STA302 / STA1001 Midterm Test

Dept of Statistical Sciences, University of Toronto

26 October 2017, LEC5101: Professor Mark Ebden

First Name: _____ Surname: _____ Student number: _____

Test location: ES 1050 [X] BR 200 [ ]

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
- On multiple choice questions, there's no penalty for wrong answers

Question	Value	Mark
6	8	
7	10	
8	13	
9	18	
10	8	
Total	57	

This test should have eight pages including this page. There are no questions 1 to 5.

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

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

---

6. Warmups (8 marks)

[a] (2 marks) Your classmate performs linear regression and tells you that $R^2 = 1$. Explain in one sentence what that implies about their data set.

[b] (2 marks) Write a single R command that takes a random sample from a t distribution with 4 degrees of freedom, saving the (scalar) result as `x`.

[c] (4 marks) Show that $\hat{\beta}_1$ can be expressed as a linear combination of the y_i's. Namely, for a function $g(x_i)$,

$$\hat{\beta}_1 = \sum_{i=1}^{n} g(x_i)\, y_i$$

7. Multiple-choice ($5 \times 2 = 10$ marks)

Circle one (1) letter per question. You don't need to show your work.

I. For β_0, you calculate a 99% confidence interval of (2.3, 2.9).

A. If your experiment is repeated, there's a 99% chance that the next $\hat{\beta}_0$ will fall between 2.3 and 2.9.
B. Before the experiment was conducted, there was a 99% chance that $\hat{\beta}_0$ would fall within (2.3, 2.9).
C. With 99% probability, β_0 falls within (2.3, 2.9).
D. None of the above.

II. The coefficient of determination, R^2, is equal to:

A. RSS/SST
B. RSS/SSreg
C. SSReg/SST
D. None of the above

III. The most interesting thing about Anscombe's quartet is that it demonstrates how Pearson's r can be:

A. Similar across datasets
B. Negative or positive
C. An indicator of linearity

Note: All three of the above are potential properties of r, but please read the question carefully.

IV. Cochran's theorem tells us why

A. The number of degrees of freedom of SST is $n - 2$
B. SSReg/σ^2 is a χ^2 variable
C. $\mathbb{E}(\text{SSReg}) = \mathbb{E}\left(\hat{\beta}_1^2\right) S_{xx}$
D. None of the above

V. You notice three points with a value of DFFITS > 1. Possible remedial measures may include:

A. Removing the points
B. Introducing a nonlinearity to the model
C. Applying a transformation
D. Any of the above

8. Working with R output: Part I (7 marks)

Consider the following:

```
print(c(qt(.95,6),qt(.975,6),qt(.95,7),qt(.975,7),qt(.95,8),qt(.975,8)))
```

```
## [1] 1.943180 2.446912 1.894579 2.364624 1.859548 2.306004
```

```
x<-c(0,2,3,4,6,7,8,10)
summary(x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    2.75    5.00    5.00    7.25   10.00
```

```
y<-c(10,2.5,4.5,5,6,8,10,10); summary(lm(y~x))
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3269 -1.6362 -0.5865  1.2404  4.9551
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.0449     1.8795   2.684   0.0363 *
## x             0.3910     0.3188   1.226   0.2660
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.816 on 6 degrees of freedom
## Multiple R-squared:  0.2004, Adjusted R-squared:  0.06718
## F-statistic: 1.504 on 1 and 6 DF,  p-value: 0.266
```

[a] (4 marks) Find a 95% confidence interval for the value of β_1.

[b] (3 marks) Suppose you calculate the 95% confidence interval for the value of Y at $x = 5$. A classmate calculates the corresponding 95% *prediction* interval, and they report that they got almost the same width as you, matching to three significant figures. Could they be correct? Explain why or why not. (You're allowed to include the actual intervals in your solution, only if you wish.)

8. Working with R output: Part II (6 marks)

Consider the following R command run on two 5-vectors `y1` and `y2`:

```
t.test(y2,y1,var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  y2 and y1
## t = 0.42164, df = 8, p-value = 0.6844
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.575335  5.175335
## sample estimates:
## mean of x mean of y
##      21.6      20.8
```

[c] (2 marks) What is the main conclusion you can draw from this output?

[d] (2 marks) Multiple choice, no need to show work.

Your classmate performs dummy-variable regression as we did in lecture. Namely,

```
y <- c(y1,y2); x <- c(rep(0,5),rep(1,5))
summary(lm(y~x))
```

In the R output (not shown), for $H_0 : \beta_1 = 0$, the p-value will be:

[i] 0.6844

[ii] Probably below 0.6844

[iii] More information is needed

[e] (2 marks) Multiple choice, no need to show work. This is a challenging question.

Another classmate performs linear regression of `y2` versus `y1`: `summary(lm(y2~y1))`. The R output (not shown) will show an F-statistic of:

[i] $(0.42164)^2$

[ii] $\sqrt{0.42164}$

[iii] 0.42164

[iv] More information is needed

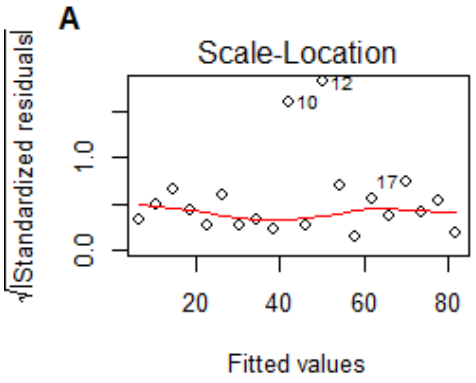
9. Diagnostics ($6 \times 3 = 18$ marks): Consider the following ten issues, I through X:

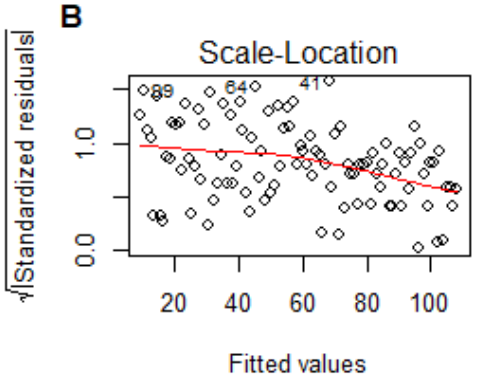
I. Heavy tails II. Left skew III. Right skew IV. Light tails

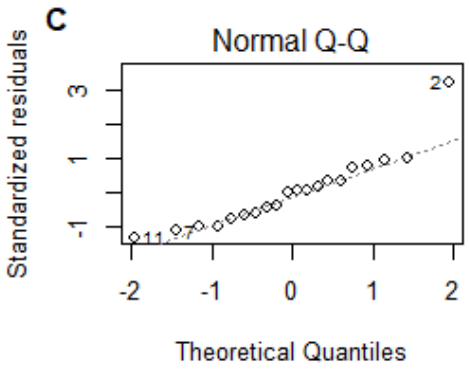
V. Non-constant variance VI. Outlier(s) VII. Leverage point(s) VIII. Influential point(s)

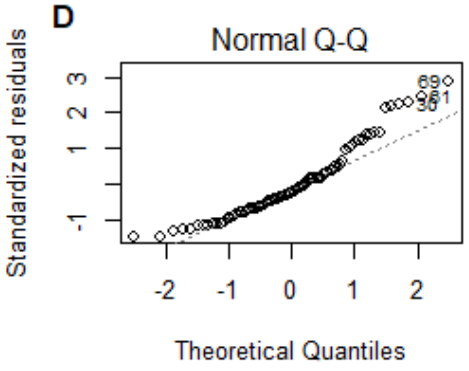
IX. Non-independent errors X. None of the above

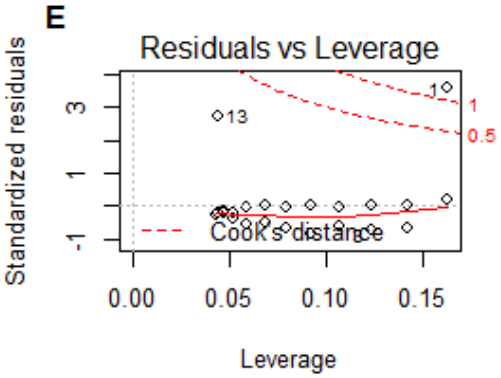
The following graphs were produced from **six different data sets** (of different sizes n). In each case, list all of the above issues which apply. You may repeat or ignore choices as needed. You may assume the datasets are "small" for the purposes of the aid sheet formulae. You don't need to show your work.

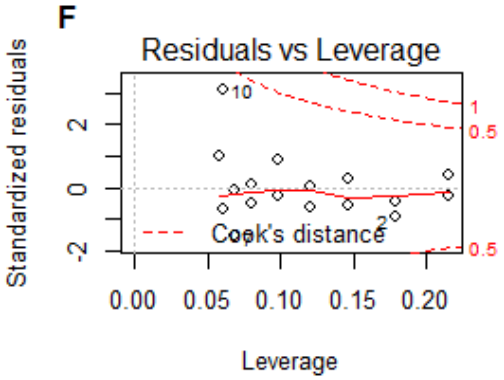












A. (n=20) _____ B. (n=100) _____

C. (n=20) _____ D. (n=80) _____

E. (n=23) _____ F. (n=17) _____

10. Theory (8 marks)

[a] (2 marks) In a model under the Gauss-Markov conditions, will the sample mean of the residuals always equal the true mean of the error term? Explain.

[b] (6 marks) Show that, for $i \neq j$, we have $\text{cov}(\hat{y}_i, \hat{e}_j) = 0$. As usual, you can assume that a value of $X = x^*$ is fixed/given, implicitly. Hint: You may find it useful to recall that

$$\hat{y}_i = \sum_{k=1}^{n} h_{ik}\, y_k$$

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, $D_i < 1$