UNIVERSITY OF TORONTO SCARBOROUGH
Department of Computer and Mathematical Sciences
STAC67H3 - Assignment - March 2018

Student 1

Last name: _____ First name: _____

Student number: _____

Student 2

Last name: _____ First name: _____

Student number: _____

Student 3

Last name: _____ First name: _____

Student number: _____

Instructions

- You may collaborate with **at most** two other students who are currently enrolled in STAC67H3. If you collaborate with other students, you must **submit only one assignment** with all students' names and number, ordered by last name. All the partners will receive the same mark. Collaboration involving more than 3 students **is not allowed**.
- For help with your assignment, you may consult only the instructor, your assignments partner(s) (if any), your textbook, your class notes, and the R help and documentation. You may not consult any other source and doing so represents an **academic offence**.
- **Print this document and complete all questions in pen**. Any questions completed in pencil will not be eligible to be remarked even if there was a marking error.
- Please **provide the answers in the corresponding space**. If you do not have enough space, please use the back of a nearby page and indicate clearly where to look.
- Please round the final results to the **nearest third decimal**.
- There are 4 exercises. Read carefully the instructions at the beginning of each exercise. Please check that you have all the consecutively numbered pages of this assignment. There are 11 pages.
- In order to receive full credit for a problem, you should show all of your work and explain your reasoning.
- The deadline to submit the assignment is **Monday April 2 at 1PM**. Drop your assignment no later than the deadline in the dropbox of the course (course code STAC67 is indicated on the dropbox) located to the right after the main entrance of CMS department (IC building, 4th floor). Note: the dropbox area is locked outside office hours and **no late assignment will be accepted**.

Question:	1	2	3	4	Total
Points:	18	13	14	35	80
Score:					

1. **Weighted least squares in simple linear regression**

When diagnostics indicate that some assumptions of the linear regression model are violated, remedial measures may need to be taken. The assumption of common variance of the error term plays a key role in least squares. An approach to handling heterogeneous variances, called heteroscedasticity, is the use of weighted least squares, which we study in this exercise.

Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where the errors ε_i are pairewise independent with mean 0, and the variance of ε_i is $c_i^2\sigma^2$ for some non-null coefficients c_i that are not all equal (we have $c_i \neq c_j$ for some i, j). Since coefficients c_i are not all equal, there is heteroscedasticity. Consider $w_i = 1/c_i$, $i = 1, \ldots, n$. The weighted least squares method is a generalization of the least squares that accounts for heteroscedasticity.

(a) (8 points) The *weighted least squares estimators* $\widehat{\beta}_{w0}$ and $\widehat{\beta}_{w1}$ of β_0 and β_1, respectively, minimize the weighted least squares criterion

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} w_i (Y_i - \beta_0 - \beta_1 X_{i1})^2.$$

Show that the weighted least squares estimators are

$$\widehat{\beta}_{w1} = \frac{\sum_{i=1}^n w_i X_i Y_i \sum_{i=1}^n w_i - \sum_{i=1}^n w_i X_i \sum_{i=1}^n w_i Y_i}{\sum_{i=1}^n w_i X_i^2 \sum_{i=1}^n w_i - (\sum_{i=1}^n w_i X_i)^2},$$

$$\widehat{\beta}_{w0} = \frac{\sum_{i=1}^n w_i Y_i}{\sum_{i=1}^n w_i} - \widehat{\beta}_{w1} \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i}.$$

Show all your work. Partial work will not receive full credit.

(b) (5 points) Are the weighted least squares estimators $\widehat{\beta}_{w0}$ and $\widehat{\beta}_{w1}$ unbiased estimators of β_0 and β_1, respectively? Justify and show all your work.

(c) (5 points) Show that the weighted least squares estimators are equal to the ordinary (unweighted) least squares estimators, i.e.

$$\widehat{\beta}_{w1} = \widehat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - \frac{1}{n}\sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i^2 - \frac{1}{n}(\sum_{i=1}^n X_i)^2},$$
$$\widehat{\beta}_{w0} = \widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X},$$

when the errors have common variance σ^2. Show all the details of your proof.

2. **Application of weighted least squares in simple linear regression**

Read pages 421–426 (Section 11.1 Unequal Error Variance Remedial Measures - Weighted Least Squares) of Kutner et al. (2004) **before** you solve this exercise.

Choose the correct answer for parts (a) to (c).

(a) (1 point) Weighted least squares is a procedure applied when
- ◯ The relation between the true mean of the response and the explanatory variables is not linear.
- ◯ The errors are not independent.
- ◯ The errors do not have common variance.
- ◯ The errors are not normally distributed.

(b) (1 point) Which of the following statements is correct?

(I) The weighted least squares is a generalization of the ordinary least squares where equal weights of 1 are replaced with unequal weights.

(II) When the variances of the errors are known, the weighted least squares estimators usually exhibit less variability than the ordinary least squares estimators.

- ◯ Only statement (I) is true.
- ◯ Only statement (II) is true.
- ◯ Both statements (I) and (II) are true.
- ◯ None of these statements is true.

(c) (1 point) In practice, the variances of the errors are often unknown. A method to estimate these variances is based on the idea that they vary in a regular fashion with the explanatory variables or with the mean response. Suppose we know that there is a relationship between the variances of the errors and one specific explanatory variable and we want to estimate this unknown relationship. Which of the following is the appropriate tool?
- ◯ A scatterplot of the residuals against the explanatory variable
- ◯ The ordinary least squares regression fit of the response variable on the explanatory variable
- ◯ The weighted least squares regression fit of the response variable on the explanatory variable
- ◯ The ordinary least squares regression fit of the square or absolute value of the residuals on the explanatory variable

We consider the cars dataset of Ezekiel (1930), which gives the speed in mph (variable speed) and the distance taken to stop in ft (variable dist) of 50 cars. Part of the dataset is shown below:

Car	1	2	3	4	$\cdots$	50
dist	2	10	4	22	$\cdots$	85
speed	4	4	7	7	$\cdots$	25

We consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where Y_i and X_i are the distance and speed of car i, the errors ε_i are pairewise independent with mean 0, and the variance of ε_i is $c_i^2\sigma^2$ for some non-null coefficients c_i. Using R, we followed the estimation process described at the top of page 426 of Kutner et al. (2004). The estimation process and R outputs are provided below.

Step 1 We fit a simple linear regression of distance on speed with ordinary least squares and obtain diagnostics plots. The summary of the fitted model is below and Figure 1 shows the diagnostics plots.

```
Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

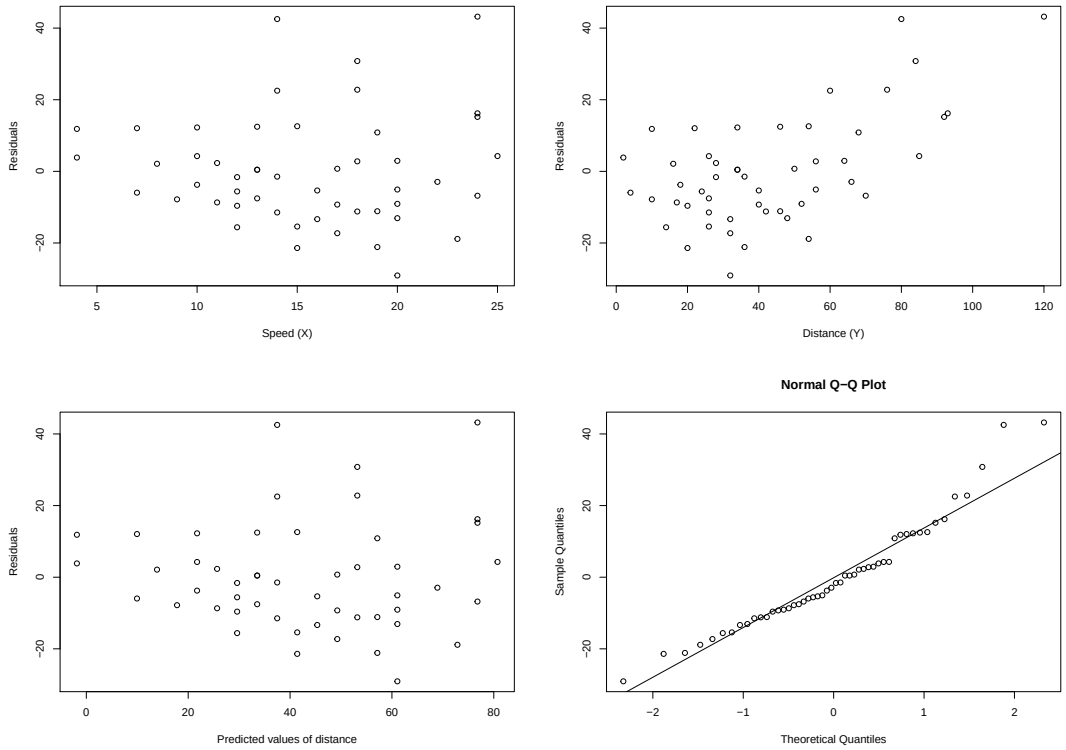

Figure 1: Plot of the residuals against speed (top left panel); plot of the residuals against dist (top right panel); plot of the residuals against the predicted values (bottom left panel); normal quantile plot of the residuals (bottom right panel)

Step 2 In order to estimate the functional relationship between the variance of the errors and the explanatory variable, we fit a simple linear regression of the absolute value of the residuals of the regression of Step 1 (abs(res.ols)) on speed with ordinary least squares. The summary of the fitted model is below.

```
Call:
lm(formula = abs(res.ols) ~ speed)

Residuals:
    Min      1Q  Median      3Q     Max
-12.349  -6.311  -1.075   3.905  31.680

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.4987     4.1453   0.844   0.4029
speed         0.5248     0.2549   2.059   0.0449 *
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1  1

Residual standard error: 9.433 on 48 degrees of freedom
Multiple R-squared:  0.08116,  Adjusted R-squared:  0.06202
F-statistic:  4.24 on 1 and 48 DF,  p-value: 0.04494
```

Step 3 We use the fitted values of the regression of Step 2 and obtain the weights w_i to use in weighted least squares.

Step 4 Using the weights w_i, we compute

$$\sum_{i=1}^{n} w_i Y_i = 14.774, \quad \sum_{i=1}^{n} w_i X_i = 5.721, \quad \sum_{i=1}^{n} w_i X_i Y_i = 232.738,$$

$$\sum_{i=1}^{n} w_i X_i^2 = 84.814, \quad \sum_{i=1}^{n} w_i = 0.46,$$

where X_i, Y_i, and w_i are the speed, the stopping distance, and the weight of car i, respectively.

Use the information and R outputs provided above to answer questions (d) to (g).

(d) (2 points) Explain why we should use weighted least squares regression rather than ordinary (unweighted) least squares regression for this data set. Be concise and specific and explain on which element(s) of the diagnostics plots/R output you base your answer.

(e) (2 points) Give the value of the residuals of the simple linear model fitted in Step 1 for the first 2 cars.

(f) (3 points) Compute the value of the estimated weights w_i that are used in weighted least squares for the first 2 cars.

(g) (3 points) Compute the weighted least squares estimates of β_0 and β_1.

3. **Multiple linear regression**

Do a person's brain size and body size predict his/her intelligence? Interested in answering this question, a group of researchers record the performance IQ scores (Y), the brain size (X_1) based on the count obtained from MRI scans, the height (X_2) in inches, and the weight (X_3) in pounds for a random sample of students.

The researchers consider the following multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where the errors ε_i are pairewise independent with mean 0 and common variance σ^2, Y_i, X_{i1}, X_{i2}, X_{i3} are the observed values of Y, X_1, X_2, X_3 taken on student i, and β_0, β_1, β_2 are parameters. This model can be rewritten in the matrix form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

The researchers obtain the following R output. Use this output to answer the questions below.

```
> iqdata = read.table("iqsize.txt", header = TRUE, sep = "")
> X = as.matrix(cbind(1, iqdata[,c("Brain","Height","Weight")]))
> cor(iqdata)
                PIQ      Brain       Height       Weight
PIQ      1.000000000 0.3778155 -0.09315559 0.002512154
Brain    0.377815463 1.0000000  0.58836684 0.513486971
Height -0.093155590 0.5883668  1.00000000 0.699614004
Weight   0.002512154 0.5134870  0.69961400 1.000000000
> t(X) %*% X
              1     Brain    Height    Weight
1         38.00   3445.68    2600.0    5740.0
Brain   3445.68 314387.94 236387.9 523715.8
Height  2600.00 236387.94 178484.9 395164.1
Weight  5740.00 523715.83 395164.1 887438.0
> round(solve( t(X) %*% X ),4)
              1   Brain  Height  Weight
1       10.1204 -0.0207 -0.1505  0.0138
Brain   -0.0207  0.0008 -0.0007  0.0000
Height  -0.1505 -0.0007  0.0039 -0.0004
Weight   0.0138  0.0000 -0.0004  0.0001
>
> fit = lm(PIQ ~ Brain + Height + Weight, data = iqdata)
> summary(fit)

Call:
lm(formula = PIQ ~ Brain + Height + Weight, data = iqdata)

Residuals:
   Min     1Q Median     3Q    Max
-32.74 -12.09  -3.84  14.17  51.69

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.114e+02  6.297e+01   1.768 0.085979 .
Brain        2.060e+00  5.634e-01   3.657 0.000856 ***
Height      -2.732e+00  1.229e+00  -2.222 0.033034 *
Weight       5.599e-04  1.971e-01   0.003 0.997750
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1          1

Residual standard error: 19.79 on 34 degrees of freedom
Multiple R-squared:  0.2949,   Adjusted R-squared:  0.2327
F-statistic: 4.741 on 3 and 34 DF,  p-value: 0.007215
```

(a) (4 points) Should one or several variables be dropped from the model? If you answer "yes", state which variable you would drop first and justify your choice using the summary of the fitted model.

(b) (2 points) Does the correlation matrix of the four variables confirm your answer to the previous part? Justify.

(c) (8 points) Use a statistical test (level of significance 5%) to test whether Height and Weight can both be dropped from the model **simultaneously**, while Brain is retained. State the null and alternative hypotheses, give the value of the test statistic, give the critical value, and conclude **in the context of the data being analyzed**. To obtain the value of the test statistic, you may use the software of your choice (R, matlab, excel, other). If you decide to use a statistical software, clearly indicate it and explain which formula you used. If you decide not to use a software, do not round the numbers.

4. (35 points) The fourth exercise is to be completed with R Studio and R Markdown package. To do so, you will need to install R, R Studio, and R Markdown package on your computer.

   Once this is done, download file "STAC67w18_assign_ex4.Rmd" from Blackboard and open it with R Studio. Carefully read and follow the instructions on this file. A pdf version of this file is also available on Blackboard.

References

Ezekiel, M. (1930). *Methods of Correlation Analysis.* Wiley.

Kutner, M. H., Nachtstein, C. J., and Neter, J. (2004). *Applied Linear Regression Models.* McGraw-Hill, 4th edition.