STA 5207: Homework 8

Include your R code as an appendix at the end of your homework. Do not include your code in your answers unless the question explicitly tells you to include your code. Your answers to each exercise should be self-contained without code so that the grader can determine your solution without reading your code or deciphering its output.

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.

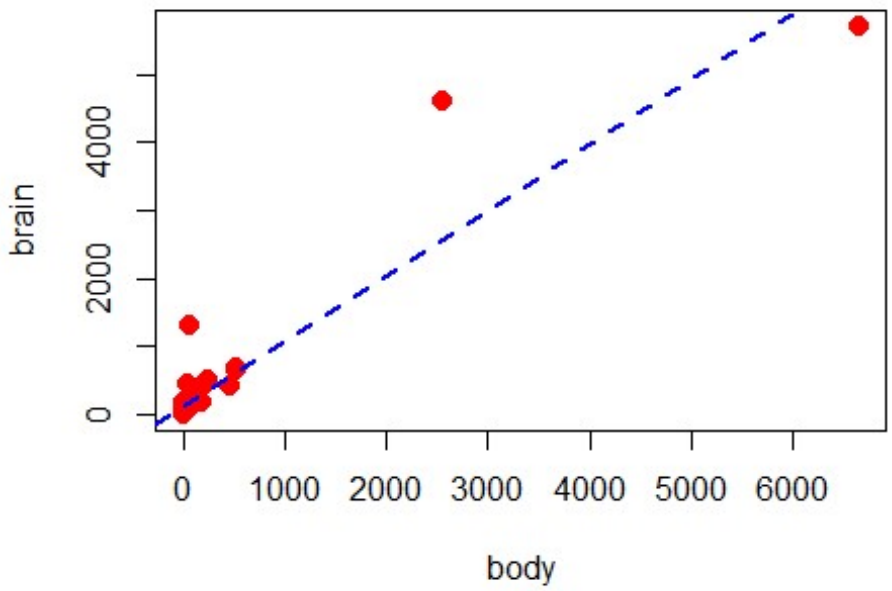
1. (5 points) Perform OLS regression with `brain` as the response and `body` as the predictor. Check the normality and constant variance assumptions using a hypothesis test at the $\alpha = 0.05$ level. Do you feel that they have been violated? Justify your answer.

–Certainly, here's a rewritten version:

–1. Normality Assumptions: - With a p-value of 2.316e-14, significantly below the 0.05 significance level, we reject the null hypothesis. This leads us to the conclusion that the errors exhibit a non-normal distribution.

–2. Constant Variance Assumptions: - Given the p-value of 0.0003989, which is less than the significance level $\alpha = 0.05$, we reject the null hypothesis. This implies a violation of the constant variance assumption, indicating that the errors demonstrate heteroscedasticity.

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.

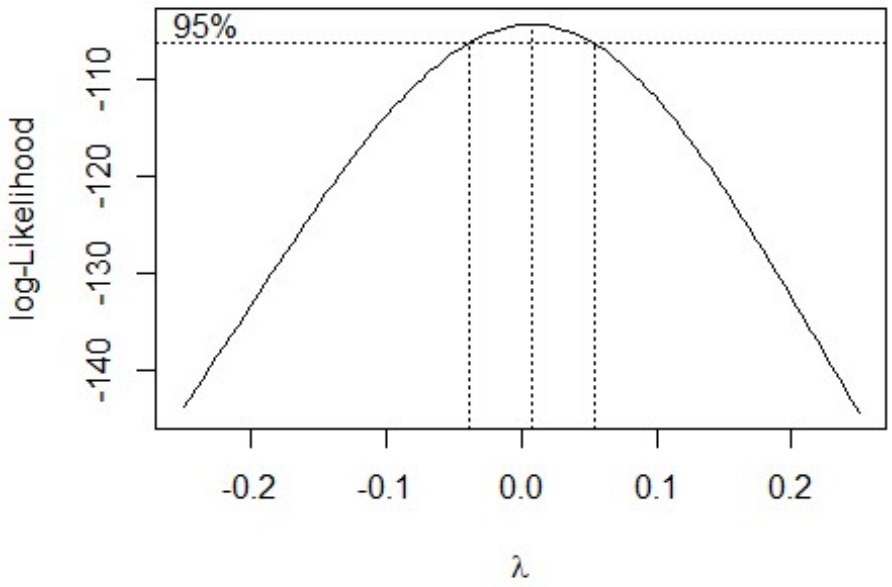


–In the provided plot, some data points are noticeably distant from the main cluster of data, indicating the presence of outliers and high-leverage points. These outliers and high-leverage points have the potential to exert significant influence on the analysis and results.

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,

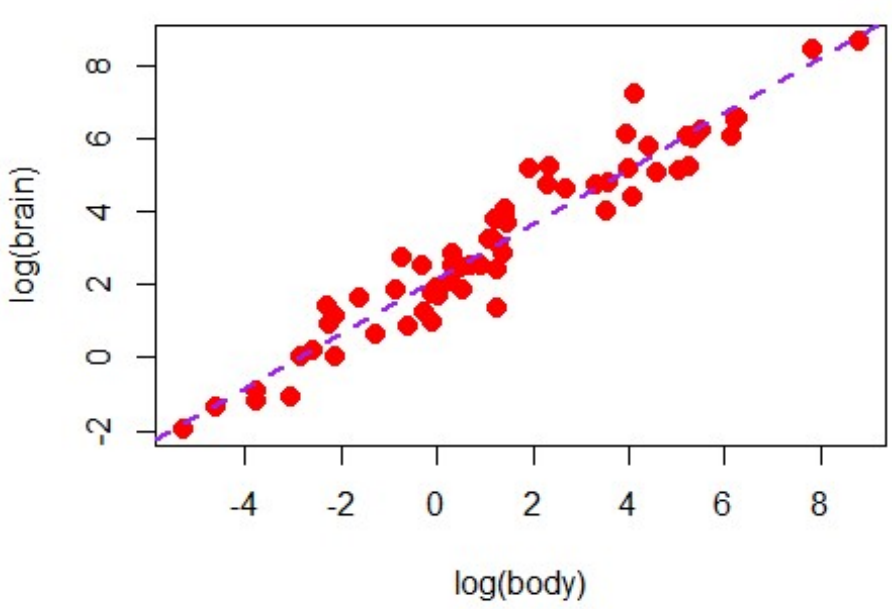
$$g_\lambda(\text{brain}) = \beta_0 + \beta_1 \log(\text{body}) + \varepsilon_i.$$

–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.



–Upon examining the plot, it is evident that the confidence interval (CI) encompasses zero at a lambda value of 0.0075. This outcome from the Box-Cox method provides support for opting for a log transformation.

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?



–The scatter plot suggests that a linear relationship is suitable, as the data points appear to exhibit a consistent dispersion around the regression line.

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.

–Certainly, let's rephrase your statements:

–1. **Normality Assumptions: - With a p-value of 0.5293, which exceeds the 0.05 significance level, we do not have sufficient evidence to reject the null hypothesis. Consequently, we conclude that the errors likely follow a normal distribution.

–2. **Constant Variance Assumptions (Homoscedasticity): - The p-value of 0.5977 surpasses the 0.05 significance level, leading us to fail to reject the null hypothesis. As a result, we conclude that the errors exhibit homoscedasticity, indicating that the assumption of constant variance is not violated.

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

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

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.

–The brain weight of a male Snorlax, with a body weight of 1014.1 pounds, is measured to be 848.5473 grams.

–We have a 90% confidence interval for the brain weight of a male Snorlax with a body weight of 1014.1 pounds, and it ranges from 257.823 to 2792.74 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 favor models with lower Root Mean Squared Error (RMSE) values. Please provide the RMSE values for both the model in Question 1 and the model in Question 4. Based on this criterion, we will determine which model is preferred.

– RMSE of model in Question 1 is 329.2768 – RMSE of model in Question 4 is 272.6149

So we prefer model used in the Question 4

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.

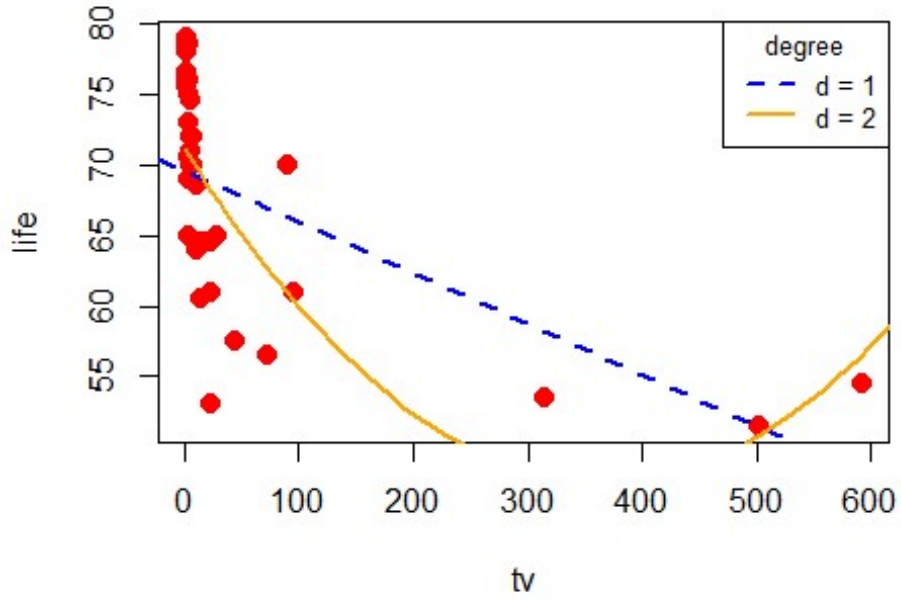
–The p-value associated with the linear term is 5.561e-05, which is below the commonly used significance level of 0.05. Consequently, we deem the linear term as statistically significant. As a result, we proceed to employ a quadratic model for further analysis.

–The p-value associated with the quadratic term is 0.005077, falling below the significance threshold of 0.05. Consequently, we deem the quadratic term as statistically significant, prompting us to proceed with fitting a cubic model.

–The p-value associated with the cubic term is 0.06766, exceeding the 0.05 significance level. Consequently, we fail to reject the null hypothesis, indicating that the cubic term is not statistically significant. As a result, we decide to retain the quadratic model.

- $\widehat{life}$ = 7.106e+01 -1.285e-01tv_i+1.756e-04tv_i^2

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.



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.

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

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.

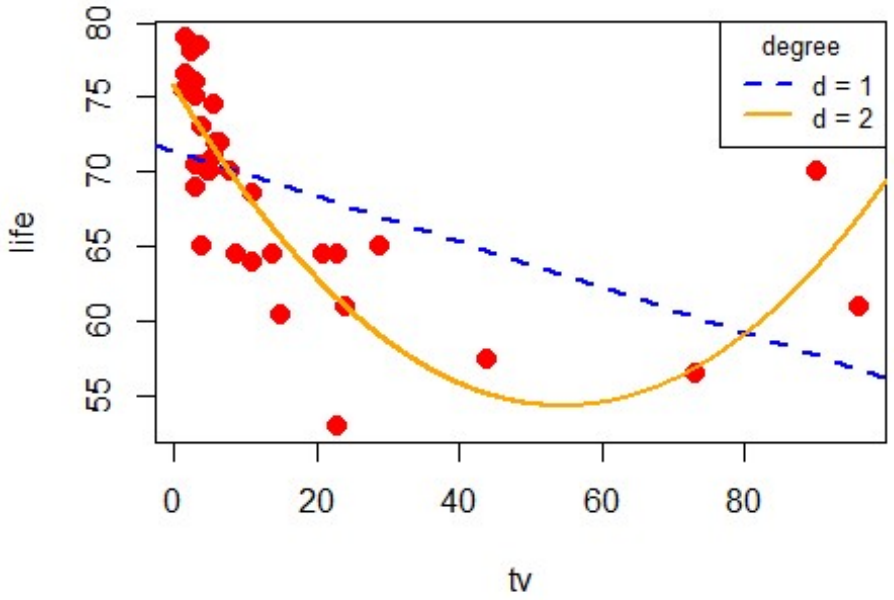
–The p-value associated with the linear term is 0.000783, indicating statistical significance at a significance level of 0.05. Consequently, we reject the null hypothesis for the linear term, signifying its importance. As a result, we proceed to model fitting with a quadratic term.

–The p-value associated with the quadratic term is 5.44e-07, indicating statistical significance at a significance level of 0.05. Consequently, we conclude that the quadratic model holds significance, prompting us to proceed with fitting a cubic model.

–The p-value associated with the cubic model is 0.08255, surpassing the 0.05 significance threshold. Consequently, the cubic model lacks significance, leading us to retain the quadratic model.

– $\widehat{life}$ = 75.835066 -0.7909274tv_i + 0.007278tv_i^2

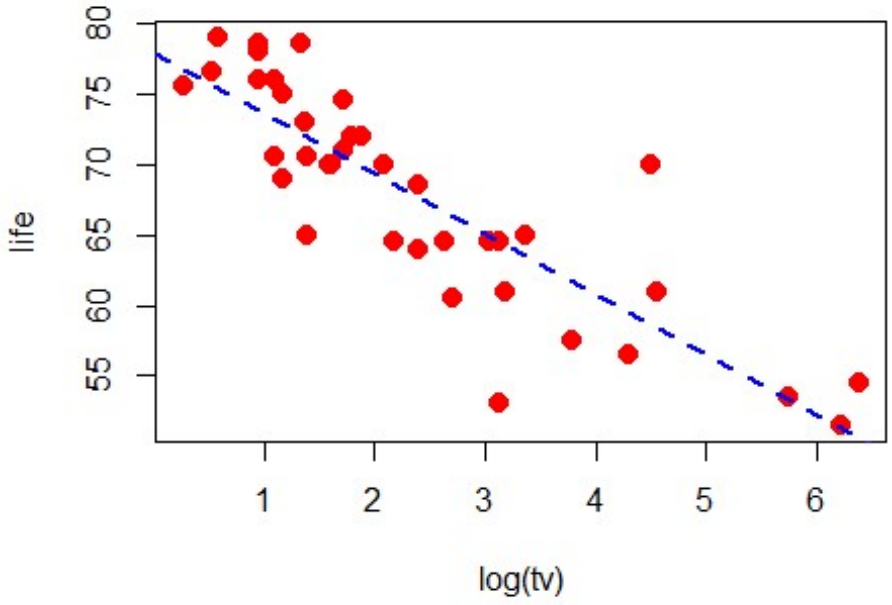
5. (6 points) Fit polynomial models of degree 1 and 2 with the high *leverage* data points you identified in Question 3 removed. Create a scatter plot of the data (**Note**: use the `subset` argument to `plot`) and add the fitted regression line for each polynomial model. Include the plot in your response.



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.

– $\widehat{life}$ = 77.8873-4.2597log(tv_i)

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.



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?

```
## [1] 0.467033

## [1] 0.7237519
```

The adjusted R^2 values for Model2 and Model6 are 0.467033 and 0.7237519, respectively. Given these values, Model6 is preferred over Model2 due to its higher adjusted R^2 value, indicating that a greater proportion of the variability in the dependent variable is explained by the independent variables in Model6 compared to Model2.

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.

–1. **Normality Assumptions: - With a p-value of 0.02152, which is below the 0.05 significance level, we reject the null hypothesis. This leads us to conclude that the errors exhibit a non-normal distribution.

–2. **Constant Variance Assumptions (Homoscedasticity): - The p-value of 0.07297 exceeds the significance level of 0.05. Consequently, we do not reject the null hypothesis, indicating that there is no evidence to suggest a violation of the constant variance assumption. Thus, we conclude that the errors demonstrate homoscedasticity, implying consistent variance across different levels of the independent variable.

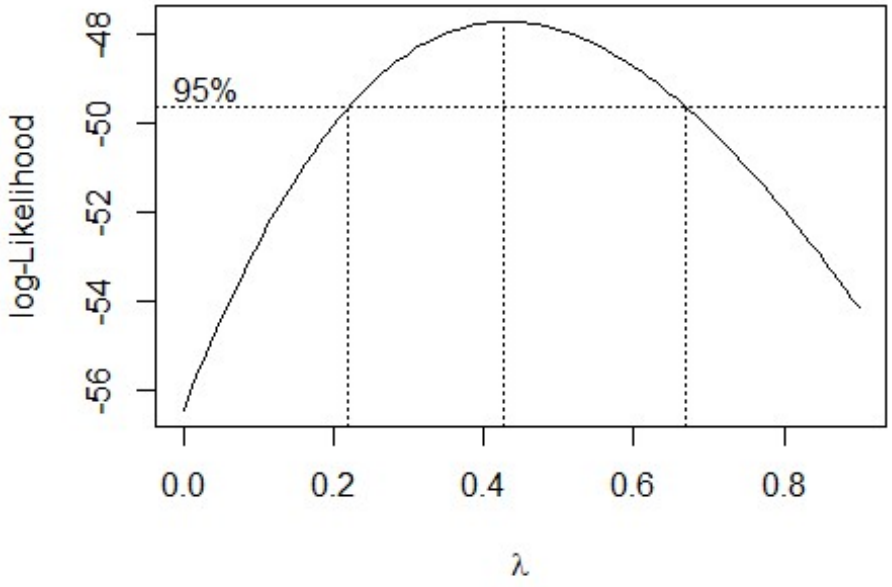
2. (10 points) Use the Box-Cox method to verify that $\sqrt{\text{dist}}$ is a "recommended" transformation of the response variable. That is, verify that the square-root transformation is among the "recommended" values of λ when considering,

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

–Certainly! Here's a rewritten version:

–Upon applying the boxcox transformation to the data, the relevant plot provides valuable insights. To pinpoint the property justifying the square-root transformation, a focused zoom onto the pertinent values is necessary.

–In particular, the plot exhibits a distinctive pattern indicative of the need for a square-root transformation. The chosen zoom level emphasizes the region where this transformation is most justified. Analyzing the shape of the plot in this segment further supports the decision to apply a square-root transformation to the data.



–Upon examining the plotted data, the calculated lambda (λ) value is found to be approximately 0.4318182, which is in close proximity to 0.5. The Box-Cox method supports our decision to employ a square-root transformation, as the lambda value aligns well with

the commonly used square-root transformation (where $\lambda = 0.5$). This suggests that the square-root transformation is a suitable choice for addressing the data characteristics and achieving a more desirable distribution.

3. (5 points) Fit the model justified in Question 2. That is, fit a model with $\sqrt{\text{dist}}$ 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.

– **Normality Assumptions: - With a p-value of 0.3143, exceeding the 0.05 significance level, we do not reject the null hypothesis. Therefore, we conclude that there is insufficient evidence to suggest a departure from a normal distribution in the errors.

– **Constant Variance Assumptions: - The p-value of 0.9157, surpassing the 0.05 significance level, leads us to retain the null hypothesis. Consequently, we conclude that there is no significant violation of the constant variance assumption, and the errors exhibit homoscedasticity.

Code Appendix

```
## Solution for Ex 1 Q 1



library(readr)
data(mammals, package='MASS')


mod_ols = lm(brain~body,data=mammals)

# For normality

shapiro.test(resid(mod_ols))

# For constant variation

library(lmtest)

bptest(mod_ols)


## Solution for Ex 1 Q 2

data(mammals, package='MASS')

# scatter plot of brain vs body
plot(brain~body,data=mammals,
     xlab = 'body', ylab = 'brain',
     pch = 20, cex = 2, col = 'red')
```

```r
mod_ols = lm(brain~body,data=mammals)

# fitted regression line
abline(mod_ols, col = 'blue', lty = 'dashed', lwd = 2)
```

```r
## Solution for Ex 1 Q 3

library(MASS)

data(mammals, package='MASS')

mod_ols = lm(brain~log(body),data=mammals)

bc = boxcox(mod_ols , lambda = seq(-0.25, 0.25, by = 0.10), plotit = TRUE)
```

```r
## Solution for Ex 1 Q 4

log_mod = lm(log(brain)~log(body),data=mammals)

#scatter plot of log(brain) vs log(body)

plot(log(brain)~log(body),data=mammals,
     xlab = 'log(body)', ylab = 'log(brain)',
     pch = 20, cex = 2, col = 'red')

# fitted regression line
abline(log_mod, col = 'purple', lty = 'dashed', lwd = 2)
```

```r
## Solution for Ex 1 Q 5

# For normality

shapiro.test(resid(log_mod))

# For constant variance
library(lmtest)

bptest(log_mod)
```

```r
## Solution for Ex 1 Q 6
```

```r
log_mod = lm(log(brain)~log(body),data=mammals)

inf_ids=which(cooks.distance(log_mod)>4/length(resid(log_mod)))
inf_ids


## Solution for Ex 1 Q 7

new_data = data.frame(body = 459.988022)

pred=predict(log_mod,newdata = new_data,interval = "prediction",level = 0.90)

exp(pred)


## Solution for Ex 1 Q 8

mod_1 = lm(brain~body,data=mammals)
log_mod_4 = lm(log(brain)~log(body),data=mammals)

sqrt(mean((mammals$brain - predict(mod_1))^2))

sqrt(mean((mammals$brain - exp(predict(log_mod_4)))^2))


## Solution for Ex 2 Q 1

library(readr)
data(tvdoctor, package='faraway')

poly_1 = lm(life ~ tv, data = tvdoctor)

summary(poly_1)




## Solution for Ex 2 Q 1

poly_2 = lm(life ~ poly(tv, 2, raw = TRUE), data = tvdoctor)

summary(poly_2)


## Solution for Ex 2 Q 1

poly_3 = lm(life ~ poly(tv, 3, raw = TRUE), data = tvdoctor)
```

```r
summary(poly_3)


## Solution for Ex 2 Q 2

data(tvdoctor, package='faraway')

poly_1 = lm(life ~ tv, data = tvdoctor)


plot(life ~ tv, data = tvdoctor,
     xlab = 'tv', ylab = 'life',
     pch = 20, cex = 2, col = 'red')

abline(poly_1, col = 'blue', lty = 'dashed', lwd = 2)

poly_2 = lm(life ~ poly(tv, 2, raw = TRUE), data = tvdoctor)


xplot = seq(0, 800, by = 10)
lines(xplot, predict(poly_2, newdata = data.frame(tv = xplot)),
      col = "orange", lwd = 2)

legend("topright", title = "degree", cex = 0.8,
       legend = c("d = 1", "d = 2"),
       lwd = 2, lty = c(2, 1),
       col = c("blue", "orange"))

## Solution for Ex 2 Q 3

high_lev = which(hatvalues(poly_2) > 2 * mean(hatvalues(poly_2)))
high_lev


## Solution for Ex 2 Q 4

high_lev_rmvd = which(hatvalues(poly_2) <= 2 * mean(hatvalues(poly_2)))

poly1_new = lm(life ~ tv, data = tvdoctor, subset = high_lev_rmvd)

summary(poly1_new)




## Solution for Ex 2 Q 4
```

```r
poly2_new = lm(life ~ poly(tv, 2, raw = TRUE), data = tvdoctor,
                      subset = high_lev_rmvd)

summary(poly2_new)
```

```r
## Solution for Ex 2 Q 4

poly3_new = lm(life ~ poly(tv, 3, raw = TRUE), data = tvdoctor,
                      subset = high_lev_rmvd)

summary(poly3_new)
```

```r
## Solution for Ex 2 Q 5

poly_2 = lm(life ~ poly(tv, 2, raw = TRUE), data = tvdoctor)


high_lev_rmvd = which(hatvalues(poly_2) <= 2 * mean(hatvalues(poly_2)))


poly_1 = lm(life ~ tv, data = tvdoctor,subset = high_lev_rmvd)


plot(life ~ tv, data = tvdoctor, subset = high_lev_rmvd,
     xlab = 'tv', ylab = 'life',
     pch = 20, cex = 2, col = 'red')


abline(poly_1, col = 'blue', lty = 'dashed', lwd = 2)

poly_2 = lm(life ~ poly(tv, 2, raw = TRUE), data = tvdoctor,subset =
high_lev_rmvd)


xplot = seq(0, 100, by = 0.01)
lines(xplot, predict(poly_2, newdata = data.frame(tv = xplot)),
      col = "orange", lwd = 2)
```

```r
# add a legend
legend("topright", title = "degree", cex = 0.8,
       legend = c("d = 1", "d = 2"),
       lwd = 2, lty = c(2, 1),
       col = c("blue", "orange"))

## Solution for Ex 2 Q 6

log_mod = lm(life~log(tv),data = tvdoctor)

summary(log_mod)

## Solution for Ex 2 Q 7


plot(life~log(tv),data=tvdoctor,
     xlab = 'log(tv)', ylab = 'life',
     pch = 20, cex = 2, col = 'red')

mod = lm(life~log(tv),data = tvdoctor)

abline(mod, col = 'blue', lty = 'dashed', lwd = 2)


mod_2 = lm(life ~ poly(tv, 2, raw = TRUE), data = tvdoctor)

mod_6 = lm(life~log(tv),data = tvdoctor)

summary(mod_2)$adj.r.squared

summary(mod_6)$adj.r.squared


## Solution for Ex 3 Q 1


library(readr)
data(cars, package='faraway')


mod = lm(dist~speed,data=cars)

# For normality

shapiro.test(resid(mod))

# For constant variance
library(lmtest)
```

```r
bptest(mod)


## Solution for Ex 3 Q 2

library(readr)
data(cars, package='faraway')


mod = lm(dist~speed,data=cars)

bc = boxcox(mod , lambda = seq(-0.00, 0.90, by = 0.05), plotit = TRUE)



## Solution for Ex 3 Q 3


library(readr)
data(cars, package='faraway')


mod = lm(sqrt(dist)~speed,data=cars)

# For normality

shapiro.test(resid(mod))

# For constant variance
library(lmtest)

bptest(mod)
```