STA302 / STA1001 Midterm Test

Dept of Statistical Sciences, University of Toronto

24 October 2017, LEC0101: Professor Mark Ebden

First Name: _____ Surname: _____ Student number: _____

Test location: OI G162 [X] UC 266 [ ] UC 273 [ ]

Your course: STA302 (undergraduate) [ ] STA1001 (graduate) [ ]

Instructions:

- Time allowed: 105 minutes
- Answer all questions, in pen or pencil
- Aids allowed: You are allowed a nonprogrammable calculator

Question	Value
1	7
2	10
3	20
4	14
5	6
Total	57

This test should have eight pages including this page

Notation: This test uses the regular notation from class and from Simon Sheather's textbook:

$$Y_i = \beta_0 + \beta_1 x_i + e_i$$

You may assume in this test that a "small" dataset (for the purposes of the aid sheet formulae) has about 200 or fewer data points.

---

1. Warmup questions (7 marks)

[a] (1 mark) In R, for vector `y` what is the command to take its transpose to the power of 0.7?

[b] (2 marks) Write a single R command to evaluate at $x = 8$ the cumulative distribution function for a $\chi^2_5(x)$, saving the result as `p`.

[c] (4 marks) Show that $\sum_{i=1}^{n} \hat{e}_i \hat{y}_i = 0$. You may or may not choose to use $\sum_{i=1}^{n} \hat{e}_i x_i = 0$ in your proof.

---

2. Multiple choice and True/false ($5 \times 2 = 10$ marks)

I. Suppose you have numerical variables $\{Z, W\}$ in your global environment in R. Which of the following R commands will correctly fit the SLR model $E(Z) = \beta_0 + \beta_1 W$?

A. `fit(Z ~ W)`

B. `lm(Z ~ W)`

C. `fit(W ~ Z)`

D. `lm(W ~ Z)`

II. True / false: A Type II error is when we incorrectly reject the null hypothesis.

True False

III. Which of the following is *not* one of the Gauss-Markov conditions?

A. The e_i's are independent of each other

B. $\mathbb{E}(e_i) = 0$

C. $\text{var}(e_i) = \text{constant for all } \ i$

D. None of the above

IV. Multiple linear regression is an example of a remedial measure in which case?

A. A point exists with DFBETA > 1

B. You suspect an important predictor variable Z is missing from your analysis.

C. A log transformation of the data has failed to work as intended.

D. None of the above is likely to be an appropriate case.

V. You perform a two-sample t-test and get a p-value of 0.012. You then use the same data to fit a linear regression model with dummy variable $X \in \{0, 1\}$ as we did in class. Using confidence level $\alpha = 0.05$:

A. If $H_0 : \beta_0 = 0$, you would necessarily reject the null hypothesis.

B. If $H_0 : \beta_1 = 0$, you would necessarily reject the null hypothesis.

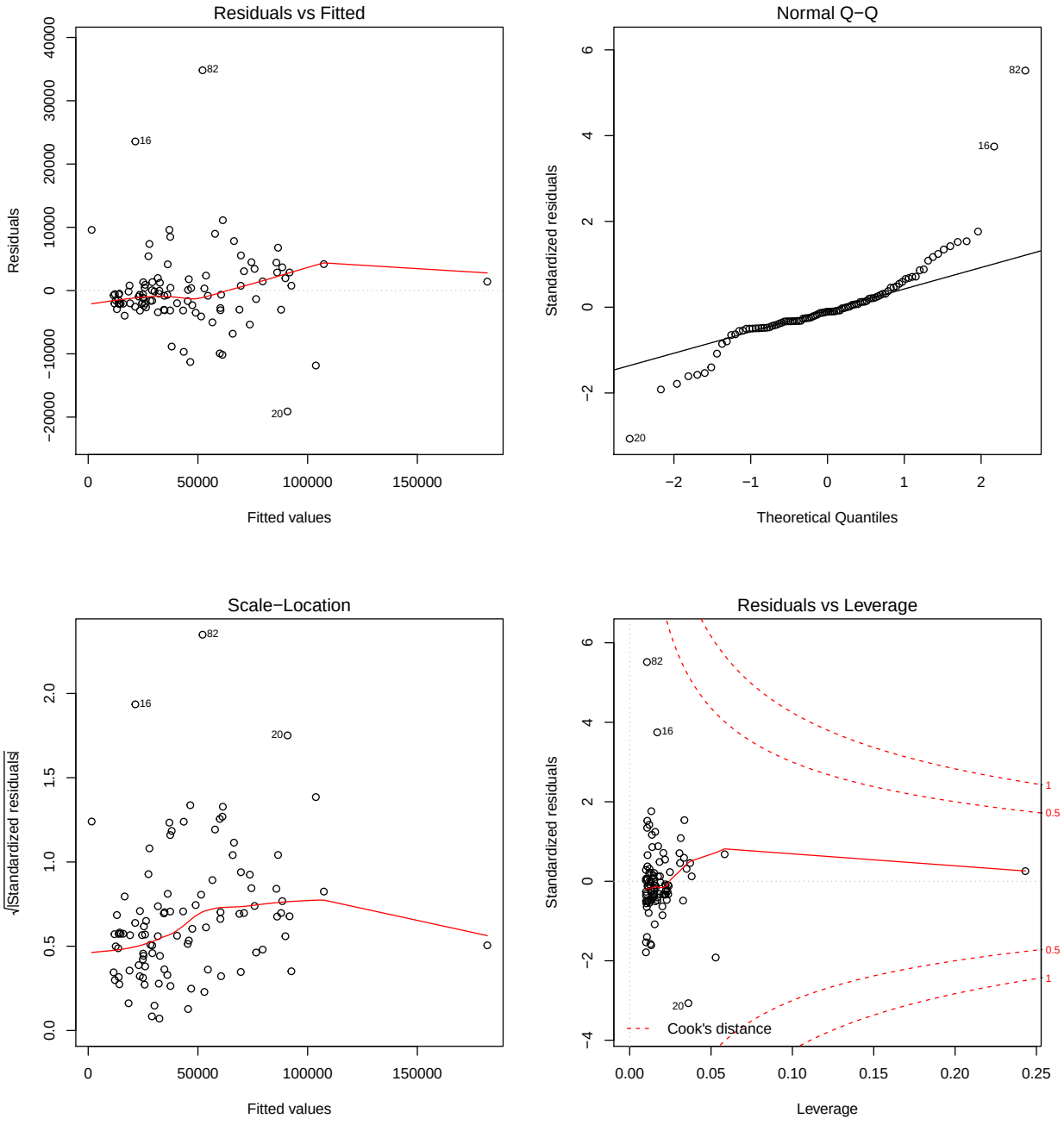
C. Neither of the above is certain.

3. R code (20 marks) In the Toronto Public Library system, there are $n = 100$ libraries, each of which provides a shelf for users to pick up books they have reserved (placed on Hold).

For the ith library, let x_i represent the number of Holds available for pickup in 2014, and let y_i represent the number of Holds available for pickup in 2015.

The R^2 between x and y is approximately 0.95. Consider the following code and output.

```
x<-X$Holds2014; y<-X$Holds2015
par(mfrow=c(2,2))
plot(lm(y~x))
```



[a] (6 marks) Explaining your answer, is there evidence for the data containing:

[i] Outliers

[ii] Leverage points

[iii] Influential points

[b] (6 marks) In the first plot, Residuals vs Fitted, suppose the rightmost data point is removed and the linear regression is rerun. Which of the following quantities will change by a substantial ($> 1\%$) amount? In such cases, indicate the direction of the change.

[i] The slope, $\hat{\beta}_1$

[ii] The mean square error, S^2

[iii] The coefficient of determination, R^2

[c] (3 marks) Assess the assumption of constant variance in the model error term, e_i.

[d] (3 marks) Assess the assumption of normal errors.

[e] (2 marks) Identify briefly two remedial measures you might take, explaining why.

1.

2.

4. R code (14 marks) Consider the following code and output.

```
x <- seq(1,5); y <- c(1,-1,3,5,9)
summary(lm(y~x))
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    1    2    3    4    5
##  2.0  [A] -0.4 -0.6  1.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -3.20       [C]  -1.608   0.2062
## x               2.20       [B]   3.667   0.0351 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.897 on [C] degrees of freedom
## Multiple R-squared:  0.8176, Adjusted R-squared:  0.7568
## F-statistic:  [D]  on 1 and [C] DF,  p-value: [E]
```

I. (5×2 marks) Find the five missing values, [A] through [E]. You do not need to show your work.

[A] = _____ [B] = _____

[C] = _____ [D] = _____

[E] = _____

II. (2 marks) What is the fitted regression line? You do not need to show your work.

III. (2 marks) In the line beginning `(Intercept)`, what is the null hypothesis?

5. Theory (6 marks) Derive the formula for $\text{cov}(\hat{\beta}_0, \hat{\beta}_1)$ given on the aid sheet. You may use any other formulae except those whose derivations require knowing $\text{cov}(\hat{\beta}_0, \hat{\beta}_1)$. As usual, you can assume that a value of $X = x^*$ is fixed/given, implicitly.

This aid sheet will be provided to you along with your test, on the day. (You can't bring your own copy to the test.)

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}, \qquad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$\mathrm{var}(\hat{\beta}_0) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right], \qquad \mathrm{var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}, \qquad \mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2\bar{x}}{S_{xx}}$$

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{SST}} = \underbrace{b_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2}_{\text{SSReg}} + \underbrace{\sum_{i=1}^{n}\hat{e}_i^2}_{\text{RSS}}, \qquad \text{SSReg} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2, \qquad F_{\text{obs}} = \frac{\text{MSReg}}{\text{MSE}}$$

$$\mathrm{var}(\hat{y}^*) = \sigma^2\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right], \qquad \mathrm{var}(Y^* - \hat{y}^*) = \sigma^2\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]$$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \qquad h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}}$$

$$\text{DFBETA}_{ik} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\text{s.e. of } \hat{\beta}_{k(i)}}, \qquad \text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\text{s.e. of } \hat{y}_{i(i)}}, \qquad D_i = \frac{\sum_{j=1}(\hat{y}_{j(i)} - \hat{y}_j)^2}{2S^2} = \frac{r_i^2 h_{ii}}{2(1 - h_{ii})}$$

$$\text{where } S^2 = \text{MSE} = \tfrac{\text{RSS}}{n-2} \text{ and } r_i = \tfrac{\hat{e}_i}{S\sqrt{1-h_{ii}}}.$$

Criteria for ordinary data points on small datasets: $r_i < 2$, $h_{ii} < 4/n$, DFBETA < 1, DFFITS < 1, $D_i < 1$