Stat 364 Lab 1

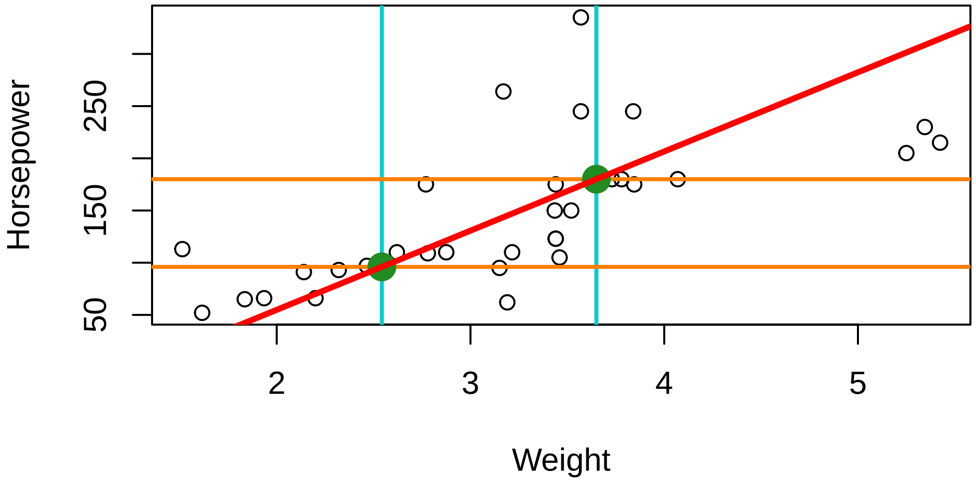
Brad Staples

Question 2

Create a a scatterplot of vehicle weight and horsepower from the mtcars data frame. Overlay your plot with a median-median line in red.

```
## this is an example, your code will involve finding the proper
##    slope and intercept, not just putting in two numbers
attach(mtcars)
plot(wt, hp, xlab="Weight", ylab="Horsepower")

median_wt<-median(wt)
grp1<-wt<median_wt

med_wt1<-median(wt[grp1])
med_wt2<-median(wt[!grp1])

med_hp1<-median(hp[grp1])
med_hp2<-median(hp[!grp1])

abline(v=med_wt1, col="cyan3", lwd=2)
abline(v=med_wt2, col="cyan3", lwd=2)

abline(h=med_hp1, col="darkorange1", lwd=2)
abline(h=med_hp2, col="darkorange1", lwd=2)

points(c(med_wt1, med_wt2), c(med_hp1, med_hp2),pch=16, cex=2, col="forestgreen")

rise<-med_hp2-med_hp1
run<-med_wt2-med_wt1
slope<-rise/run
```

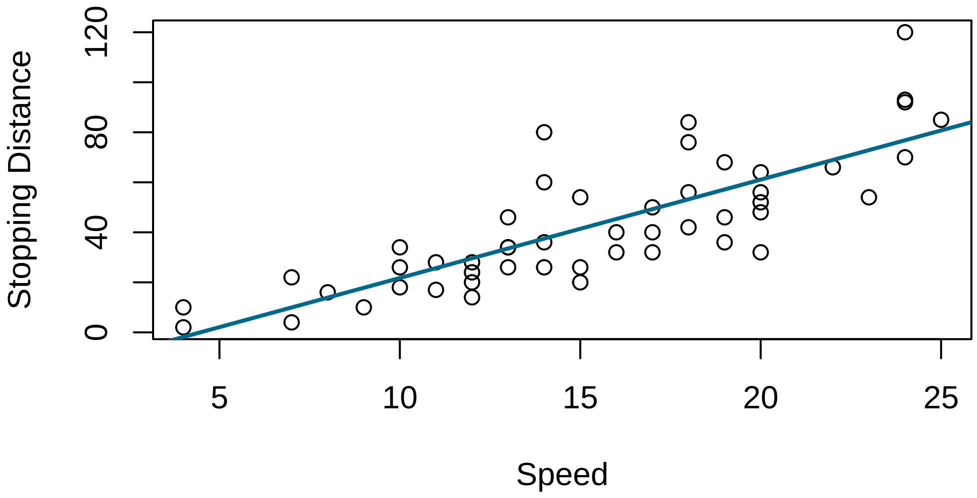
```
intercept = med_hp1-(slope*med_wt1)
abline(intercept, slope, col="red", lwd=3)
```



Question 3

Create a scatterplot of vehicle speed and stopping distance from the cars data frame. Overlay your plot with a least-squares regression line in the color of your choice.

```
attach(cars)
plot(speed, dist, xlab="Speed", ylab="Stopping Distance")

reg<-lm(dist~speed)
abline(reg, col="deepskyblue4", lwd=2)
```



```r
reg$coefficients
```

```
(Intercept)       speed
 -17.579095    3.932409
```

Question 5

Fit a least-squares model predicting vehicle hwy_mpg as a function of vehicle weight using the cars04 data frame in the openintro library. Which estimate, d or e, do you feel better about? Why?

I feel more confident in C as an estimate than in D because 3,802 lbs (our c datapoint) is closer to the center of the dataset used to fit our model. Predictions made using data near the middle of the observed range are generally more reliable for estimating highway MPG at a given weight. Conversely, 1,700 lbs is near the edge of the dataset, making predictions at that weight more uncertain and potentially affected by extrapolation.

Question 6a

```
median_median <-function(x,y){

  median_wt<-median(x)
  grp1<-x<median_wt

  med_x1<-median(x[grp1])
  med_x2<-median(x[!grp1])

  med_y1<-median(y[grp1])
  med_y2<-median(y[!grp1])

  rise<-med_y2-med_y1
  run<-med_x2-med_x1
  slope<-rise/run
  intercept <- med_y1-(slope*med_x1)
  #print(slope)
  #print(intercept)
  return(list(slope=slope, intercept=intercept))
}
#testing data for the function
X <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12)
Y <- c(2, 4, 5, 7, 6, 8, 9, 11, 12, 10, 14, 15)
median_median(X, Y)
```

```
$slope
[1] 1

$intercept
[1] 2
```

Question 6b

```
median_median(x=cars$speed, y=cars$dist)
```

```
$slope
[1] 4

$intercept
[1] -22
```

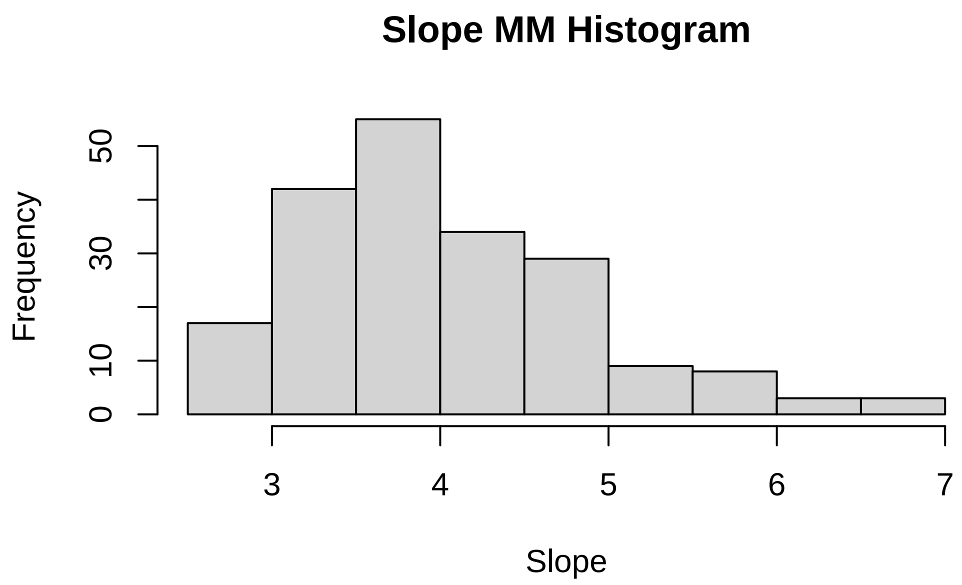
Question 6c

```r
rows_sample <- sample(1:nrow(cars), replace=TRUE)
sampled_result<-median_median(x=cars$speed[rows_sample], y=cars$dist[rows_sample])
original_result<-median_median(cars$speed, cars$dist)
```

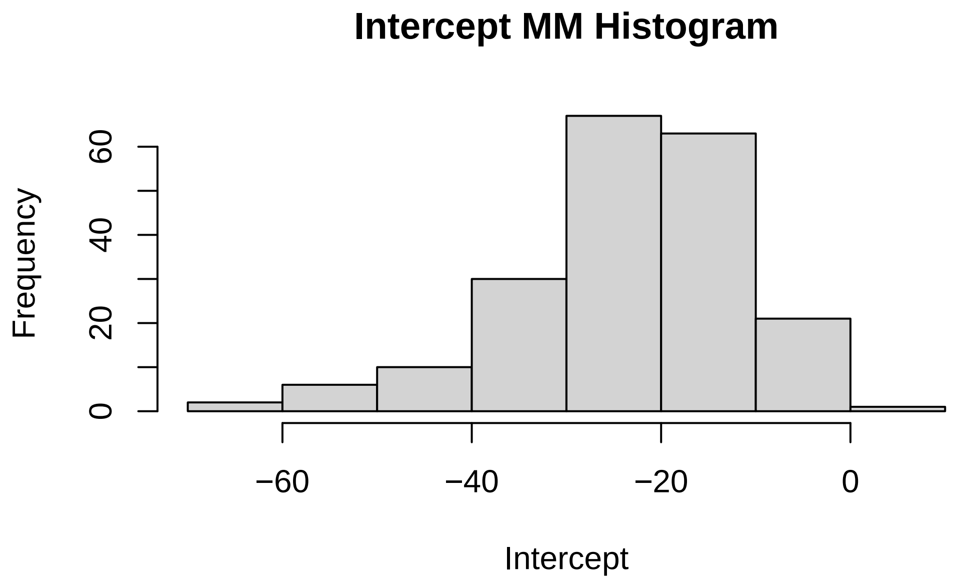
When comparing the median-median line between the original and the sampled data, the sampled slope tends to remain close to the original value but does show some slight variability. In my runs, the slope was often in the range of 3–4, but sometimes as high as 6. For example, the original slope was 4 and the sampled slope was 5.7142857 The intercept shows even greater variability than the slope, with my personal tests on the sampled intercept ranging from -12.5 to -38 compared to the original value of -22. As another example in action, original intercept was -22 compared to our sampled intercept was -39.7142857. This demonstrates the variability in sampling random data and how we can get many different estimates of slope and intercept from the same, singular data set, all depending on the sample we randomly generate.

Question 6d

```r
slope_grp<-c()
intercept_grp<-c()
for(i in 1:200){
  rows_sample <- sample(1:nrow(cars), replace=TRUE)
  results<-median_median(x=cars$speed[rows_sample], y=cars$dist[rows_sample])
  slope_grp<-c(slope_grp, results$slope)
  intercept_grp<-c(intercept_grp, results$intercept)
}
hist(slope_grp, main ="Slope MM Histogram", xlab="Slope")
```



```
hist(intercept_grp, main ="Intercept MM Histogram", xlab="Intercept")
```



The slope histogram has a fairly tight spread, with most values between 3 and 5 and a peak around 3.5–4, which is close to the original slope. There are a few outliers between 6 and 8, but those are rare. This shows that the slope is relatively stable across samples.

The intercept histogram shows a much wider spread of values compared to the slope, with most intercepts falling between -80 and 10 and the bulk clustered around -40 to -10. This indicates that the intercept is much more variable than the slope in this dataset.

Question 7a

```
lm_reg <- lm(cars$dist~cars$speed)
coef(lm_reg)
```

```
(Intercept)  cars$speed 
 -17.579095    3.932409 
```

Question 7b

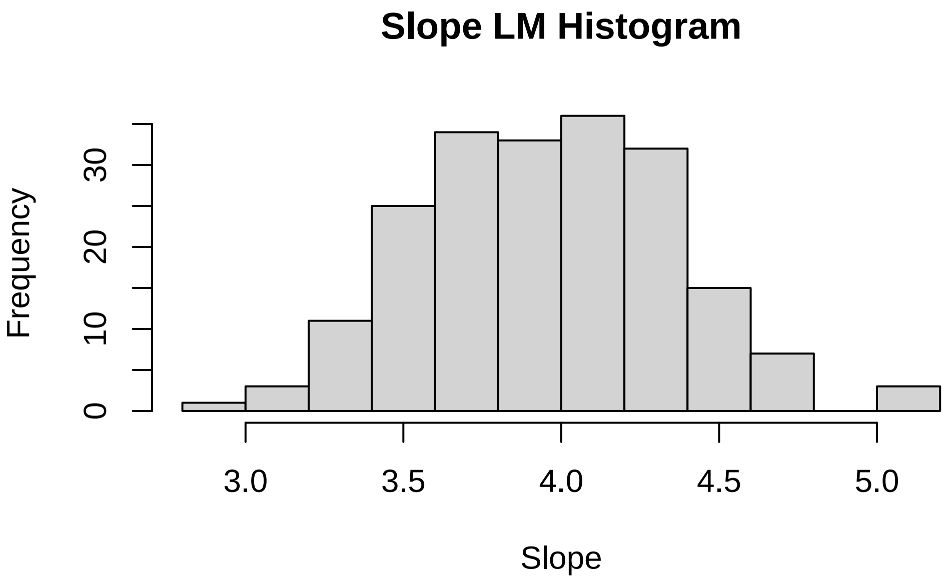
```
rows_sample <- sample(1:nrow(cars), replace=TRUE)
lm_sample<-lm(cars$dist[rows_sample]~cars$speed[rows_sample])
coef(lm_sample)
```

```
            (Intercept) cars$speed[rows_sample] 
             -28.166269                4.546295 
```

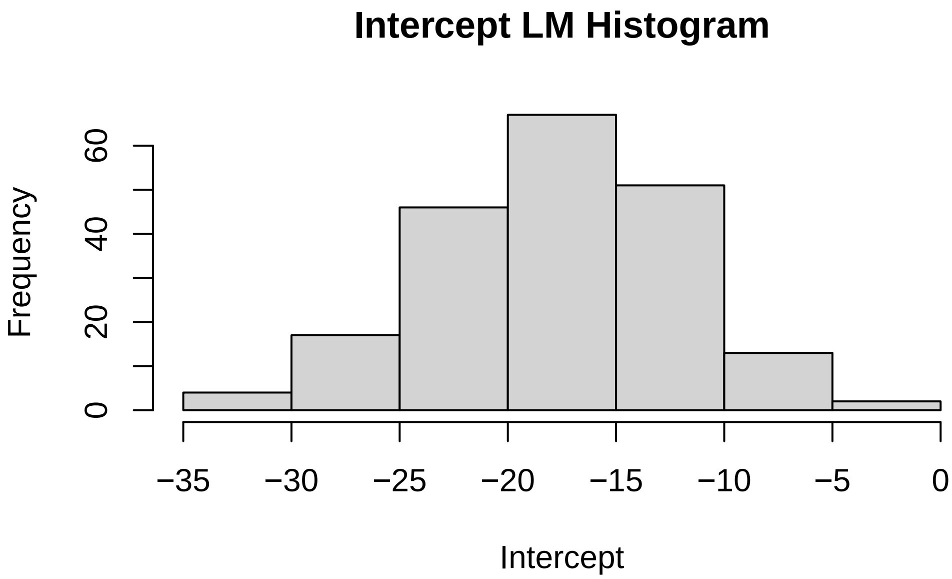
Much like our last sampling runs with the median-median lines, the sample data show a greater variability in the intercept than in the slope. Our original dataset had a slope of 3.9324088 and an intercept of -17.5790949 and the sampled data from one of the runs had a slope of 4.5462946 and an intercept of -28.1662689.

Question 7c

```
slope_grp_lm<-c()
intercept_grp_lm<-c()
for(i in 1:200){
  rows_sample <- sample(1:nrow(cars), replace=TRUE)
  lm_result<-lm(cars$dist[rows_sample]~cars$speed[rows_sample])
  slope_grp_lm<-c(slope_grp_lm, lm_result$coefficients[2])
  intercept_grp_lm<-c(intercept_grp_lm, lm_result$coefficients[1])
}
hist(slope_grp_lm, main ="Slope LM Histogram", xlab="Slope")
```



```
hist(intercept_grp_lm, main ="Intercept LM Histogram", xlab="Intercept")
```

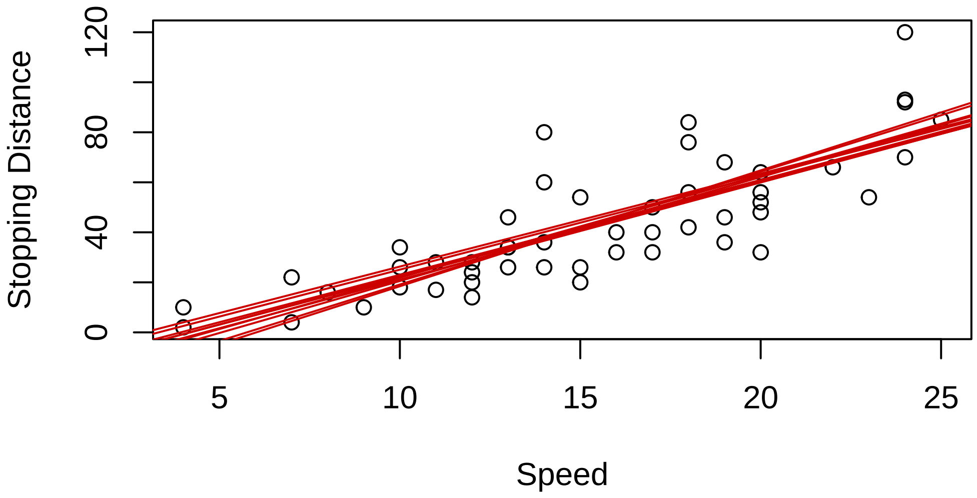


The slope histogram has a bit more variation in this histogram than compared to the median-median line histograms, with the values spread through 3.5 and 4.5, with the most values centered between 4.0 and 4.25, meaning these slopes have more variation than our median-median line approach.

The intercepts have a much more obvious bell shaped curve than the slopes, with most of the values between -25 and -15, which suggests a normal distribution of values, but there is still greater variation in the intercepts than in the slopes, which means there could be a large variation in where out regression line will start on the plot.

Question 7d

```
plot(cars$speed, cars$dist, xlab="Speed", ylab="Stopping Distance")

for(i in 1:10){
  abline(a=intercept_grp_lm[i], b= slope_grp_lm[i], col="red3", lwd=1)
}
```



The 10 fitted regression lines are not parallel but converge and get closer to each other in the middle of the plot, around the average speed, which makes sense given the variability of our slopes found in the histograms. The line placement on the other hand has a large variability in the range of x, with a decent amount of values consistenly starting between 0 and 5 on the x-axis, but there are also a decent amount of outliers that start at an x-value less than zero or greater than 5.