STA 5207: Homework 10

Due: Friday, April 12th 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.

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.

```r
model_ols = lm(stack.loss ~ ., data = stackloss)

shapiro.test(resid(model_ols))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model_ols)
## W = 0.97399, p-value = 0.8186
```

```r
summary(model_ols)$coeff
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -39.9196744 11.8959969 -3.3557234 3.750307e-03
## Air.Flow      0.7156402  0.1348582  5.3066130 5.799025e-05
## Water.Temp    1.2952861  0.3680243  3.5195672 2.630054e-03
## Acid.Conc.   -0.1521225  0.1562940 -0.9733098 3.440461e-01
```

The test statistic is.974 with a p-value of .819, which is greater than the significance level so we do not reject the null hypothesis and conclude the errors have a normal distribution.

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.

```r
model_lad = rq(stack.loss ~ ., data = stackloss)

summary(model_lad, alpha = 0.05)
```

```
##
## Call: rq(formula = stack.loss ~ ., data = stackloss)
##
## tau: [1] 0.5
##
## Coefficients:
##             coefficients lower bd  upper bd
## (Intercept) -39.68986    -53.79464 -24.49145
## Air.Flow      0.83188      0.50909   1.16751
## Water.Temp    0.57391      0.27151   3.03726
## Acid.Conc.   -0.06087     -0.27772   0.01534
```

$$stack.loss = -39.69 + .831\text{Air.flow}_i + .573\text{Water.temp}_i - .061\text{Acid.conc}_i$$

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.

```
model_hub = rlm(stack.loss ~ ., maxit = 100, data = stackloss)

summary(model_hub)
```

```
##
## Call: rlm(formula = stack.loss ~ ., data = stackloss, maxit = 100)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.91753 -1.73127  0.06187  1.54306  6.50163
##
## Coefficients:
##             Value    Std. Error t value
## (Intercept) -41.0265   9.8073    -4.1832
## Air.Flow      0.8294   0.1112     7.4597
## Water.Temp    0.9261   0.3034     3.0524
## Acid.Conc.   -0.1278   0.1289    -0.9922
##
## Residual standard error: 2.441 on 17 degrees of freedom
```

$$stack.loss = -41.03 + .830\text{Air.flow}_i + .926\text{Water.temp}_i - .128\text{Acid.conc}_i$$

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.

```
set.seed(42)

Confint(Boot(model_hub, R = 2000, method = 'residual'))
```

```
## Bootstrap bca confidence intervals
##
##                 Estimate        2.5 %      97.5 %
## (Intercept) -41.0265311 -64.1560502 -19.6266742
## Air.Flow      0.8293739   0.5672855   1.0637728
## Water.Temp    0.9261082   0.2303327   1.6068223
## Acid.Conc.   -0.1278492  -0.4154434   0.1642079
```

Intercept	(-64.16,-19.63)
Air.Flow	(.567,1.063)
Water.Temp	(.230,1.607)
Acid.Conc.	(-.415,.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	**-39.92**	**.716**	**1.30**	-.152
LAD	-39.69	.832	.574	**-.061**
HUB	-41.02	**.823**	**.926**	-.128

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.

```r
which(cooks.distance(model_ols) > 4/length(resid(model_ols)))
```

```
## 21
## 21
```

At observation 21 there is a highly influential point.

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.

```r
which(model_hub$w < 1)
```

```
## [1]  3  4 21
```

Observations 3, 4, and 21 have weights less than one. 21 is also a highly influential point.

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.

```r
model_ols = lm(stack.loss ~ ., data = stackloss, subset = -21)
summary(model_ols)$coeff
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -43.7040310  9.4915652 -4.6045125 2.930291e-04
## Air.Flow      0.8891082  0.1188476  7.4810750 1.309021e-06
## Water.Temp    0.8166199  0.3250294  2.5124489 2.308829e-02
## Acid.Conc.   -0.1071414  0.1245414 -0.8602872 4.023381e-01
```

	Intercept	Air.Flow	Water.temp	Acid.Conc.
OLS	**-39.92**	**.716**	**1.30**	-.152
OLS (refit)	**-43.7**	**.889**	**.817**	-.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.

```r
model_lad = rq(stack.loss ~ ., data = stackloss, subset = -21)

summary(model_lad, alpha = .05)
```

```
##
## Call: rq(formula = stack.loss ~ ., data = stackloss, subset = -21)
##
## tau: [1] 0.5
##
## Coefficients:
##             coefficients lower bd  upper bd
## (Intercept) -39.98645    -55.09381 -29.61200
## Air.Flow      0.83469      0.69070   1.19001
## Water.Temp    0.56369      0.18052   1.61761
## Acid.Conc.   -0.05691     -0.40670   0.02282
```

	Intercept	Air.Flow	Water.temp	Acid.Conc.
LAD	-39.69	.832	.574	**-.061**
LAD (refit)	-40	.834	.564	**-.0569**

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.

```r
model_hub = rlm(stack.loss ~ ., maxit = 100, data = stackloss[-21,])

Confint(Boot(model_hub, R = 2000, method = 'residual'))
```

```
## Bootstrap bca confidence intervals
##
##                Estimate         2.5 %     97.5 %
## (Intercept) -42.8414925 -58.713964224 -21.175925
## Air.Flow      0.9183673   0.704056299   1.154169
## Water.Temp    0.6854204  -0.008075849   1.266582
## Acid.Conc.   -0.1077664  -0.354273232   0.120069
```

```r
summary(model_hub)$coeff
```

```
##                   Value Std. Error    t value
## (Intercept) -42.8414925  8.6192837 -4.9704238
## Air.Flow      0.9183673  0.1079255  8.5092734
## Water.Temp    0.6854204  0.2951590  2.3222074
## Acid.Conc.   -0.1077664  0.1130959 -0.9528755
```

Intercept	(-58.713,-21.18)
Air.Flow	(.704,1.154)
Water.Temp	(-.008,1.267)
Acid.Conc.	(-.354,.120)

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.
HUB	-41.02	**.823**	**.926**	-.128
HUB (refit)	-42.8	**.918**	**.685**	-.108

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.

    The LAD method seems to be most resistant as the intercept and slope estimates do not change much after removing the influential points.

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.

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

shapiro.test(resid(model_ols))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model_ols)
## W = 0.98254, p-value = 0.7234
```

```
summary(model_ols)$coeff
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -6.0646629 4.27194117 -1.419650 1.630896e-01
## income       0.5987328 0.11966735  5.003310 1.053184e-05
## education    0.5458339 0.09825264  5.555412 1.727192e-06
```

The test statistic is.983 with a p-value of .723, which is greater than the significance level so we do not reject the null hypothesis and conclude 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.

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

summary(model_lad, alpha = .05)$coeff
```

```
##             coefficients    lower bd   upper bd
## (Intercept)   -6.4082569 -15.1806282 -3.1478942
## income         0.7477064   0.3875688  0.9120119
## education      0.4587156   0.1545138  0.7342714
```

$$\text{prestige}_i = -6.41 + .748\text{income}_i + .459\text{education}_i$$

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.

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

summary(model_hub)$coeff
```

```
##                  Value Std. Error   t value
## (Intercept) -7.1107028 3.88131509 -1.832034
## income       0.7014493 0.10872497  6.451593
## education    0.4854390 0.08926842  5.437970
```

$$\text{prestige}_i = -7.11 + .701\text{income}_i + .485\text{education}_i$$

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.

```
set.seed(42)

Confint(Boot(model_hub, R = 2000, method = 'residual'))
```

```
## Bootstrap bca confidence intervals
##
##               Estimate       2.5 %    97.5 %
## (Intercept) -7.1107028 -15.2358418 0.1205594
## income       0.7014493   0.4821324 0.9310555
## education    0.4854390   0.3102359 0.6698057
```

Intercept	(-15.24, .121)
Income	(.482, .931)
Education	(.310, .670)

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	-6.065	**.599**	**.546**
LAD	-6.408	.748	.459

	Intercept	Income	Education
HUB	**-7.11**	**.701**	.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.

```r
which(cooks.distance(model_ols) > 4/length(resid(model_ols)))
```

```
##  minister  reporter conductor 
##         6         9        16 
```

Minister, reporter, and conductor at observation 6,9,16 respectively 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.

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


as_tibble(cbind(
    'Occupation' = row.names(Duncan)[ord],
    'Weight' = model_hub$w[ord]       # ordered weights from huber
))
```

```
## # A tibble: 45 x 2
##    Occupation         Weight           
##    <chr>              <chr>            
##  1 minister           0.344663638975194
##  2 reporter           0.441726568838591
##  3 insurance.agent    0.53356863524905 
##  4 conductor          0.538620423453824
##  5 contractor         0.552262762725598
##  6 machinist          0.570380348743291
##  7 store.clerk        0.618704298971657
##  8 mail.carrier       0.690874843443827
##  9 factory.owner      0.706152141835959
## 10 streetcar.motorman 0.833087108675877
## # i 35 more rows
```

Minister, reporter, insurance.agent, conductor, and contractor have the lowest weights. Minister, reporter, and conductor are also the highly influential data points.

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

```r
model_ols = lm(prestige ~ . - type, data = Duncan, subset = -c(6,9,16))
summary(model_ols)$coeff
```

```
##               Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) -7.2414232 3.39052345 -2.135783 3.903151e-02
## income       0.8772736 0.11283928  7.774541 1.897411e-09
## education    0.3538098 0.09163487  3.861082 4.138738e-04
```

	Intercept	Income	Education
OLS	-6.065	**.599**	**.546**
OLS (refit)	**-7.24**	**.877**	**.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.

```
model_lad = rq(prestige ~ . - type, data = Duncan,subset = -c(6,9,16))
summary(model_lad, alpha = .05)$coeff
```

```
##             coefficients    lower bd   upper bd
## (Intercept)   -8.6163070 -15.2319356 -2.0573109
## income         0.8105516   0.7272188  1.0855510
## education      0.4448441   0.1066670  0.5248255
```

	Intercept	Income	Education
LAD	-6.408	.748	.459
LAD (refit)	-8.62	.811	.445

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.

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

set.seed(42)

Confint(Boot(model_hub, R = 2000, method = 'residual'))
```

```
## Bootstrap bca confidence intervals
##
##               Estimate       2.5 %     97.5 %
## (Intercept) -7.6103423 -14.0326678 -1.7009350
## income       0.8549488   0.6295368  1.0821971
## education    0.3836578   0.1951765  0.5627294
```

Intercept	(-14.03, -1.701)
Income	(.630, 1.082)
Education	(.195, .563)

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.

```
summary(model_hub)$coeff
```

```
##                  Value Std. Error   t value
## (Intercept) -7.6103423 3.00880350 -2.529358
## income       0.8549488 0.10013534  8.537933
## education    0.3836578 0.08131822  4.717981
```

	Intercept	Income	Education
HUB	**-7.11**	**.701**	.485
HUB (refit)	-7.610	**.855**	**.384**

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.

    I would choose the OLS method. Both LAD and HUB seem to be affected the same way as OLS after removing the highly influential points, so it would be to our benefit to choose the simpler overall model.