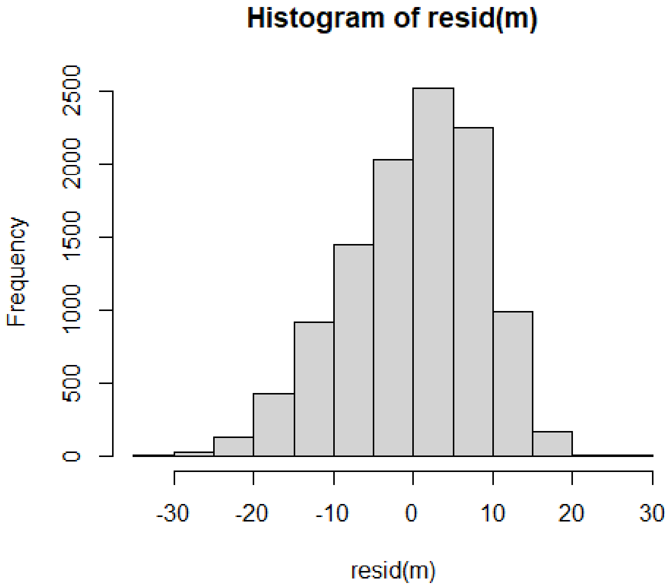
Pairwise t-tests:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	57.97211	2.17821	26.615	< 2e-16 ***
age	-0.35682	0.01317	-27.097	< 2e-16 ***
sex	-5.51081	0.30155	-18.275	< 2e-16 ***
height	0.09050	0.01482	6.106	1.06e-09 ***
weight	-0.25226	0.01367	-18.447	< 2e-16 ***
years_cl	0.74885	0.01583	47.312	< 2e-16 ***

Since all p-values for each of the features (age, sex, height, weight, and years climbing) are all < 0.05, then we have significant evidence at the $\alpha = 0.05$ level that each of the features $\neq 0$ after controlling for the other features.

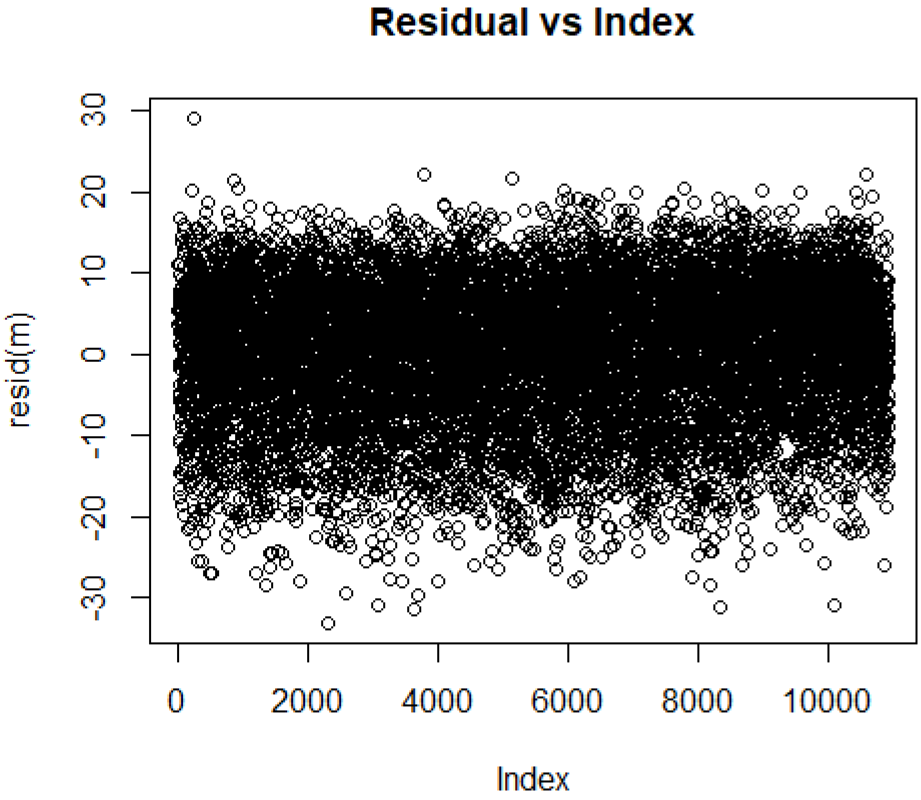
Checking Assumptions



Histogram of the residuals:



Scatter plot of residuals:



Linearity:
From the “Residuals vs Fitted” plot, we can see that the fitted value is roughly following a ‘linear trend’.
The linearity assumption has been satisfied.

Independent:
From the plot of “Residual vs Index”, we don’t see an obvious trend. The independent assumption has been satisfied

Normality:
From the histogram of the residual, we can see that the residuals roughly follow a normal distribution, with a slight left screw.
From the Q-Q plot of residuals, we can see that the residuals roughly on the line, which indicates that residuals roughly follow the normal distribution.
The normality assumption has been satisfied.

Equal variance:
From the “REsiduals vs Fitted” plot, we can see that the variance of the response slighting increases with the fitted value. This indicates that there is a slight violation of the equal variance assumption.

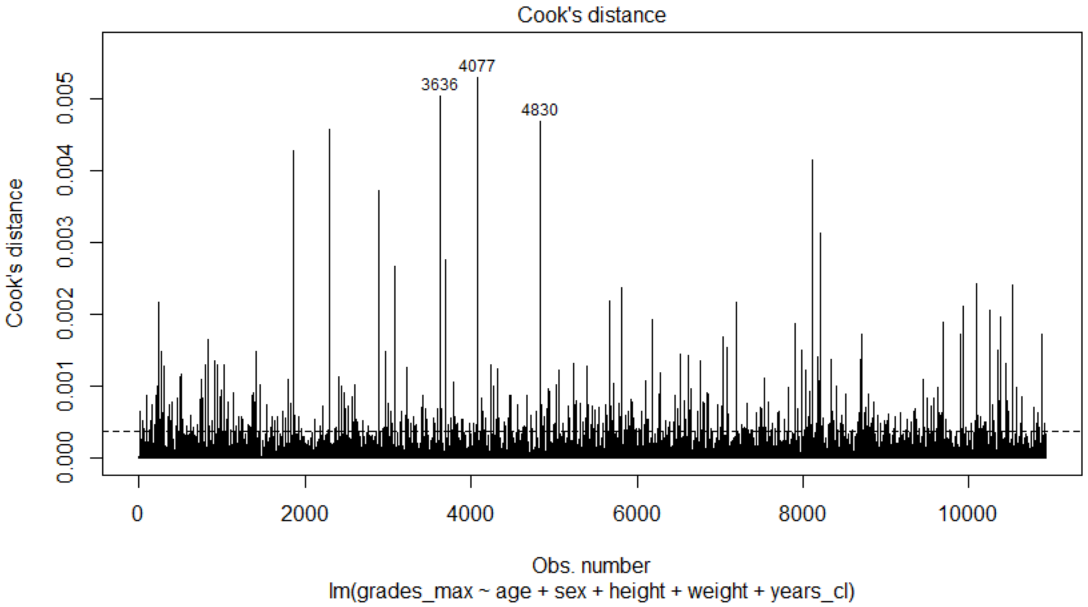
Outliers and Influence points

Outlier test:

```
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
      rstudent unadjusted p-value Bonferroni p
2294 -3.963824         7.4228e-05      0.81109
```

There is no outlier.

Influence points:



Using the Cook's distance in the plot above with the cutoff shown, there are many influence points in the data set. If we were to remove them, the new model would be significantly different, and our sample size wouldn't be a good representation of the population. We have decided to leave those influence points in the dataset and take it into consideration when evaluating the model because this is a true representation of the population.

Conclusion + Limitations

From the dataset, the multiple linear regression model is:

$$Max\ Grade = (-0.357 * AGE) + (-5.511 * SEX) + (0.091 * HEIGHT) + (-0.252 * WEIGHT) + (0.749 * YEARS\ CLIMBING)$$

Based on the adjusted coefficient of determination value from the summary of the model, the regression model only explains only 22.69% of the variation in the climbing score. However, each variable is a significant predictor of the output of the overall model. The relatively low coefficient of determination is not surprising, based on the large number of influence points that exist in the dataset. There is also a slight violation of the “constant variance” assumption. Perhaps significantly increasing the sample size will help satisfy the “constant variance” assumption and decrease the amount of influence points, which might lead to a better model with higher coefficient of determination.

R Code

Unset

```
## DATA PREPARATION
library(car)
# Read in the climbing dataset
df <- read.csv('climber_df.csv')
attach(df)

# Select subset of columns
df <- data.frame(df$sex, df$age, df$weight, df$height, df$years_cl,
df$grades_max)

# Describe the data
df$df.sex <- as.factor(df$df.sex)
sex <- as.factor(sex)
summary(df)

# Check for NA
sum(is.na(df))
## there’s no NA for the current data, run the following line if there’s NA
# df <- na.omit(df)
length(df$df.sex)

hist(age)
summary(age)

hist(height)
summary(height)

hist(weight)
summary(weight)

hist(years_cl)
summary(years_cl)

hist(grades_max)
```

```
summary(grades_max)

## BUILDING MODEL
m <- lm(grades_max~age+sex+height+weight+years_cl, data=df)
coef(m)

## STATISTICAL ANALYSIS

# Global F-test
summary(m)
qf(0.95, df1=5, df2=10921)

#pairwise t-tests
summary(m)

## CHECK ASSUMPTIONS
## Residual plots, check for linearity, equal variance
par(mfrow = c(2,2))
plot(m)
## histogram of residuals, check for normality
hist(resid(m))
## Independent: plot residual vs index
plot(resid(m), main = "Residual vs Index")

## Check for outliers
outlierTest(m)
## Check for influence point
b <- influence.measures(m)
c <- which(apply(b$is.inf, 1, any))
cd <- cooks.distance(m)
par(mfrow=c(1,1))
plot(m, which = 4)

## rule of thumb
cutoff = 4/nrow(df)
abline(h=cutoff,lty=2)
```