STA 5207: Homework 9

Due: Friday, April 5th by 11:59 PM

Include your R code in an R chunks as part of your answer. In addition, your written answer to each exercise should be self-contained so that the grader can determine your solution without reading your code or deciphering its output.

```
##
## Attaching package: 'olsrr'

## The following object is masked from 'package:MASS':
##
##     cement

## The following object is masked from 'package:datasets':
##
##     rivers

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## Attaching package: 'faraway'

## The following object is masked from 'package:olsrr':
##
##     hsb
```

Exercise 1 (Brains) [40 points]

For this exercise, we will use the `mammals` data set in the `MASS` package. You can also find the data in `mammals.csv` on Canvas. The data set contains the average brain and body weights of 62 species of land mammals. There are 62 observations and two variables:

- `body`: Average body weight in kilograms (kg).
- `brain`: Average brain weight in grams (g).

In the following exercise, we will use `brain` as the response and `body` as the predictor.

1. (5 points) Perform OLS regression with `brain` as the response and `body` as the predictor. Check the normality and constant variance assumptions using a hypothesis test at the $\alpha = 0.05$ level. Do you feel that they have been violated? Justify your answer.

```
model = lm(brain ~ body, data = mammals)

shapiro.test(resid(model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.41112, p-value = 2.316e-14
```

```
bptest(model)
```

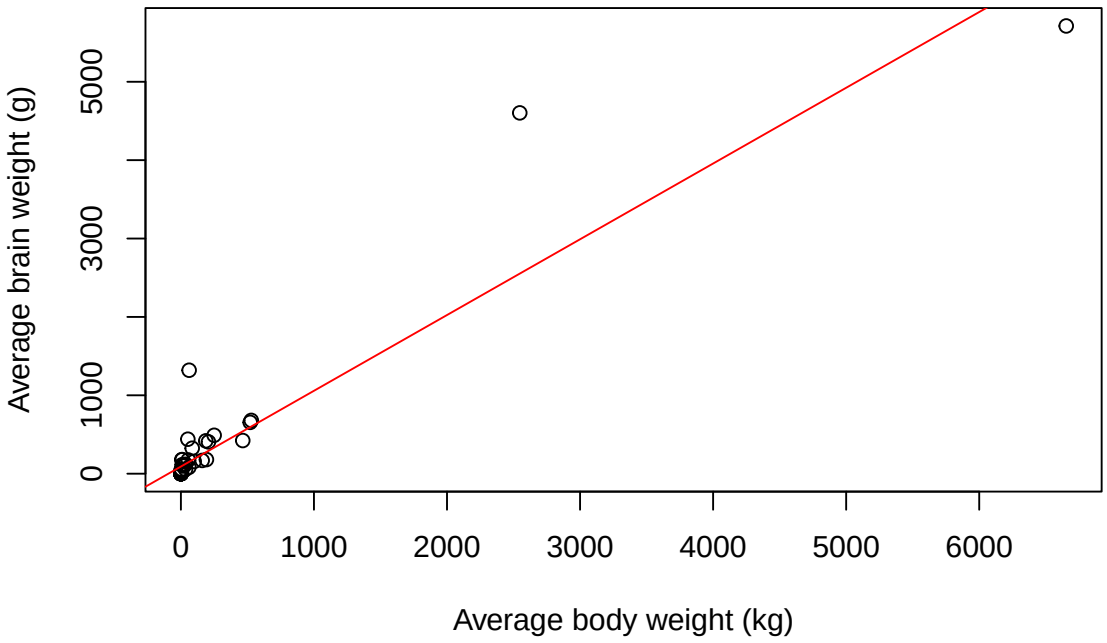
```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 12.537, df = 1, p-value = 0.0003989
```

With a test statistic of .411 and p-value of $2.316e^{-14}$, we reject the null hypothesis and conclude the errors have a non-normal distribution based on the Shapiro-Wilk test.

With a test statistic of 12.537 and a p-value of .0003989, we reject the null hypothesis, and conclude the constant variance assumption is violated based on the Breusch-Pagan test.

2. (3 points) Create a scatter plot of the data and add the fitted regression line. Based on this plot, does their appear to be any outliers, high-leverage points, or high influential data points? Include the plot in your response.

```
plot(brain ~ body, data = mammals, xlab = "Average body weight (kg)", ylab = "Average brain weight

abline(model, col = "red")
```



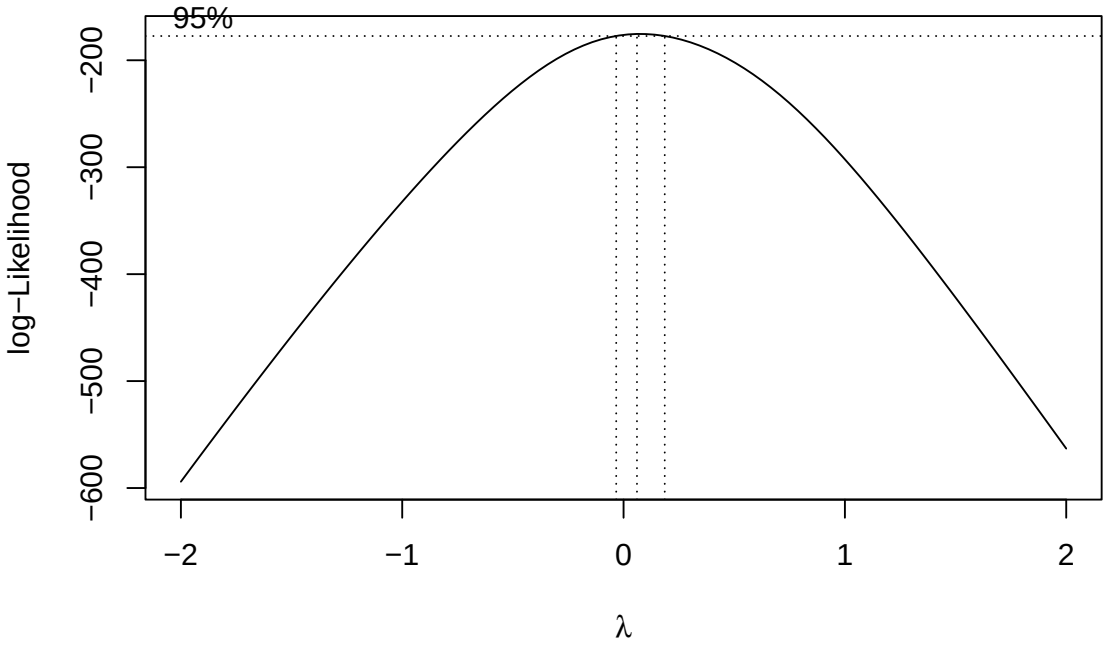
There appear to be two outliers around 2500 kg and over 6000 kg.

3. (6 points) Since the body weights range over more than one order of magnitude and are strictly positive, we will use log(`body`) as our *predictor*, with no further justification (Recall *the log rule*: if the values of a variable range over more than one order of magnitude and the variable is strictly positive, then replacing the variable by its logarithm may be helpful). Use the Box-Cox method to verify that log(`brain`) is then a "recommended" transformation of the *response* variable. That is, verify that the log transformation is amoung the "recommended" values of λ when considering,

$$g_\lambda(\texttt{brain}) = \beta_0 + \beta_1 \log(\texttt{body}) + \varepsilon_i.$$

Report the relevant plot returned by the `boxcox` function and use the appropriate zoom onto the relevant values. Indicating the property of the plot that justifies the log transformation.

```
bc = boxcox(model, plotit = TRUE)
```



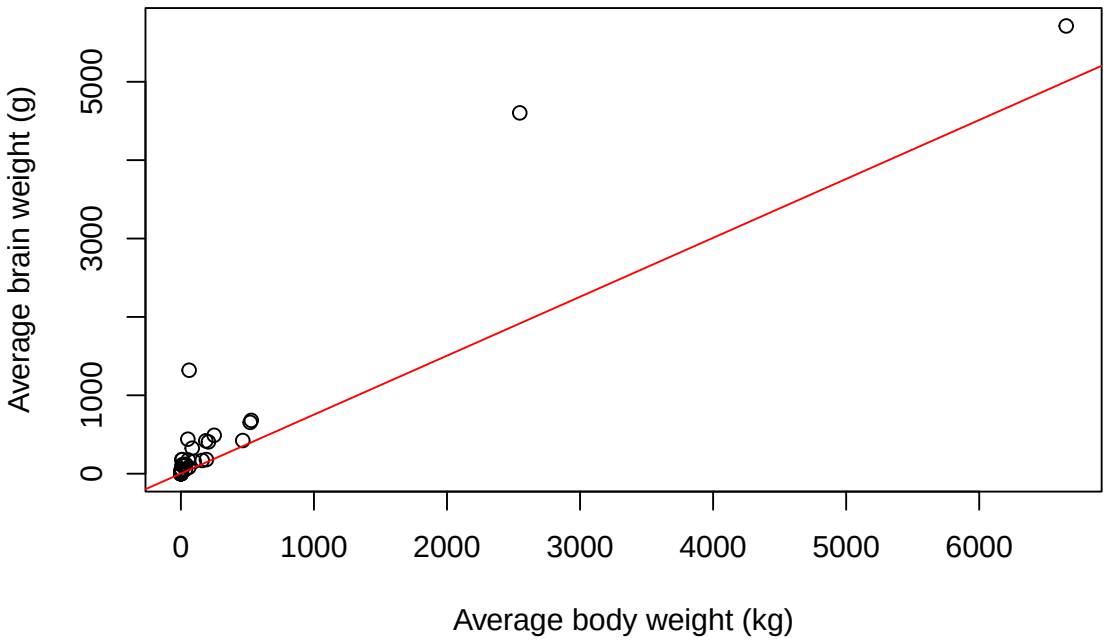

```
bc$x[which.max(bc$y)]
```

```
## [1] 0.06060606
```

$\hat{\lambda} = .0606$, which suggests a logarithmic transformation would be appropriate.

4. (5 points) Fit the model justified in Question 3. That is, fit a model with log(`brain`) as the response and log(`body`) as the predictor. Create a scatter plot of the data and add the fitted regression line for this model. Does a linear relationship seem to be appropriate here?

```
model_log = lm(log(brain) ~ log(body), data = mammals)
plot(brain ~ body, data = mammals, xlab = "Average body weight (kg)", ylab = "Average brain weight

abline(model_log, col = "red")
```



A linear relationship does appear to be appropriate here.

5. (3 points) Based on the model from Question 4, check the normality and constant variance assumptions using a hypothesis test at the $\alpha = 0.05$ level. Do you feel that they have been violated? Justify your answer.

```
shapiro.test(resid(model_log))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model_log)
## W = 0.98268, p-value = 0.5293
```

```
bptest(model_log)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_log
## BP = 0.27849, df = 1, p-value = 0.5977
```

With a test statistic of .982 and p-value of 0.5293, we do reject the null hypothesis and conclude the errors have a normal distribution based on the Shapiro-Wilk test.

With a test statistic of .278 and a p-value of .5977, we do not reject the null hypothesis, and conclude the constant variance assumption is not violated based on the Breusch-Pagan test.

6. (6 points) Using the model from Question 4, check for any high influential observations. Report any observations you determine to be highly influential.

```
high_inf_ids = which(cooks.distance(model_log) > 4 /length(resid(model_log)))
high_inf_ids
```

```
## Human
##    32
```

Human is highly influential.

7. (6 points) Use the model in Question 4 to predict the brain weight of a male Snorlax, which has a body weight of 1014.1 *pounds*. (A Snorlax would be a mammal, right?) Construct a 90% prediction interval.

```
new_data <- data.frame(body = 1014.1)
predict(model_log, newdata = new_data, interval = "prediction", level = 0.90)
```

```
##        fit      lwr     upr
## 1 7.337776 6.138762 8.53679
```

We are 90% confident that the brain weight of a Snorlax would be between 6.139 and 8.536 grams.

8. (6 points) A common measure of model performance is the root mean squared error (RMSE), which is defined as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

We prefer models with a lower RMSE. Report the RMSE values for the model in Question 1 and Question 4. Based on this criteria, which model do you prefer?

```
sqrt(mean((mammals$brain - predict(model))^2))
```

```
## [1] 329.2768
```

```
sqrt(mean((mammals$brain - exp(predict(model_log)))^2))
```

```
## [1] 272.6149
```

I prefer the log transformed model as it has a lower RMSE.

Exercise 2 (TV and Health) [40 points]

For this exercise, we will use the `tvdoctor` data set in the `faraway` package. You can also find the data in `tvdoctor.csv` on Canvas. The data set contains information on life expectancy, doctors, and televisions collected in 38 countries in 1993. There are 38 observations on three variables:

- `life`: Life expectancy in years.
- `tv`: Number of people per television set.
- `doctor`: Number of people per doctor.

In the following exercise, we will use `life` as the response and `tv` as the predictor.

1. (6 points) Use forward selection based on t-tests at the $\alpha = 0.05$ level to select a d-th degree polynomial model with `life` as the response and `tv` as the predictor. Report the estimated regression equation for your chosen model.

```
model_poly1 = lm(life ~ tv, data = tvdoctor)

summary(model_poly1)
```

```
##
## Call:
## lm(formula = life ~ tv, data = tvdoctor)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.8141  -4.6061   0.5876   5.3647   9.4171
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 69.648132   1.101058  63.256  < 2e-16 ***
## tv          -0.036264   0.007937  -4.569 5.56e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.293 on 36 degrees of freedom
## Multiple R-squared:  0.3671, Adjusted R-squared:  0.3495
## F-statistic: 20.88 on 1 and 36 DF,  p-value: 5.561e-05
```

```
summary(model_poly1)$coeff
```

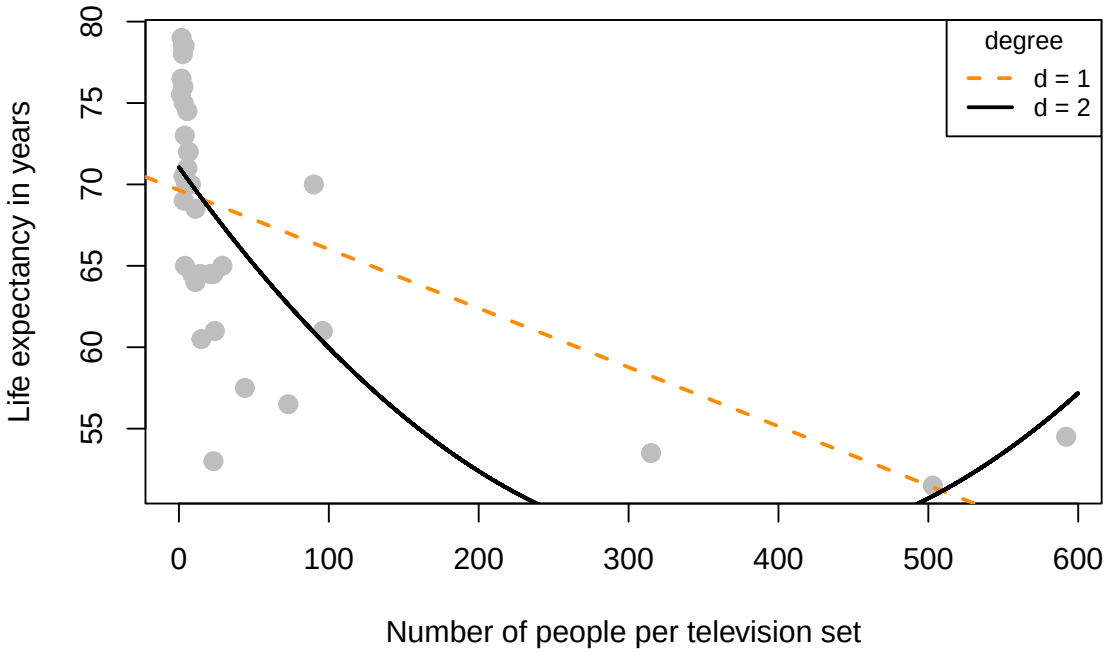
```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 69.64813242 1.101057927 63.255648 1.684483e-38
## tv          -0.03626419 0.007936857 -4.569087 5.560533e-05
```

$$\text{life} = 69.65 - .0000561\text{tv}$$

2. (6 points) Fit polynomial models of degree 1 and 2. Create a scatter plot of the data and add the fitted regression line for each polynomial model. Include the plot in your response.

```
model_poly2 = lm(life ~ poly(tv, 2, raw = TRUE), data = tvdoctor)

plot(life ~ tv, data = tvdoctor, xlab = "Number of people per television set", ylab = "Life expecta

# fitted regression line
abline(model_poly1, col = 'darkorange', lty = 'dashed', lwd = 2)

# fitted regression line for the quadratic model
xplot = seq(0, 600, by = 0.01)
lines(xplot, predict(model_poly2, newdata = data.frame(tv = xplot)),
      col = "black", lwd = 2)
```

```r
# add a legend to label the two lines
legend("topright", title = "degree", cex = 0.8,
       legend = c("d = 1", "d = 2"),
       lwd = 2, lty = c(2, 1),
       col = c("darkorange", "black"))
```



3. (3 points) Check for any high *leverage* points using the quadratic model you fit in Question 2. Report any observations you determine to have high leverage.

```r
high_lev_ids = which(hatvalues(model_poly2) > 2 * mean(hatvalues(model_poly2)))
high_lev_ids
```

```
## Bangladesh   Ethiopia    Myanmar
##          2          8         21
```

There are the high leverage points at the countries of Bangladesh, Ethiopia and Myanmar

4. (6 points) Use forward selection based on t-tests at the $\alpha = 0.05$ level to select a d-th degree polynomial model with `life` as the response and `tv` as the predictor with the high *leverage* data points you identified in Question 3 removed. Report the estimated regression equation for your chosen model.

```r
non_inf_ids = which(hatvalues(model_poly2) < 2 * mean(hatvalues(model_poly2)))

model_poly1_fix = lm(life ~ tv, data = tvdoctor, subset = non_inf_ids)

summary(model_poly1_fix)
```

```
##
## Call:
## lm(formula = life ~ tv, data = tvdoctor, subset = non_inf_ids)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9671  -3.7987  -0.3712   4.2035  12.2737
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 71.48255    1.18159  60.497  < 2e-16 ***
## tv          -0.15285    0.04132  -3.699 0.000783 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.769 on 33 degrees of freedom
## Multiple R-squared:  0.2931, Adjusted R-squared:  0.2717
## F-statistic: 13.68 on 1 and 33 DF,  p-value: 0.0007835
```

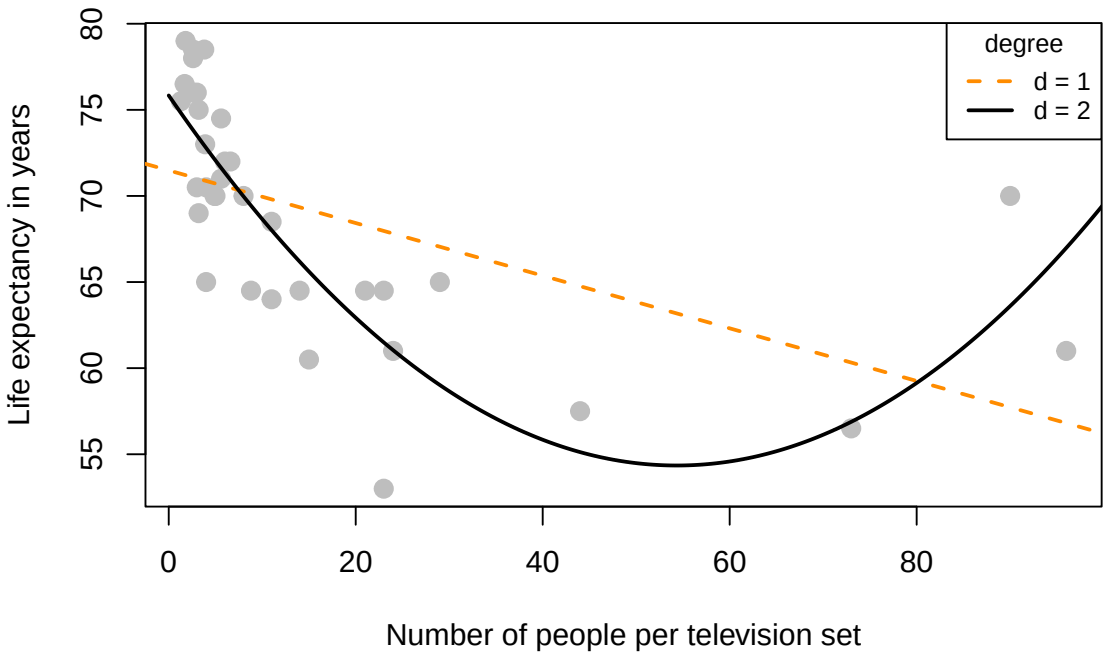
```
summary(model_poly1_fix)$coeff
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 71.4825525 1.18159320 60.496754 2.163861e-35
## tv          -0.1528474 0.04131996 -3.699117 7.834575e-04
```

$$\text{life} = 71.48 - .153\text{tv}$$

5. (6 points) Fit polynomial models of degree 1 and 2 with the high *leverage* data points you identified in Question 3 removed. Create a scatter plot of the data (**Note**: use the `subset` argument to `plot`) and add the fitted regression line for each polynomial model. Include the plot in your response.

```
model_poly1_fix = lm(life ~ tv, data = tvdoctor, subset = non_inf_ids)

model_poly2_fix = lm(life ~ poly(tv, 2, raw = TRUE), data = tvdoctor, subset = non_inf_ids)

plot(life ~ tv, data = tvdoctor, subset = non_inf_ids, xlab = "Number of people per television set"

# fitted regression line
abline(model_poly1_fix, col = 'darkorange', lty = 'dashed', lwd = 2)

# fitted regression line for the quadratic model
xplot = seq(0, 600, by = 0.01)
lines(xplot, predict(model_poly2_fix, newdata = data.frame(tv = xplot)),
      col = "black", lwd = 2)

# add a legend to label the two lines
legend("topright", title = "degree", cex = 0.8,
       legend = c("d = 1", "d = 2"),
       lwd = 2, lty = c(2, 1),
       col = c("darkorange", "black"))
```



6. (5 points) Since the number of people per television set (`tv`) ranges over more than one order of magnitude and are strictly positive, we might use log(`tv`) as our predictor. Fit an OLS regression model with `life` as the response and log(`tv`) as the predictor. Report the estimated regression equation for this model.

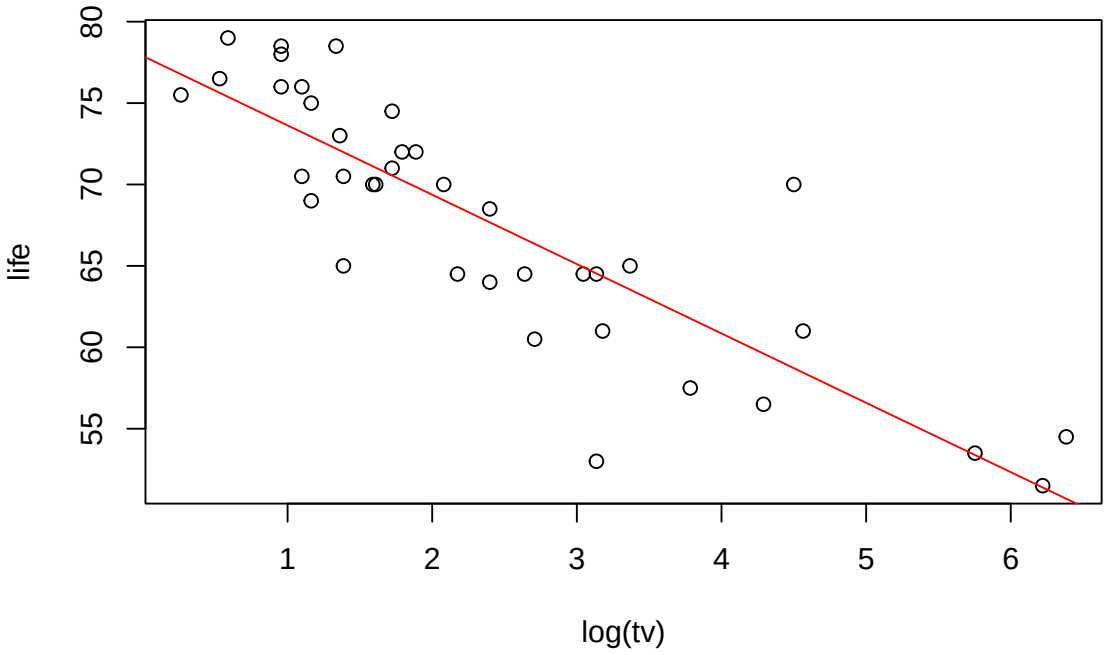
```
model_log = lm(life ~ log(tv), data = tvdoctor)
summary(model_log)$coeff
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 77.887278  1.2202883 63.826946 1.222096e-38
## log(tv)     -4.259678  0.4304296 -9.896341 8.200536e-12
```

$$\text{life} = 77.89 - 4.26\text{tv}$$

7. (3 points) Create a scatter plot of the `life` vs log(`tv`) and add the fitted regression line for model you fit in Question 6. Include the plot in your response.

```
plot(life ~ log(tv), data = tvdoctor)
abline(model_log, col = "red")
```



8. (5 points) Report the adjusted R^2 values for the quadratic model you fit in Question 2 and the model you fit in Question 6. Based on this criteria, which model do you prefer?

```
SST = sum((tvdoctor$life - mean(tvdoctor$life))^2)

1 - sum((tvdoctor$life - predict(model_poly2)^(1/2))^2) / SST
```

```
## [1] -59.74478
```

```
1 - sum((tvdoctor$life - exp(predict(model_log)))^2) / SST
```

```
## [1] -2.505096e+63
```

I prefer the log transformed model as it has a much higher R^2 value.

Excercise 3 (The `cars` Data Set) [20 points]

For this exercise, we will use the built-in `cars` data set. You can also find the data in `cars.csv` on Canvas. In the following exercise, we will use `dist` as the response and `speed` as the predictor.

1. (5 points) Perform OLS regression with `dist` as the response and `speed` as the predictor. Check the normality and constant variance assumptions using a hypothesis test at the $\alpha = 0.05$ level. Do you feel that they have been violated? Justify your answer.

```r
model = lm(dist ~ speed, data = cars)

shapiro.test(resid(model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.94509, p-value = 0.02152
```

```r
bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 3.2149, df = 1, p-value = 0.07297
```

With a test statistic of .945 and p-value of 0.02152, we do reject the null hypothesis and conclude the errors have a normal distribution based on the Shapiro-Wilk test.

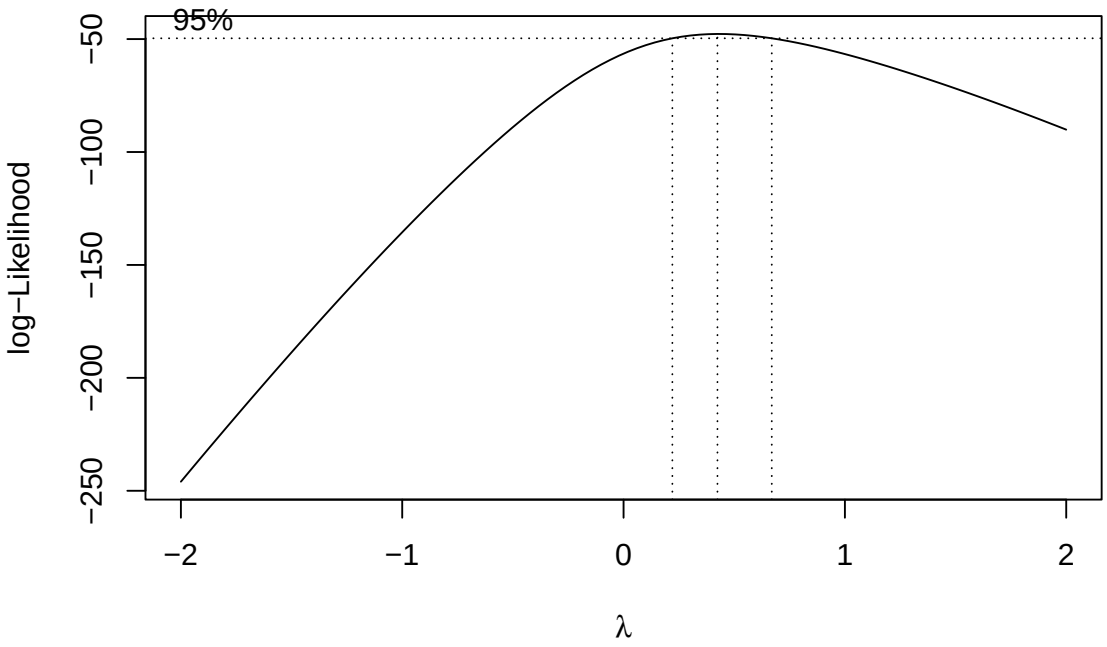
With a test statistic of 3.215 and a p-value of .07297, we do not reject the null hypothesis, and conclude the constant variance assumption is not violated based on the Breusch-Pagan test.

2. (10 points) Use the Box-Cox method to verify that $\sqrt{\texttt{dist}}$ is a "recommended" transformation of the *response* variable. That is, verify that the square-root transformation is among the "recommended" values of λ when considering,

$$g_\lambda(\texttt{dist}) = \beta_0 + \beta_1 \texttt{speed} + \varepsilon_i.$$

Report the relevant plot returned by the `boxcox` function and use the appropriate zoom onto the relevant values. Indicating the property of the plot that justifies the square-root transformation.

```r
bc = boxcox(model, plotit = TRUE)
```




```
bc$x[which.max(bc$y)]
```

```
## [1] 0.4242424
```

$\hat{\lambda} = .4242$, which suggests a square root transformation would be appropriate.

3. (5 points) Fit the model justified in Question 2. That is, fit a model with $\sqrt{\texttt{dist}}$ as the response and `speed` as the predictor. Check the normality and constant variance assumptions using a hypothesis test at the $\alpha = 0.05$ level. Do you feel that they have been violated? Justify your answer.

```
model = lm(sqrt(dist) ~ speed, data = cars)

shapiro.test(resid(model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.97332, p-value = 0.3143
```

```
bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 0.011192, df = 1, p-value = 0.9157
```

With a test statistic of .973 and p-value of 0.3143, we do reject the null hypothesis and conclude the errors have a normal distribution based on the Shapiro-Wilk test.

With a test statistic of .0111 and a p-value of .9157, we do not reject the null hypothesis, and conclude the constant variance assumption is not violated based on the Breusch-Pagan test.