# Homework 3

Alexander Hernandez

11/03/2022

```
library(pwr)
library(MASS)
```

## 1) "HairEyeColor" of 592 students

### a) Is hair color independent of eye color for men?

```
# H0: Hair and eye color are independent
# Ha: Hair and eye color are   dependent
chisq.test(HairEyeColor[, , "Male"])
```

```
## Warning in chisq.test(HairEyeColor[, , "Male"]): Chi-squared approximation may
## be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  HairEyeColor[, , "Male"]
## X-squared = 41.28, df = 9, p-value = 4.447e-06
```

```
# With a p-value of 4.447e-06, there is enough evidence to reject the null,
# and it can be claimed that hair color is dependent on eye color for males.
```

### b) Is hair color independent of eye color for women?

```
# H0: Hair and eye color are independent
# Ha: Hair and eye color are   dependent
chisq.test(HairEyeColor[, , "Female"])
```

```
## Warning in chisq.test(HairEyeColor[, , "Female"]): Chi-squared approximation may
## be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  HairEyeColor[, , "Female"]
## X-squared = 106.66, df = 9, p-value < 2.2e-16
```

```
# With a p-value of 2.2e-16, there is enough evidence to reject the null,
# and it can be claimed that hair color is dependent on eye color for females.
```

## 2) Diets A and B. How many subjects are needed in each group, assuming equal sized groups (a=0.05, Power=0.8)?

```
pwr.t.test(d=(0-10)/16.03, power=.8, sig.level=0.05, type="two.sample", alt="two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 41.31968
##               d = 0.6238303
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
# Subjects required for each group is at least 21.
```

## 3) Fire Damage versus distance of fire from Fire Station

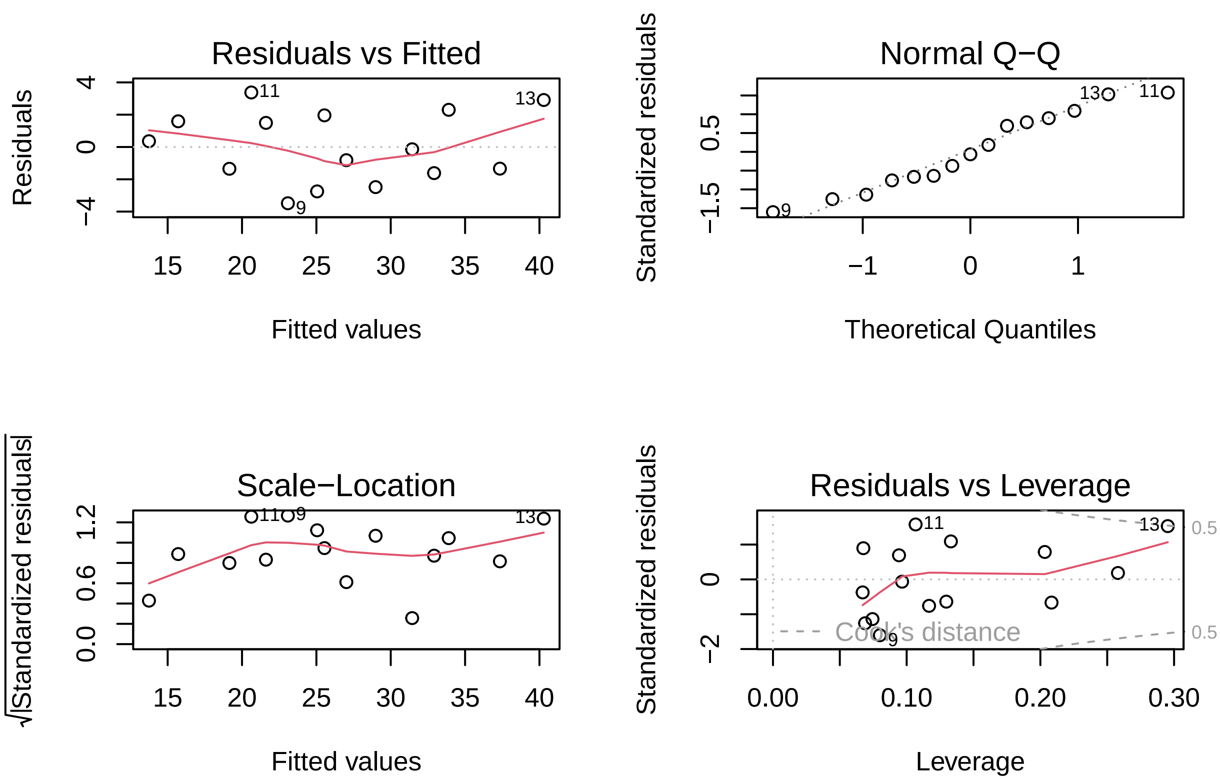
```r
fires = read.table("C:\\repos\\STAT 50001\\Homework 3\\q2.txt",
                   header=TRUE)
attach(fires)
```

### a) Fit a simple linear regression model and analyze the residual plots.

```r
model_3a = lm(Damage ~ Distance)
model_3a
```

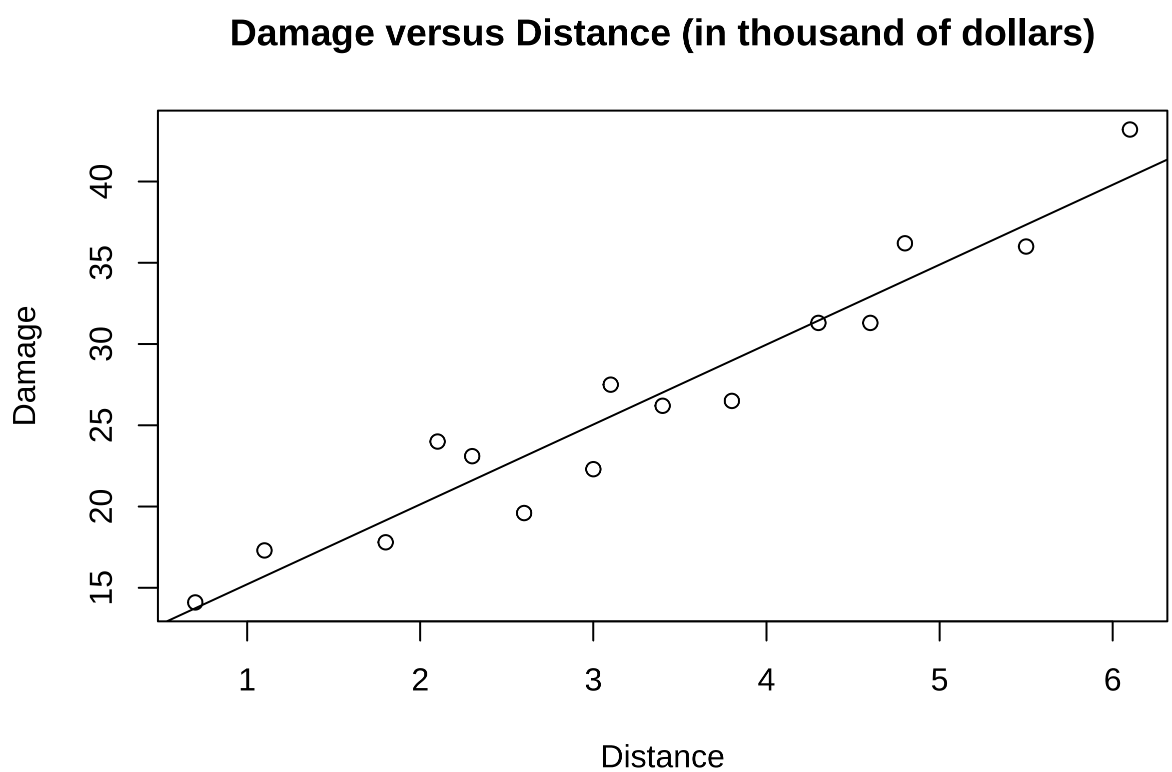
```
##
## Call:
## lm(formula = Damage ~ Distance)
##
## Coefficients:
## (Intercept)     Distance
##      10.300        4.917
```

```r
par(mfrow=c(2,2))
plot(model_3a)
```




```r
# Plot 4 demonstrates that the residuals may not exhibit a linear pattern.
```

```r
par(mfrow=c(1,1))
plot(fires,
     main="Damage versus Distance (in thousand of dollars)")
abline(model_3a)
```




**b) What is the expected Damage if the fire station is 4 miles away?**

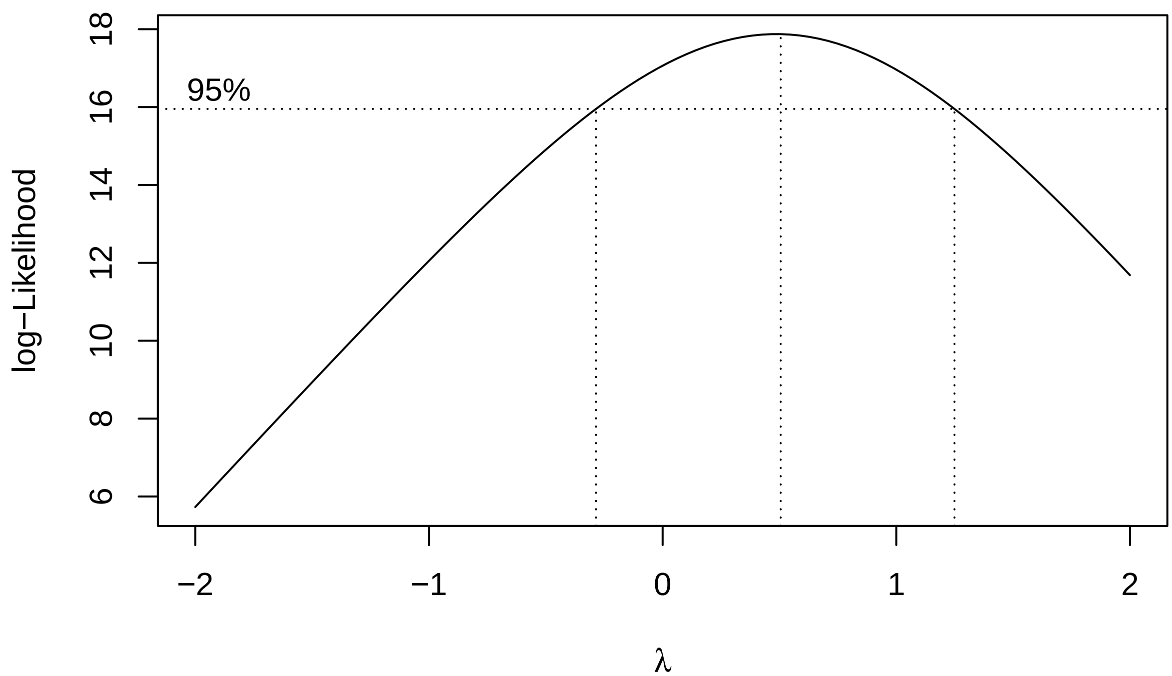
```r
predict(model_3a, data.frame(Distance = 4), interval="conf", level=0.95)
```

```
##       fit      lwr      upr
## 1 29.9666 28.56955 31.36365
```

```r
# Expected damage interval with 95% confidence: (28.56955, 31.36365)
```

**c) Use the Box-Cox transformation to choose an appropriate value of to improve the model.**

```
boxcox(model_3a)
```




```
# The graph shows a lambda value of 0.5 should be chosen.
```

**d) Fit a simple linear regression model after transformation.**

```
model_3b = lm(Damage**0.5 ~ Distance)
model_3b
```

```
##
## Call:
## lm(formula = Damage^0.5 ~ Distance)
##
## Coefficients:
## (Intercept)     Distance
##      3.5183       0.4777
```

### e) Compare and contrast models in (a) and (d).

```
summary(model_3a)$r.squared
```

```
## [1] 0.9265571
```

```
summary(model_3b)$r.squared
```

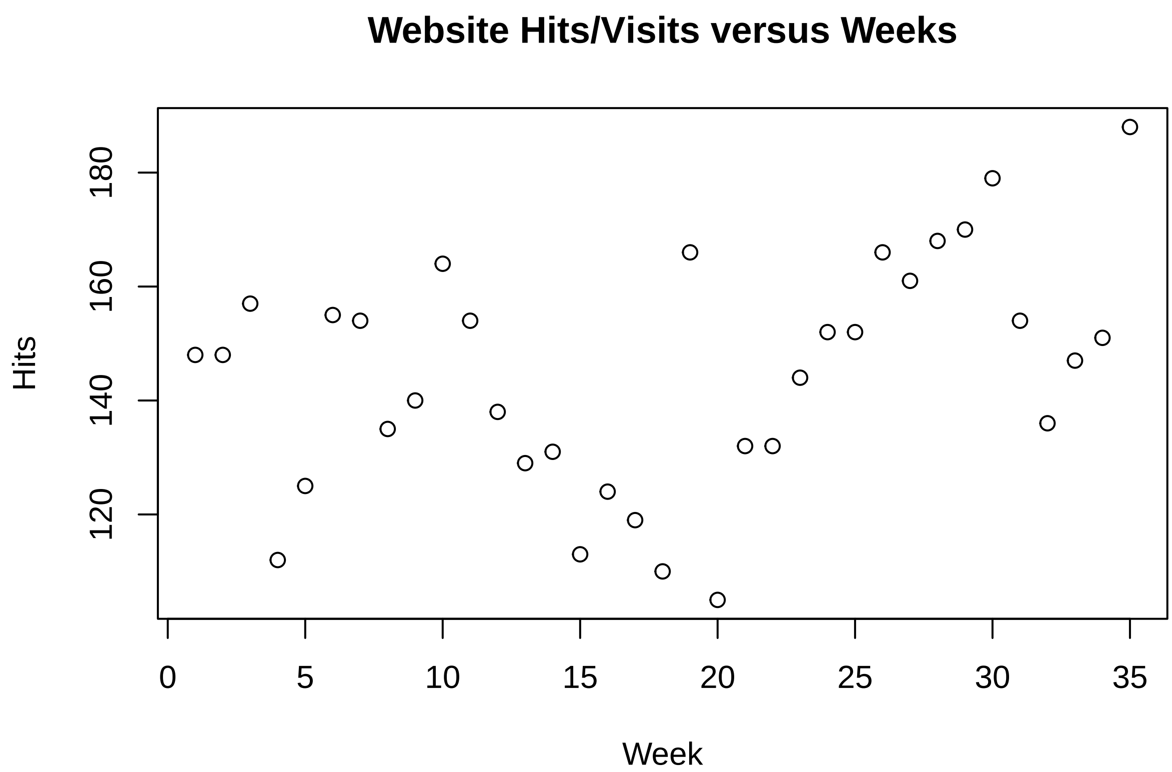
```
## [1] 0.9315393
```

```
# The new model in part d has a lambda=0.5 with a higher R**2 of 0.9315393
# compared to the part a model which has a R**2 of 0.9265571.
# This shows that the transformed model may be a better fit.
```

## 4) Website weeks versus hits/visits

### a) Display the data using a scatterplot.

```
website = read.table("C:\\repos\\STAT 50001\\Homework 3\\q4.txt",
                     header=TRUE)
plot(website,
     main= "Website Hits/Visits versus Weeks")
```




### b) Calculate the correlation coefficient to measure the association between the week and the number of hits on the website. Check whether rank correlation is more appropriate than Pearson correlation

```
# Pearson Test
cor.test(website$Week, website$Hits)
```

```
##
##  Pearson's product-moment correlation
##
## data:  website$Week and website$Hits
## t = 2.1952, df = 33, p-value = 0.03529
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.02691344 0.61682992
## sample estimates:
##       cor
```

```
## 0.3569585
```

```r
# Spearman Rank Test
cor.test(website$Week, website$Hits,
         method="spearman", exact=FALSE)
```

```
##
##  Spearman's rank correlation rho
##
## data:  website$Week and website$Hits
## S = 4842.7, p-value = 0.05945
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.3217489
```

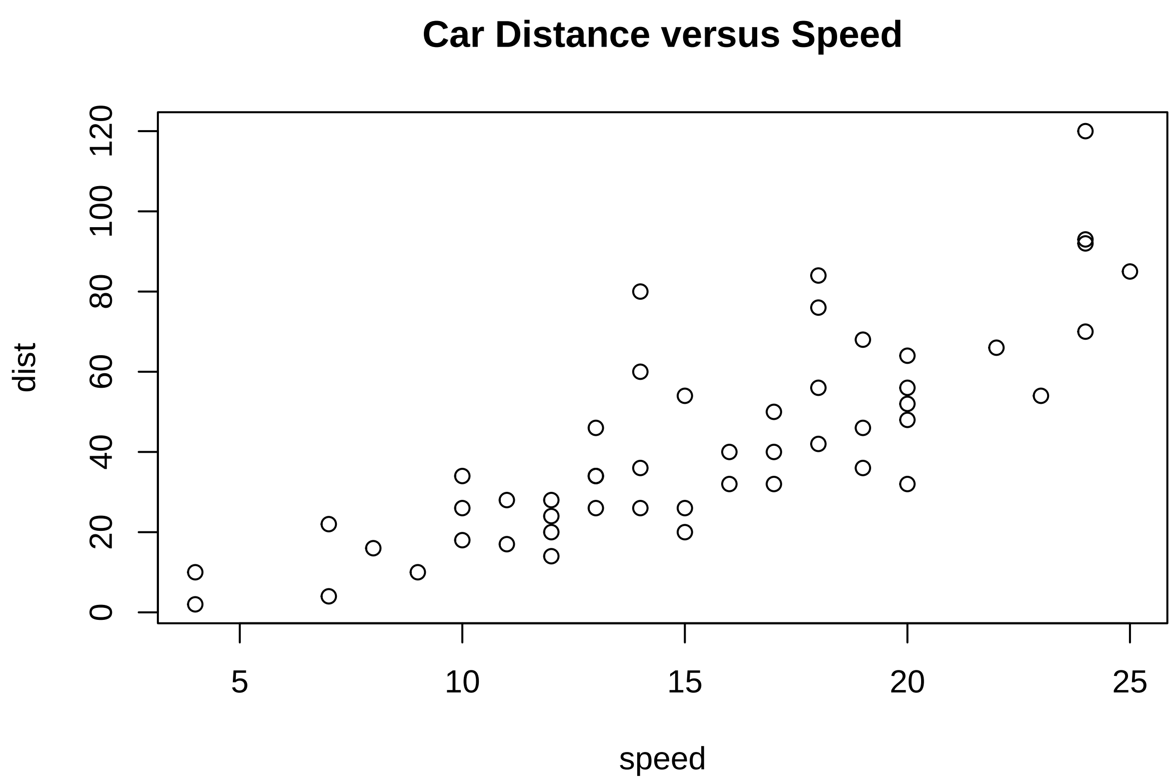
### c) Test for the significance of the correlation at 0.05 level.

```r
# Ho: Correlation coefficient is not significantly different from 0 (p = 0)
# Ha: Correlation coefficient is significantly different from 0     (p != 0)
# Given the values are not ranked/ordinal, the Pearson Correlation Test
# is used, which results in using the p-value, 0.357. This means
# the null hypothesis is rejected, claiming there is a significant
# linear relationship between website weeks and website hits.
```

## 5) Cars: speed versus distance

### a) Display the data using scatter plot.

```
plot(cars, main="Car Distance versus Speed")
```




### b) Fit a simple regression model using speed as a predictor variable.

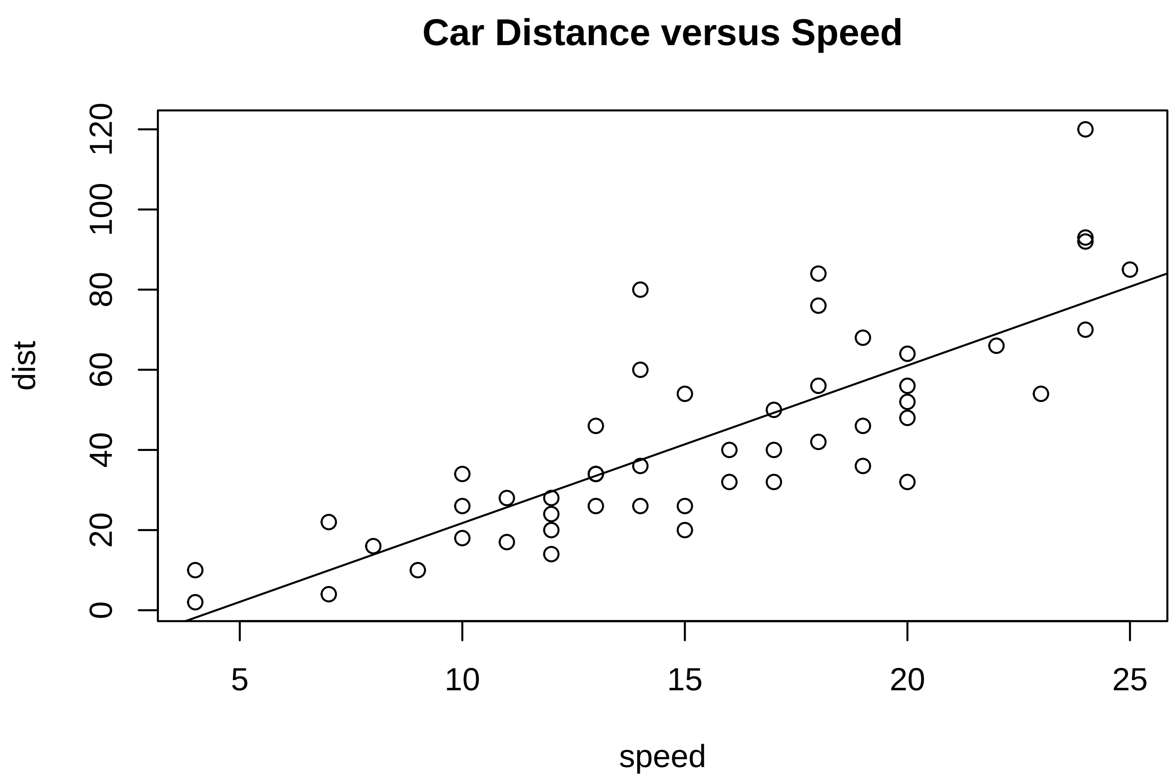
```
attach(cars)
model_5 = lm(dist ~ speed)
model_5
```

```
##
## Call:
## lm(formula = dist ~ speed)
##
## Coefficients:
## (Intercept)        speed
##     -17.579        3.932
```

```
# dist = -17.579 + 3.932(speed)
```

**c) Add the fitted line to the scatter plot.**

```
plot(cars, main="Car Distance versus Speed")
abline(model_5)
```




**d) Calculate the residuals and fitted values and print only first five observations of the residuals and fitted values.**

```
head(resid(model_5),  5)
```

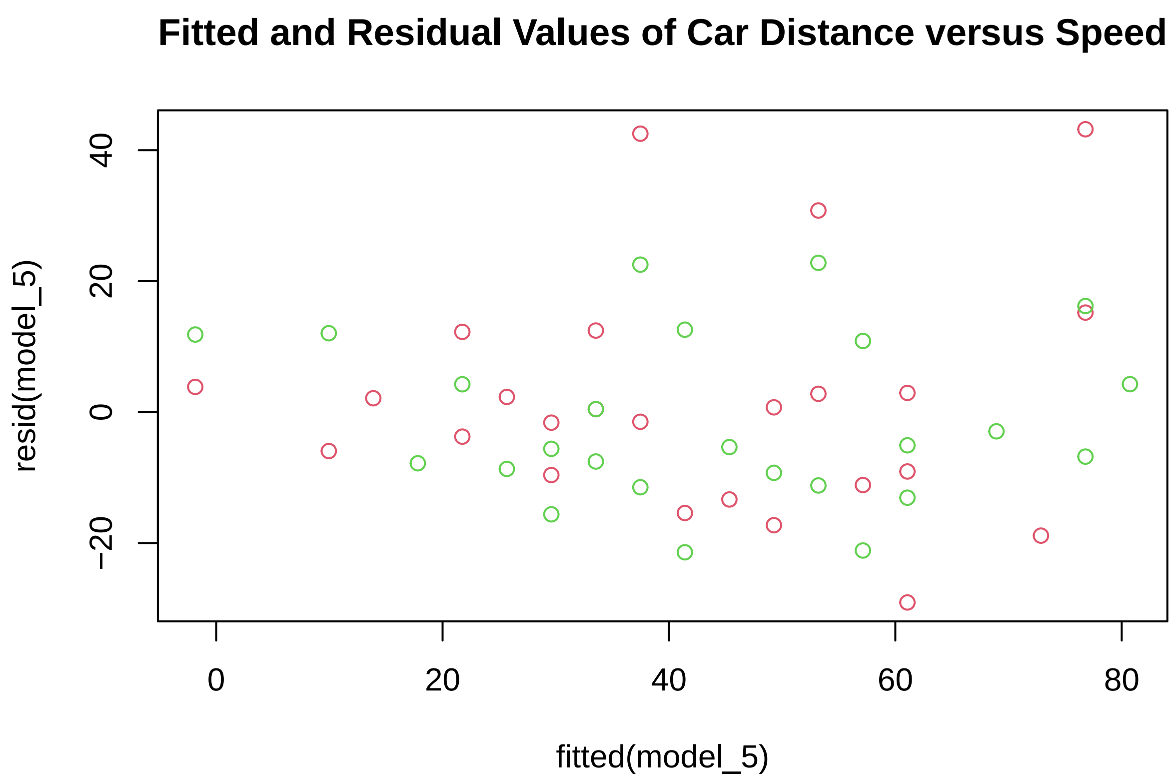
```
##         1         2         3         4         5
##  3.849460 11.849460 -5.947766 12.052234  2.119825
```

```
head(fitted(model_5), 5)
```

```
##         1         2         3         4         5
## -1.849460 -1.849460  9.947766  9.947766 13.880175
```

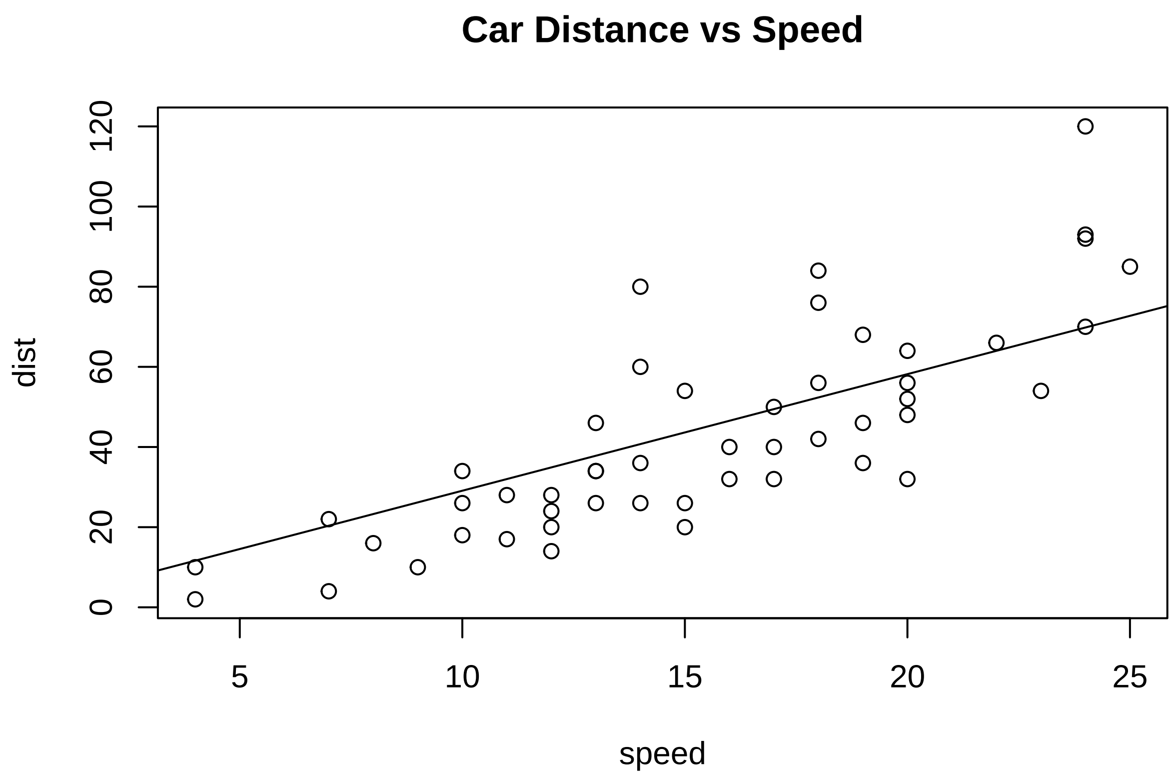
**e) Create a scatter plot of the residuals and fitted values.**

```
plot(fitted(model_5), resid(model_5),
     col  = c(2,3),
     main = "Fitted and Residual Values of Car Distance versus Speed")
```



**f) Assuming that no intercept model is appropriate fit a simple linear regression model.**

```r
model_5 = lm(dist ~ -1 + speed)
plot(cars, main="Car Distance vs Speed")
abline(model_5)
```




**g) Calculate and compare the coefficient of determination for both with intercept and no-intercept models.**

```r
summary(lm(dist ~ speed))$r.squared
```

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

```r
summary(lm(dist ~ -1 + speed))$r.squared
```

```
## [1] 0.8962893
```

```r
# Compared to the intercept R**2 of 0.65, the no-intercept
# R**2 is significantly stronger at 0.90 (rounded),
# which indicates a valid strength of data.
```

### h) Using your fitted model predict the stopping distance for a car with an speed of 21 mph.

```
predict(lm(dist ~ speed), data.frame(speed=21), interval="conf", level=0.9)
```

```
##        fit      lwr      upr
## 1 65.00149 59.65934 70.34364
```

```
# Intercept model with a 90% confidence interval of (59.65934, 70.34364),
# predicts a fit value of 65.00149.
predict(lm(dist ~ -1 + speed), data.frame(speed=21), interval="conf", level=0.9)
```

```
##        fit      lwr      upr
## 1 61.09178 56.11453 66.06902
```

```
# No-intercept model with a 90% confidence interval of (56.11453, 66.06902),
# predicts a fit value of 61.09178.
```