STAT 308 – In Class Exercise

For the problems in which calculations are needed, please include your `R` code with your answers, otherwise you would not be given full credit on the exam.

The `cars` dataset is automatically available in base `R`. The dataset provides information on 50 randomly selected cars and information how long it takes the cars to come to a complete stop when travelling at a particular speed. The variables in the dataset are `speed`, the speed (in mph) at which the car was travelling when it began to stop and `dist`, the distance (in feet) it took the car to come to a complete stop.

1. Suppose we think that speed has nothing to do with the distance the car takes to stop.

a. Report the mean and standard deviation of the stopping distance

```
est <- mean(cars$dist)
s <- sd(cars$dist)
est
```

```
## [1] 42.98
```

```
s
```

```
## [1] 25.76938
```

Sample mean stopping distance is 42.98 feet. Sample standard devitation for stopping distance is 25.77 feet.

b. Calculate a 95% confidence interval for the mean stopping distance. Interpret this interval in the context of the problem.

```
est <- mean(cars$dist)
n <- nrow(cars)
se <- s/sqrt(n)
alpha <- 0.05
crit <- qt(1-alpha/2,df=n-1)
est + c(-1,1)*crit*se
```

```
## [1] 35.65642 50.30358
```

(35.66,50.30). We are 95% confident that the true mean stopping distance for all cars is between 35.66 and 50.30 feet.

c. Suppose it is known that if a car can stop in 35 feet or less, the car is deemed to be safe. Perform a hypothesis test where we wish to test our null hypothesis that, on average, the cars are safe. Be sure to properly specify your null and alternative hypotheses, test statistic, p-value, and decision and conclusion in the context of the problem.

$H_0 : \mu \leq 35$, $H_a : \mu > 35$.

```
t.test(cars$dist,mu=35,alternative="greater")
```

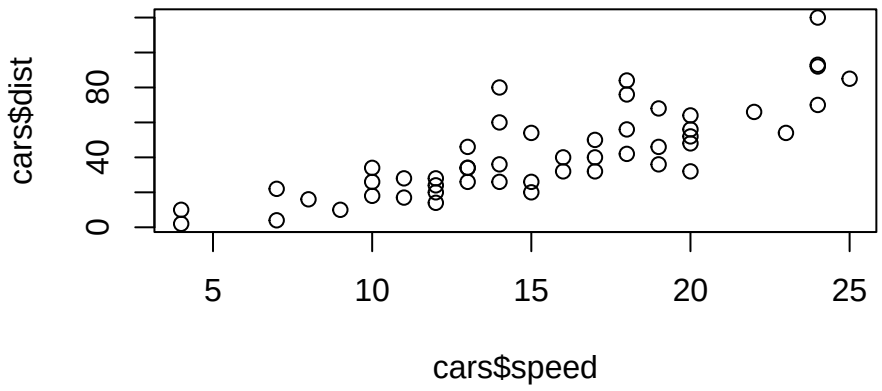
```
##
##  One Sample t-test
##
## data:  cars$dist
## t = 2.1897, df = 49, p-value = 0.01667
## alternative hypothesis: true mean is greater than 35
## 95 percent confidence interval:
##  36.87008      Inf
## sample estimates:
## mean of x
##     42.98
```

$t = 2.19$, p-value $= 0.01667$, p-value $< \alpha$, so we reject H_0. We have statistically significant evidence that the true mean stopping distance of all cars is greater than 35 feet.

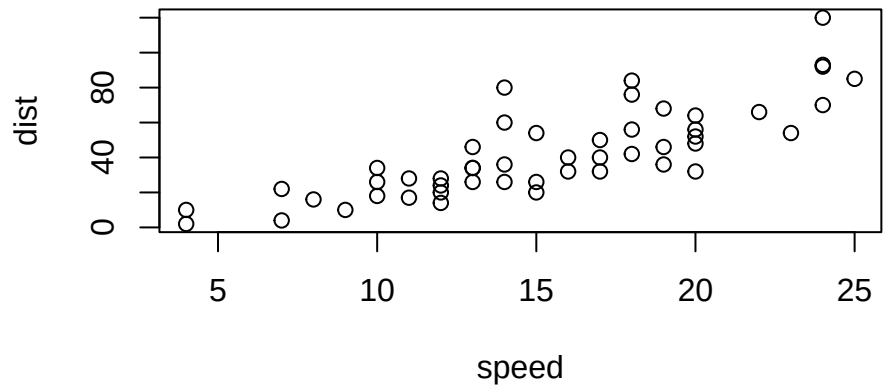
2. Now, suppose someone comes to you and says if we know how fast the car was travelling when it started to stop, we can make better predictions about the stopping distance of the car. We want to start by creating a linear model for stopping distance based on the cars' speed.

a. Create a scatterplot of speed vs. distance. Determine if a linear model is valid for the given data.

```
plot(cars$speed,cars$dist)
```




```
plot(dist ~ speed,data=cars)
```



Stopping distance appears to be increasing linearly with the car's speed, so I think a linear model is appropriate.

b. State the least squares regression line for speed vs. distance.

```
mod <- lm(dist ~ speed,cars)
summary(mod)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

$\hat{Y}_x = -17.58 + 3.93 * x$

$\hat{Dist} = -17.58 + 3.93 * Speed$

c. Interpret the intercept and the slope of the regression lines. Comment on the validity of these interpretations.

When a car is travelling at 0 mph, the expected stopping distance is -17.58 feet. This does not make sense because a car at a complete stop cannot travel a negative distance.

When a car's speed increases by 1 mph, the expected stopping distance increases by 3.93 feet.

d. What is the predicted stopping distance for a car that is travelling at 17 mph? Calculate and interpret 90% confidence and prediction intervals for this prediction.

```
newdata <- data.frame(speed=17)
predict(mod,newdata)
```

```
##        1
## 49.27185
```

```
predict(mod,newdata,interval="confidence",level=0.9)
```

```
##        fit      lwr      upr
## 1 49.27185 45.45728 53.08643
```

```
predict(mod,newdata,interval="prediction",level=0.9)
```

```
##        fit      lwr      upr
## 1 49.27185 23.19631 75.34739
```

We predict that a car travelling 17 mph will take 49.27 feet to stop.

90% CI is (45.46,53.09). We are 90% confident that the average stopping distance for all cars travelling at 17 mph is between 45.46 and 53.09 feet.

90% PI is (23.20,75.35). We are 90% confidenct that a randomly selected car travelling at 17 mph has a stopping distance between 23.2 and 75.35 feet.

e. Suppose now I wish to test whether or not the average stopping distance increases linearly with the car's speed? Perform this hypothesis test, stating the correct null and alternative hypotheses, the test statistic, p-value, and decision and conclusion in the context of the problem.

$H_0 : \beta_1 = 0$, $H_a : \beta_1 > 0$

```
summary(mod)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
```

```
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

$t = 9.464$ p-value = 1.49e-12/2 = 7.499e-13. p-value $< \alpha = 0.05$, so we reject H_0. We have statistically significant evidence of a positive linear relationship between the car's speed and its stopping distance.

f. Create an ANOVA table from the least squares regression. Use the table to answer the following questions.

```
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: dist
##           Df Sum Sq Mean Sq F value   Pr(>F)
## speed      1  21186 21185.5  89.567 1.49e-12 ***
## Residuals 48  11354   236.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(i). What are the sums of squares for the model?

```
SSM <- anova(mod)[1,2] # Extracts value from 1st row and
# 2nd column of ANOVA table
SSM
```

```
## [1] 21185.46
```

(ii). What are the total sums of squares?

```
SSE <- anova(mod)[2,2] # Extracts value from 2nd row and
# 2nd column of ANOVA table
SST <- SSM + SSE
SST
```

```
## [1] 32538.98
```

(iii). What is the r^2 for the model? Interpret this value in the context of the given problem

```
r2 <- SSM/SST
r2
```

```
## [1] 0.6510794
```

$r^2 = 0.6511$. 65.11% of the variation in the car's stopping distance can be explained by its linear relationship with the car's speed.

(iv). Using r^2, calculate the estimate of the correlation coefficient, r.

```
r <- sqrt(r2)
r
```

```
## [1] 0.8068949
```

$r = 0.8069$

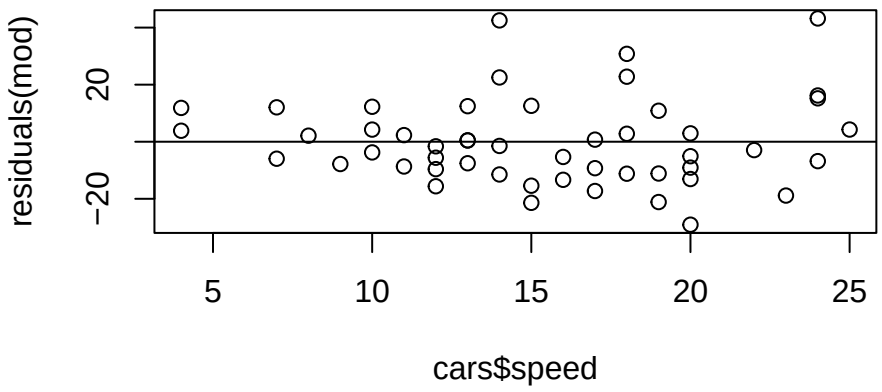
(v). What is the estimate of the regression variance? (Mean Squared Error)

```
s2 <- anova(mod)[2,3] # Extracting the value from the 2nd row
# of the 3rd column
s2
```

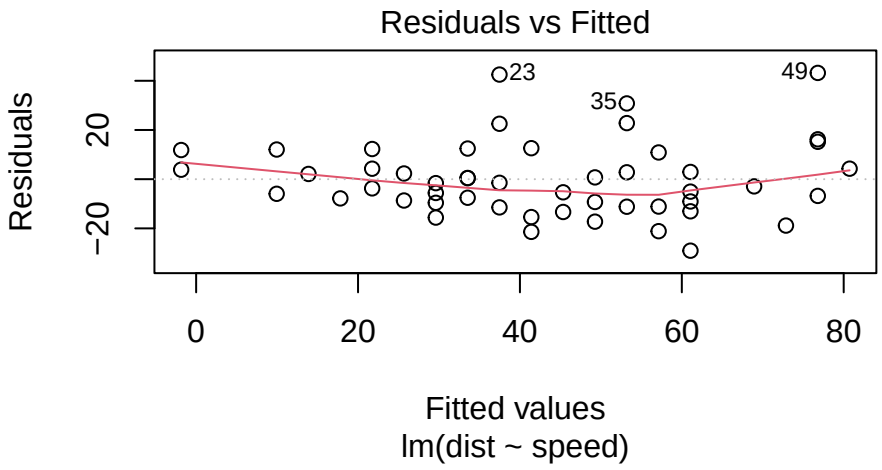
```
## [1] 236.5317
```

g. Determine if the assumption of homoscedasticity is violated.

```
plot(cars$speed,residuals(mod))
abline(h=0)
```




```
plot(mod,1) # Another way to produce residual plot
```

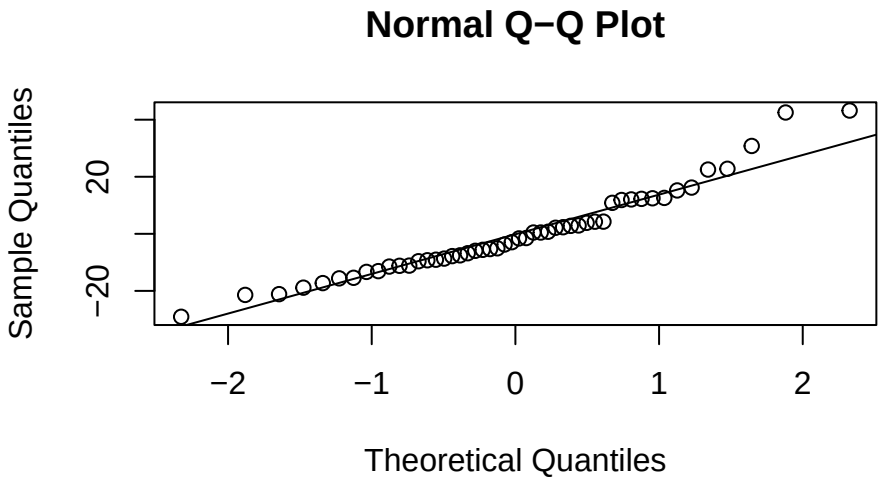



```
# This plots the fitted hat{y} values on the x-axis instead
# of X
```

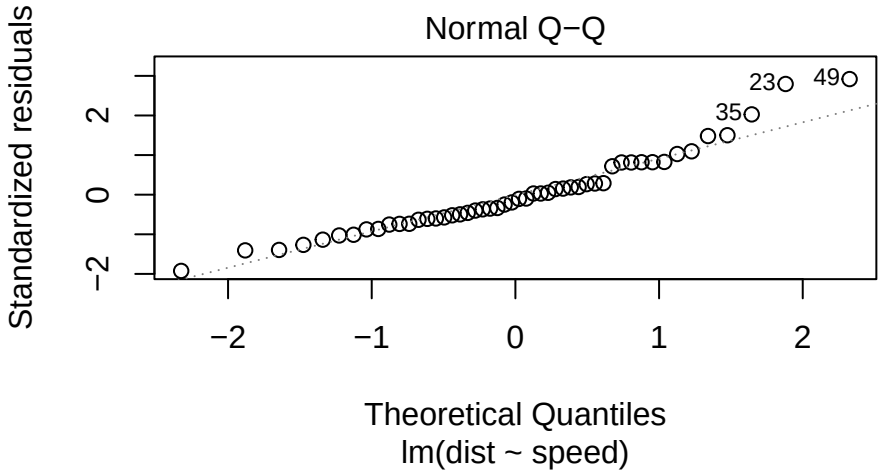
There appears to be some kind of megaphone shape as the fitted values gets larger. Thus, we believe the assumption of homscedasticity/ common variance is violated.

h. Determine if the assumption of normally distributed residuals is violated.

```
qqnorm(residuals(mod))
qqline(residuals(mod))
```




```
plot(mod,2) # This is another method of producing the qq-plot
```



The majority of points fall on the 45-degree line, and therefore, the assumption of normally distributed residuals is not violated.