Lab 3

Ve406

Due: **18 November 2018, 11:40am**

Instructions

- Please report your findings in this **R Markdown** file by removing any text that you do not need. Include all your **R**-code in `chunks` and your comments and findings as texts.
- Recall **R**-chunks that are not necessary to report (like the package loading and the working directory path) can be exempt from printing by using the option `echo=FALSE` in the setting up of the `chunk`.
- This lab is about unususal points, heteroskedasticity and correlated errors.

Task 1 (8 points)

The data `chem_pro` is the dataset about a particular chemical process we considered in class.

(a) (1 point)

Succesfully render this file.

(b) (1 point)

Clean `chem_pro.df` according to what we have discussed in class.

```
chem_pro.df = read.table(file = chem_pro.csv, sep = ",", header = TRUE)
```

(c) (1 point)

Produce the pairs plot of all the variables in `chem_pro.df` like the one I showed in class.

(d) (1 point)

Construct the following model, then produce all the usual regression diagnostic plots for `chem_pro.LM`.

```
chem_pro.LM = lm(yield~conversion+flow+ratio, data = chem_pro.df)
```

- Standardised residual Vs fitted value
- Standardised residual Vs conversion
- Standardised residual Vs flow

- Standardised residual Vs ratio
- Residual Vs Previous Residual
- Residual Autcorrelation (ACF)
- Q-Q Normal

(e) (1 point)

Compute VIF for `chem_pro.LM` according to the definition, then compare it with the values found in class.

(f) (1 point)

Produce a boxplot of Leverage Scores for `chem_pro.LM` like the one I showed in class.

(g) (1 point)

Produce the plot of standardised residual Vs leverage score for `chem_pro.LM` like the one I showed in class.

(h) (1 point)

Produce a table of influence measures for `chem_pro.LM` like the one I showed in class.

Task 2 (6 points)

The data `USA_real_estate` is about the median price of houses sold in different areas of USA in 2006.

Variable	Description
`mppsf`	Median Price Per Square Foot
`ns`	Number Homes from which the Median Price is computed
`pnh`	Percentage of Homes sold that are build in 2005 or 2006
`pms`	Percentage of Mortgage Foreclosure Sales

Each data point is for one such area of USA in 2006.

(a) (1 point)

Check for the presence of heteroskedasticity in the model `usare.LM`.

```
usare.df = read.table(file = USA_real_estate.txt, sep = "", header = TRUE)
usare.LM = lm(mppsf~pnh+pms, data = usare.df)
```

(b) (1 point)

Estiamte the weights for using weighted least squares for the following linear model

$$\text{mppsf}_i = \beta_0 + \beta_1 \text{pnh}_i + \beta_2 \text{pms}_i + \sigma_i \varepsilon$$

(c) (1 point)

Construct the linear model using weighted least squares with your estimated weights, name it `usare.WLS`.

$$\text{mppsf}_i = \beta_0 + \beta_1 \text{pnh}_i + \beta_2 \text{pms}_i + \sigma_i \varepsilon$$

(d) (1 point)

Explain why `ns` might also be an appropriate estimate for the weights.

(e) (1 point)

Construct the linear model using weighted least squares with the weights based on `ns`, name it `usare.ns.WLS`.

$$\text{mppsf}_i = \beta_0 + \beta_1 \text{pnh}_i + \beta_2 \text{pms}_i + \sigma_i \varepsilon$$

(f) (1 point)

Compare `usare.WLS` with `usare.ns.WLS`. Which of the two models do you prefer? Explain your answer.

Task 3 (5 points)

The data `grossboxoffice` is about yearly gross box office receipts from moives screened in Australia.

(a) (1 point)

Load the data file `grossboxoffice.txt` into R, and construct the following model, name it as `gbo.LM`.

$$\text{GrossBoxOffice}_i = \beta_0 + \beta_1 \text{year}_i + \varepsilon$$

Comment on the validity of `gbo.LM`.

(b) (1 point)

Explore the possibility of using AR(1), AR(2), and AR(3).

(c) (1 point)

Obtain a final model for predicting `GrossBoxOffice` for `year=1975`, name it as `gbo.final.M`.

(d) (1 point)

Produce diagnostic plots to justify your choice of model.

(e) (1 point)

Describe any weakness in your `gbo.final.M`.

(f) (1 point)

Use your model `gbo.final.M` to identify any outliers.