STA 5207: Homework 9

Manjunath Aineni. MA23BI

Due: Wednesday, November 29th by 11:59 PM

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 (The `stackloss` Data Set) [50 points]

For this exercise, we will use the `stackloss` data set from the `faraway` package. You can also find the data in `stackloss.csv` on Canvas. The data set contains operational data of a plant for the oxidation of ammonia to nitric acid. There are 21 observations and the following 4 variables in the data set

- `Air Flow`: Flow of cooling air.
- `Water Temp`: Cooling Water Inlet Temperature.
- `Acid Conc.`: Concentration of acid [per 1000, minus 500].
- `stack.loss`: Stack loss.

In the following exercise, we will use `stack.loss` as the response and `Air Flow`, `Water Temp`, and `Acid Conc.` as predictors.

1. (4 points) Perform OLS regression with `stack.loss` as the response and the remaining variables as predictors. Check the normality assumption using a hypothesis test at the $\alpha = 0.05$ significance level. Report the p-value of the test and your conclusions.

With a p-value of 0.8186 and a test statistic value of 0.97399, the results surpass the significance level of 0.05. We come to the conclusion that the errors have a normal distribution and do not reject the null hypothesis.

2. (4 points) Perform LAD regression with `stack.loss` as the response and the remaining variables as predictors. Report the estimated regression equation for this model.

The estimated regression equation =

$\hat{stack}.loss$ = -39.68986 + 0.83188 $Air.Flow_i$ + 0.57391 $Water.Temp_i$ - 0.06087 $AcidConc._i$

3. (4 points) Perform robust regression using Huber's method with `stack.loss` as the response and the remaining variables as predictors. Use `maxit = 100` iterations for IRWLS. Report the estimated regression equation for this model.

The estimated regression equation =

$\hat{stack}.loss$ = -41.0265 + 0.8294 $Air.Flow_i$ + 0.9261 $Water.Temp_i$ - 0.1278 $AcidConc._i$

4. (4 points) Calculate and report the 95% confidence intervals for the intercept and the slope parameters of the model you fit in Question 3 using the residual bootstrap. Use $R = 2000$ bootstrap samples, `method = 'residual'`, and set a seed of 42.

95% confidence intervals in below table:

Intercept	(-64.156,-19.627)

Air.Flow	(0.567,1.064)
Water.Temp	(0.230,1.607)
Acid.Conc.	(-0.415,0.164)

5. (5 points) Create and report a table comparing the OLS, LAD, and Huber estimates for the intercept *and* slope parameters. Bold entries in the table that are significant at the $\alpha = 0.05$ significance level (for OLS use the standard t-test). Recall that for LAD, you should set `alpha = 0.05` in the model `summary`.

	Intercept	Air.Flow	Water.Temp	Acid.Conc.
OLS	-39.919	**0.715**	**1.295**	-0.152
LAD	-39.689	**0.831**	**0.573**	-0.060
Huber	-41.026	**0.829**	**0.926**	-0.127

6. (3 points) Use the OLS model from Question 1 to check for any highly influential data points. Report the observations you determine are highly influential.

In this case, observation 21 is highly influential.

7. (3 points) Identify the observations with weights less than one in the Huber fit from Question 3. Report these observations along with their weights. Which (if any) of these observations also have high influence according to Question 6.

```
## # A tibble: 3 x 2
##   Observation Weight
##   <chr>       <chr>
## 1 21          0.368141062112395
## 2 4           0.504940940772636
## 3 3           0.785887104227537
```

In this case, we see that observation 21, 4 and 3 have weights less than 1.

From these only obervation 21 is also highly influential according to Question 6

8. (5 points) Fit an OLS regression model with the observations that were highly influential removed. Create a table comparing the OLS estimates from the model in Question 1 with these new estimates. Bold entries in the table that are statistically significant at the $\alpha = 0.05$ significance level according to a standard t-test for each model.

	Intercept	Air.Flow	Water.Temp	Acid.Conc.
OLS	-39.919	**0.715**	**1.295**	-0.152
OLS [Refit]	-43.704	**0.889**	**0.816**	-0.107

9. (5 points) Fit an LAD regression model with the observations that were highly influential removed. Create a table comparing the parameter estimates from the LAD model in Question 2 with these new estimates. Bold entries in the table that are statistically significant at the $\alpha = 0.05$ significance level.

	Intercept	Air.Flow	Water.Temp	Acid.Conc.
LAD	-39.689	**0.831**	**0.573**	-0.060
LAD [Refit]	-39.986	**0.834**	**0.563**	-0.056

10. (5 points) Perform robust regression using Huber's method with the observations that were highly influential removed. Use `maxit = 100` iterations of IRWLS. Calculate and report the 95% confidence

intervals for the intercept and slope parameters of this model using the residual bootstrap. Use $R = 2000$ bootstrap samples, `method = 'residual'`, and set a seed of 42.

95% confidence intervals in below table:

Intercept	(-59.951,-22.892)
Air.Flow	(0.698,1.143)
Water.Temp	(0.061,1.266)
Acid.Conc.	(-0.336,0.140)

11. (5 points) Create a table comparing the parameter estimates from the model using Huber's method in Question 3 with the new estimates from the model in Question 10. Bold entries in the table that are statistically significant at the $\alpha = 0.05$ significance level.

	Intercept	Air.Flow	Water.Temp	Acid.Conc.
Huber	-41.026	**0.829**	**0.926**	-0.127
Huber [Refit]	-42.841	**0.918**	**0.685**	-0.107

12. (3 points) Based on your answers to Questions 8 - 11 and the difference in the slope estimates, which method is most resistant to the highly influential observations. Justify your answer.

LAD method is most resistant to the highly influential observations as we observe that there is no much difference in the slope estimates after refitting the LAD model with highly influential observations removed.

In LAD –> Air.Flow estimate is almost same and Water.Temp & Acid.Conc. estimates are varied by 0.01 whereas variation in slope estimates is much higher in case of OLS & Huber methods.

Exercise 2 (The `Duncan` Data Set) [50 points]

For this exercise, we will use the `Duncan` data set from the `carData` package. You can also find the data in `Duncan.csv` on Canvas. The data set contains information on the prestige and other characteristics of 45 U.S. occuptations in 1950. There are 45 obervations and the following 4 variables in the data set

- `type`: Type of occupation (professional and managerial, white-collar, and blue-collar).
- `income`: Percentage of occupational incumbents in the 1950 U.S. Census who earned $3,500 or more per year (about $36,000 in 2017 U.S. dollars).
- `education`: Percentage of occupational incumbents in 1950 who were high school graduates.
- `prestige`: Percentage of respondents in a social survey who rated the occupation as "good" or better in prestige.

In the following exercise, we will use `prestige` as the response and `income` and `education` as predictors.

1. (4 points) Perform OLS regression with `prestige` as the response and `income` and `education` as predictors. Check the normality assumption using a hypothesis test at the $\alpha = 0.05$ significance level. Report the p-value of the test and your conclusions.

The value of the test statistic is 0.98254 with a p-value of 0.7234 , which is greater than the significance level of 0.05. We donot reject the null hypothesis and conclude that the errors have a normal distribution.

2. (4 points) Perform LAD regression with `prestige` as the response and `income` and `education` as predictors. Report the estimated regression equation for this model.

estimated regression equation is

$\hat{prestige}$ = -6.408 + 0.747 *income* + 0.458 *education*

3. (4 points) Perform robust regression using Huber's method with `prestige` as the response `income` and `education` as predictors. Use maxit = 50 iterations for IRWLS. Report the estimated regression equation for this model.

estimated regression equation is

$\hat{prestige}$ = -7.110 + 0.701 *income* + 0.485 *education*

4. (4 points) Calculate and report the 95% confidence intervals for the intercept and the slope parameters of the model you fit in Question 3 using the residual bootstrap. Use $R = 2000$ bootstrap samples, `method = 'residual'`, and set a seed of 42.

95% confidence intervals in below table:

Intercept	(-15.235,0.120)
income	(0.482,0.931)
education	(0.310,0.669)

5. (5 points) Create and report a table comparing the OLS, LAD, and Huber estimates for the intercept *and* slope parameters. Bold entries in the table that are significant at the $\alpha = 0.05$ significance level (for OLS use the standard t-test). Recall that for LAD, you should set `alpha = 0.05` in the model `summary`.

	Intercept	income	education
OLS	-6.064	**0.598**	**0.545**
LAD	-6.408	**0.747**	**0.458**
Huber	-7.110	**0.701**	**0.485**

6. (3 points) Use the OLS model from Question 1 to check for any highly influential data points. Report the observations you determine are highly influential.

In this case, minister, reporter and conductor are highly influential.

7. (3 points) Identify the five observations that have the lowest weights in the Huber fit from Question 3. Report these observations along with their weights. Which (if any) of these observations also have high influence according to Question 6.

```
## # A tibble: 5 x 2
##   Profession      Weight
##   <chr>           <chr>
## 1 minister        0.344663638975194
## 2 reporter        0.441726568838591
## 3 insurance.agent 0.53356863524905
## 4 conductor       0.538620423453824
## 5 contractor      0.552262762725598
```

All 3 observations minister,reporter and conductor are also highly influential and have lowest weights.

8. (5 points) Fit an OLS regression model with the observations that were highly influential removed. Create a table comparing the OLS estimates from the model in Question 1 with these new estimates. Bold entries in the table that are statistically significant at the $\alpha = 0.05$ significance level according to a standard t-test for each model.

	Intercept	income	education
OLS	-6.064	**0.598**	**0.545**
OLS [Refit]	-7.241	**0.877**	**0.353**

9. (5 points) Fit an LAD regression model with the observations that were highly influential removed. Create a table comparing the parameter estimates from the LAD model in Question 2 with these new estimates. Bold entries in the table that are statistically significant at the $\alpha = 0.05$ significance level.

	Intercept	income	education
LAD	-6.408	**0.747**	**0.458**
LAD [Refit]	-8.616	**0.810**	**0.444**

10. (5 points) Perform robust regression using Huber's method with the observations that were highly influential removed. Use `maxit = 50` iterations of IRWLS. Calculate and report the 95% confidence intervals for the intercept and slope parameters of this model using the residual bootstrap. Use $R = 2000$ bootstrap samples, `method = 'residual'`, and set a seed of 42.

95% confidence intervals in below table:

Intercept	(-14.032,-1.700)
income	(0.629,1.082)
education	(0.195,0.562)

11. (5 points) Create a table comparing the parameter estimates from the model using Huber's method in Question 3 with the new estimates from the model in Question 10. Bold entries in the table that are statistically significant at the $\alpha = 0.05$ significance level.

	Intercept	income	education
Huber	-7.110	**0.701**	**0.485**
Huber [Refit]	-7.610	**0.854**	**0.383**

12. (3 points) Based on your answers to Questions 8 - 11 and the difference in the slope estimates, which method is most resistant to the highly influential observations. Justify your answer.

LAD method is most resistant to the highly influential observations as we observe that there is very less variation in the slope estimates after refitting the LAD model with highly influential observations removed.

In LAD –> Income estimate is varied by 0.5 and Education estimates is varied by 0.01 compared to variation in slope estimates in case of OLS & Huber methods after highly influential observations removed.

Code Appendix

```
## Code for Problem 1, Question 1

data(stackloss,package='faraway')

stack_model_ols=lm(stack.loss ~ . , data=stackloss)

shapiro.test(resid(stack_model_ols))

## Code for Problem 1, Question 2
library(quantreg)

stack_model_lad = rq(stack.loss ~ . , data = stackloss)


coef(stack_model_lad)


## Code for Problem 1, Question 3
```

```r
library(MASS)
stack_model_hub = rlm(stack.loss ~ . , maxit = 100, data = stackloss)


coef(stack_model_hub)

## Code for Problem 1, Question 4

library(car)
set.seed(42)

Confint(Boot(stack_model_hub, R = 2000, method = 'residual'), level = 0.95)

## Code for Problem 1, Question 5

summary(stack_model_ols)

summary(stack_model_lad, alpha = 0.05)

summary(stack_model_hub)
set.seed(42)
Confint(Boot(stack_model_hub, R = 2000, method = 'residual'), level = 0.95)

## Code for Problem 1, Question 6

which(cooks.distance(stack_model_ols) > 4/length(resid(stack_model_ols)))

## Code for Problem 1, Question 7

library(tidyverse)

# order of the weights going from lowest to heighest
ord = order(stack_model_hub$w)

head(as_tibble(cbind(
    'Observation' = row.names(stackloss)[ord],
    'Weight' = stack_model_hub$w[ord]
)),3)

## Code for Problem 1, Question 8

stack_model_ols_no_infl = lm(stack.loss ~ . , data = stackloss,
                             subset = -c(21))

summary(stack_model_ols_no_infl)


## Code for Problem 1, Question 9

stack_model_lad_no_infl = rq(stack.loss ~ . , data = stackloss,
                             subset = -c(21))

summary(stack_model_lad_no_infl, alpha = 0.05)
```

```r
## Code for Problem 1, Question 10

stack_model_hub_no_infl = rlm(stack.loss ~ . , maxit = 100, data = stackloss[-c(21),])


set.seed(42)
Confint(Boot(stack_model_hub_no_infl, R = 2000, method = 'residual'))

## Code for Problem 1, Question 11


summary(stack_model_hub_no_infl)

## Code for Problem 2, Question 1

data(Duncan,package='carData')

duncan_model_ols=lm(prestige ~ . - type, data=Duncan)


shapiro.test(resid(duncan_model_ols))

## Code for Problem 2, Question 2

library(quantreg)

duncan_model_lad = rq(prestige ~ . -type, data = Duncan)


coef(duncan_model_lad)


## Code for Problem 2, Question 3

duncan_model_hub = rlm(prestige ~ . -type , maxit = 50, data = Duncan)

coef(duncan_model_hub)

## Code for Problem 2, Question 4

library(car)
set.seed(42)

Confint(Boot(duncan_model_hub, R = 2000, method = 'residual'), level = 0.95)


## Code for Problem 2, Question 5

summary(duncan_model_ols)

summary(duncan_model_lad, alpha = 0.05)

summary(duncan_model_hub)
```

```r
set.seed(42)
Confint(Boot(duncan_model_hub, R = 2000, method = 'residual'), level = 0.95)

## Code for Problem 2, Question 6

which(cooks.distance(duncan_model_ols) > 4/length(resid(duncan_model_ols)))


## Code for Problem 2, Question 7

library(tidyverse)

# order of the weights going from lowest to heighest
ord = order(duncan_model_hub$w)

head(as_tibble(cbind(
    'Profession' = row.names(Duncan)[ord],
    'Weight' = duncan_model_hub$w[ord]
)),5)


## Code for Problem 2, Question 8

duncan_model_ols_no_infl = lm(prestige ~ . -type, data = Duncan,
                           subset = -c(6,9,16))

summary(duncan_model_ols_no_infl)

## Code for Problem 2, Question 9

duncan_model_lad_no_infl = rq(prestige ~ . -type , data = Duncan,
                           subset = -c(6,9,16))

summary(duncan_model_lad_no_infl, alpha = 0.05)


## Code for Problem 2, Question 10

duncan_model_hub_no_infl = rlm(prestige ~ . -type , maxit = 50, data = Duncan[-c(6,9,16),])


set.seed(42)
Confint(Boot(duncan_model_hub_no_infl, R = 2000, method = 'residual'))

## Code for Problem 1, Question 11


summary(duncan_model_hub_no_infl)
```