NATIONAL UNIVERSITY OF SINGAPORE

ST3131 REGRESSION ANALYSIS

(Semester 1 : AY 2020/2021)

Time Allowed: 2 Hours

INSTRUCTIONS TO STUDENTS

1. Please write your student number on your answer script only. **Do not write your name**.
2. Write your answer on A4 size paper. Do not write your answer using any electronic device.
3. This exam paper contains **FOUR (4)** problems and comprises **EIGHT (8)** printed pages (including the cover page).
4. This is an OPEN BOOK exam. You may refer to your lecture notes, tutorials, textbooks or any notes that you have made, but you are not allowed to perform any online search.
5. You may use a non-programmable scientific calculator but you are not allowed to use R or any other software for computations.
6. Students are required to answer **ALL** questions. Total mark is 60.
7. During the exam, you are not allowed to communicate with any person other than the invigilators via Zoom Chat. To prevent communication with others,
    - You are not allowed to use your mobile phone or tablet for viewing the question paper or your notes. However, you can still use these devices to join the Zoom session and as the webcam.
    - You may view the exam question paper and your notes on a PC or laptop, but you are not allowed to use the keyboard or tap on the screen. This is to prevent messaging with others. You should use only your mouse for navigation.
8. If you wish to communicate with the invigilator, click the Raise Hand button on Zoom. The invigilator will then give you permission to use the keyboard, mobile phone, or tablet to type a message via the Zoom Chat function.
9. At the end of the exam, take photos of your answers and label them by your student number before submitting them to the LumiNUS Submission folder.

1. (*20 points*) For each of the following questions: choose one of the given choices that best answers the question.

(a) Let y be the response variable and x_1 be a regressor in a simple linear model which gives $SS_{res} = 7058.5$. Adding another two regressors x_2 and x_3 into the model for y, the new fitted model will have

1. SS_{res} increasing to 7100.
2. SS_{res} increasing but not enough information to determine its value.
3. SS_{res} decreasing.
4. SS_{res} not changing.

(b) Let y be the response variable and $\hat{y}$ be the fitted value from a linear regression model. Consider the correlation coefficient $\text{Cor}(y, \hat{y})$. Which of the following results is incorrect:

1. $\text{Cor}(y, \hat{y}) = 0.589$
2. $\text{Cor}(y, \hat{y}) = -0.589$
3. $\text{Cor}(y, \hat{y}) = 0.9$
4. $\text{Cor}(y, \hat{y}) = 1$

(c) Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the least squares estimators of the parameters of the simple linear regression model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, ..., n$. We express $\hat{\beta}_0$ and $\hat{\beta}_1$ as $\hat{\beta}_0 = \sum_{i=1}^n b_i y_i$ and $\hat{\beta}_1 = \sum_{i=1}^n a_i y_i$. Which of the following statements is incorrect?

1. $\sum_{i=1}^n a_i^2 = 1$.
2. $\sum_{i=1}^n b_i = 1$.
3. $\sum_{i=1}^n b_i x_i = 0$.
4. $\sum_{i=1}^n a_i = 0$.

(d) Which of the following is not the case when an intercept model should be used?

1. $H_0 : \beta_0 = 0$ is not rejected in a formal hypothesis test.
2. The data are assumed to come from a single random sample.
3. None of the predictor variables is of interest.
4. $H_0 : \beta_1 = 0$ is not rejected in a formal hypothesis test.

(e) Calculating variance inflation factors (VIFs) can help to detect the presence of multicollinearity in a data set. Suppose a data set has four quantitative regressors, and the VIFs are obtained as follows: 0.8, 1.6, 4 and 100. Which value is calculated wrongly?

1. 0.8
2. 1.6
3. 4
4. 100

(f) Let Cor$(y, \hat{y})$ be the correlation between the vector of the response values and that of the fitted values based on a simple linear regression fit of y on x. If Cor$(y, \hat{y}) > 0$, which of the following must be false?

1. $\text{Cor}(x, y) = 0$
2. $\text{Cor}(x, y) > 0$
3. $\text{Cor}(x, y) < 0$
4. $\text{Cor}(x, y) = \text{Cor}(y, \hat{y})$

(g) Let y and x_1 have a strong nonlinear relationship, y and x_2 have a strong linear relationship, and x_1 and x_2 be negatively linearly correlated. A multiple linear regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ is fitted to the data. What linear regression assumption(s) is/are violated for this data set?

1. Linearity assumption only.
2. Constant variance assumption only.
3. Uncorrelated regressors assumption only.
4. Linearity and constant variance assumptions.
5. Linearity and uncorrelated regressors assumptions.

(h) If any of the off-diagonal values $|r_{ij}|$, $i \neq j$, of $\boldsymbol{X}'\boldsymbol{X}$ (in correlation form) is larger than the value below then severe multicollinearity is present.

1. 0.5
2. 0.7
3. 0.9

4. There is no threshold for $|r_{ij}|,\ i \neq j$ to ensure the presence of multicollinearity.

(i) If $\boldsymbol{X}'\boldsymbol{X}$ in correlation form does not have any large $|r_{ij}|,\ i \neq j$, then

1. there is surely no multicollinearity in the model.
2. model still might have mulicollinearity.
3. there is mild multicollinearity in the model.
4. all regressors are uncorrelated.

(j) Which of the following simple linear regression models has residuals that may not sum up to zero?

1. The slope model
2. The general simple linear regression model
3. The intercept model
4. All of the models listed in the other three options

2. (*6 points*) Provide the answer for each of the questions below.

(a) When severe multicollinearity is present, explain why estimation of parameters is poor. Describe how ridge regression deals with the presence of multicollinearity in the data.

(b) You are given a data set where the variance of the response variable is not stable. Describe the purpose of using weighted least squares (WLS) method instead of ordinary least squares (OLS) method to estimate the coefficients in a linear regression model for this data. What are the advantage(s) and disadvantage(s) of the WLS method?

(c) Suppose a linear regression model with $k = 2$ regressors has been fitted to $n = 43$ observations and $R^2 = 0.48$ is obtained. Test for the significance of the model (report the test statistic, p-value and conclusion) at significance level 0.01.

3. (*28 points*) Consider a data set of 173 random female horseshoes crabs where the **weight** (gram), the carapace **width** (cm) and the **spine** condition of each crab were recorded. The spine condition is divided into three groups: both good (1); one worn or broken (2) and both worn or broken (3). We are interested in explaining the weight of the crabs based on their carapace width and spine condition.

(a) A histogram of the variable weight and a scatter plot of weight vs width are given in Figure 1. Comment on the suitability of fitting a linear regression model for the variable weight with regressor width.

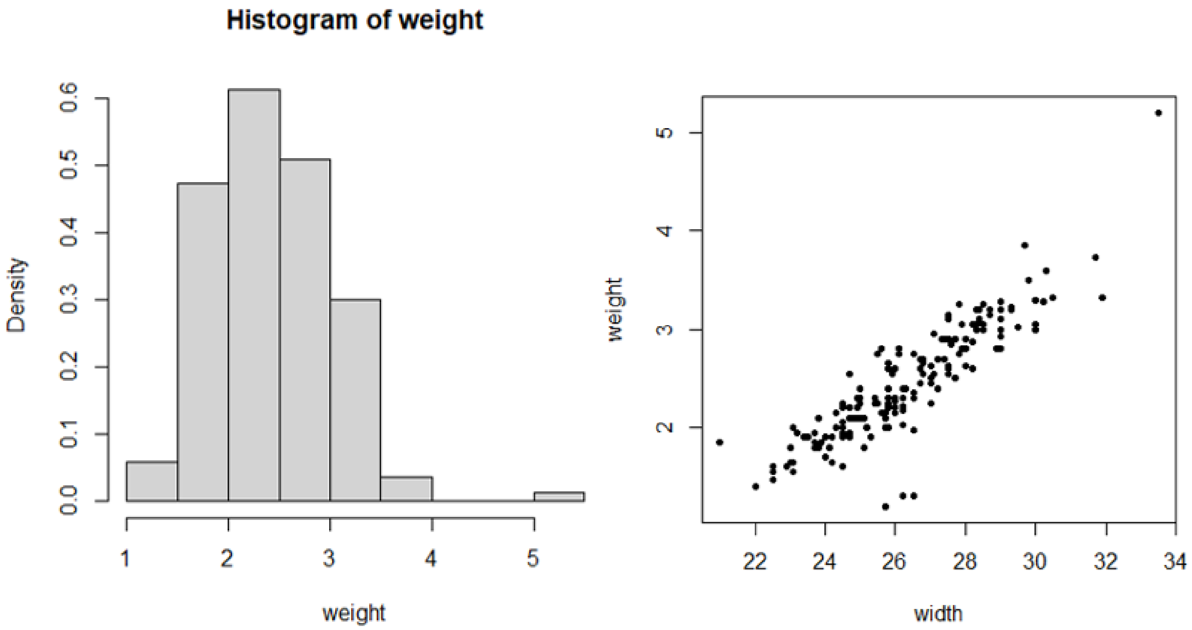
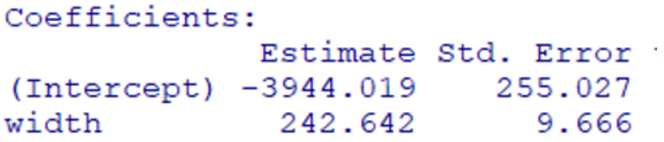


Figure 1: Plots before fitting a model

PART I A simple linear regression model is fitted (called Model 1), where the variable width is the regressor. The coefficients of Model 1 is given in Figure 2. The residual plots and some further outputs of Model 1 are given in Figure 3 and Figure 4.

(b) Conduct a detailed check of the adequacy of Model 1 (check model assumptions using residual plots, and check for outliers and influential points).

(c) Test the significance of Model 1 at significance level 0.01.



```
Coefficients:
             Estimate Std. Error
(Intercept) -3944.019    255.027
width         242.642      9.666
```

Figure 2: Coefficients of Model 1

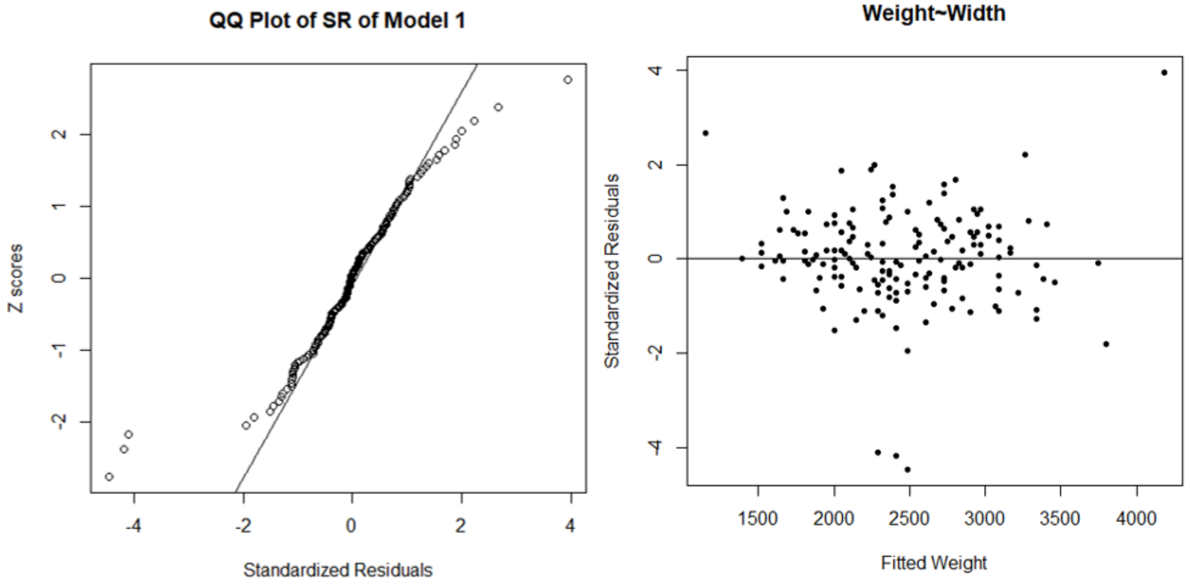
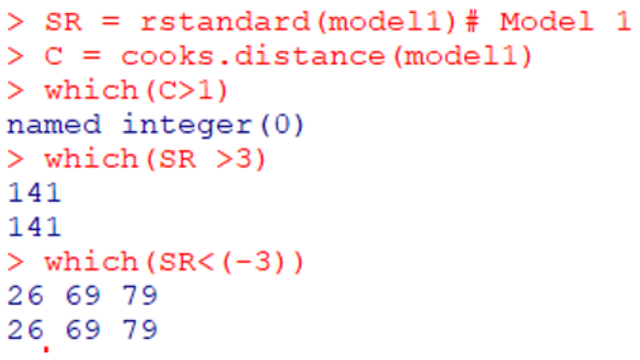


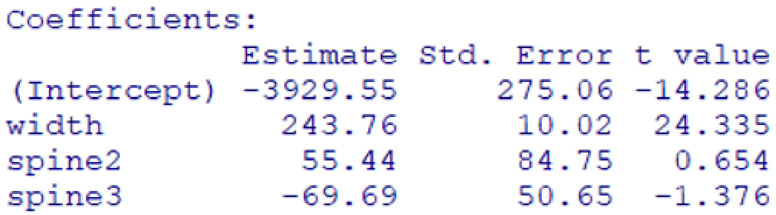
Figure 3: Residual plots of Model 1



```
> SR = rstandard(model1)# Model 1
> C = cooks.distance(model1)
> which(C>1)
named integer(0)
> which(SR >3)
141
141
> which(SR<(-3))
26 69 79
26 69 79
```

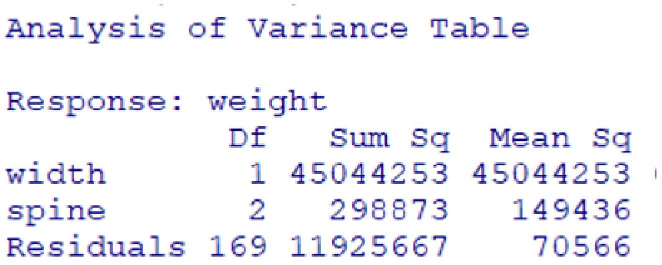
Figure 4: Output of Model 1

PART II A second model (called Model 2) which has variables width and spine as regressors is fitted. Its output is given in Figure 5 and Figure 6.



```
Coefficients:
             Estimate Std. Error t value
(Intercept) -3929.55     275.06 -14.286
width         243.76      10.02  24.335
spine2         55.44      84.75   0.654
spine3        -69.69      50.65  -1.376
```

Figure 5: Coefficients of Model 2



```
Analysis of Variance Table

Response: weight
           Df   Sum Sq  Mean Sq
width       1 45044253 45044253
spine       2   298873   149436
Residuals 169 11925667    70566
```

Figure 6: Anova table of Model 2

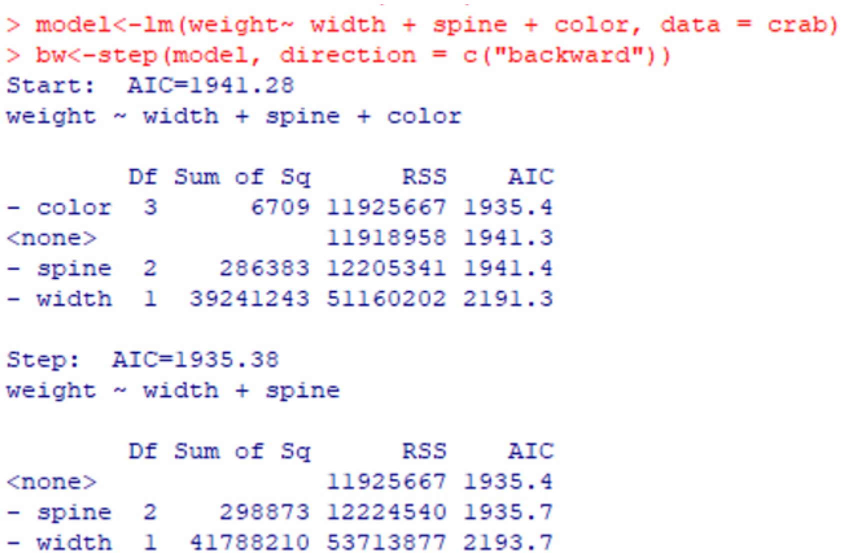
(d) Write down the fitted equation of Model 2.

(e) Perform a test to test the significance of the variable spine in Model 2, at significance level 0.05. (Report the hypotheses, test statistic value and its null distribution, p-value of the test and your conclusion.)

(f) Calculate the adjusted R^2 of Model 1 and the adjusted R^2 of Model 2. Which model do you prefer and why?

PART III There is another regressor in the data called **color** which describes the color of the crab. A full model including all three regressors (width, spine and color) is fitted. The backward variable selection method is applied to this full model. The R output is given in Figure 7.

(g) Report the regressors and the AIC of the model that is chosen by this method (called Model 3). What is the value of R^2 of Model 3?



```
> model<-lm(weight~ width + spine + color, data = crab)
> bw<-step(model, direction = c("backward"))
Start:  AIC=1941.28
weight ~ width + spine + color

        Df Sum of Sq      RSS    AIC
- color  3      6709 11925667 1935.4
<none>               11918958 1941.3
- spine  2    286383 12205341 1941.4
- width  1  39241243 51160202 2191.3

Step:  AIC=1935.38
weight ~ width + spine

        Df Sum of Sq      RSS    AIC
<none>               11925667 1935.4
- spine  2    298873 12224540 1935.7
- width  1  41788210 53713877 2193.7
```

Figure 7: Backward Variable Selection

4. (*6 points*) A data set consists of two variables x and y derived from a simple random sample. The observations of x and y are $x_1, ..., x_n$, and $y_1, ..., y_n$, respectively. Let y be the response variable. A simple linear regression model (using least squares method) is fitted to these data, which yields a coefficient of determination R^2. Let $\hat{y}_1, ..., \hat{y}_n$ denote the fitted values and $e_i = y_i - \hat{y}_i, \quad i = 1, ..., n$, denote the residuals. Using the fact that $\sum_{i=1}^n x_i\, e_i = 0$, prove that

(a) $\sum_{i=1}^n \hat{y}_i\, e_i = 0$.

(b) $\text{Cov}(y, \hat{y}) = \text{Var}(\hat{y})$.

(c) $R^2 = [\text{Cor}(y, \hat{y})]^2$.

– SOME GIVEN QUANTILES –

$F_{2,40}(0.01)$	$F_{2,41}(0.01)$	$F_{2,40}(0.005)$	$F_{2,169}(0.1)$	$F_{2,169}(0.025)$	$t_{171}(0.005)$	$t_{171}(0.01)$
5.178	5.163	6.066	2.334	3.77	2.605	2.348

– END OF PAPER –