York University

MATH 4939 – Midterm Test

Professor Georges Monette

February 14, 2018 – 10:30 am to 11:20 am (50 minutes)

WARNING

DO NOT OPEN THIS BOOKLET UNTIL YOU ARE INSTRUCTED TO DO SO

Student number: _____

Family name: (in BLOCK letters)_____

Given name: (in BLOCK letters)_____

Signature_____

Information:

This exam has 9 questions. Make sure you complete every question.

Be sure to read questions closely. Some may ask for multiple pieces of information. Make sure to respond completely. If you need more space to answer, write "**OVER**" and continue the answer on the back of the page.

The point value is shown at the end of each question. The sum of the points is 105. The exam will be graded out of 100 so that you can potentially earn 5 bonus points.

Aids allowed: Non-programmable calculator, ruler, pencils, pens, erasers.

WARNING

DO NOT OPEN THIS BOOKLET UNTIL YOU ARE INSTRUCTED TO DO SO

1. Suppose you have data on two variables, X and Y, in each of J groups. Let $\hat{\beta}_j$ represent the vector of coefficients from the least-squares regression ot Y on X within the jth group, $j = 1, \ldots, J$. **Prove** that the inverse-variance weighted combination of the within-group estimated coefficients is the same as the vector of coefficients for the least-squares regression using the pooled data. Explain why (a sketch will suffice, no formal argument is needed) the pooled estimate of the slope estimates a combination of the within-group regression slope and the between-group regression slope. *(15 points)*

2. Describe the difference in R between a 'generic function' and a method. *(10 points)*

3. Suppose we run this command:

```
a <- matrix(1:8, rnow = 2)
```

Then which of the following R expressions result in the following output?

```
[1]  8
```

(Write 'Y' for yes, 'N' for no, and 'D' or blank for 'do not know'. +1 for a correct answer, -1 for a wrong answer and 0 for 'D')

_____ `a(8)`

_____ `a[8]`

_____ `a(2,4)`

_____ `a[4,2]`

_____ `a(2,4)` *(5 points)*

4. When interpreting a study that purports to show a relationship between two variables, what do you think are the three most important questions that you should ask? Discuss as succinctly as you can the consequences of the answers to those questions. *(10 points)*

5. Discuss whether you should use the aggregate (marginal) or the segregated (conditional) data to attempt to determine the true effect in the following situation: There are two doctors in a small town. Each has performed 100 surgeries in his career, which are of two types: one very difficult surgery and one very easy surgery. The first doctor performs the easy surgery much more often than the difficult surgery and

the second doctor performs the difficult surgery more often than the easy surgery. You need surgery, but you do not know whether your case is easy or difficult. Should you consult the success rate of each doctor over all cases, or should you consult their success rates for the easy and difficult cases separately, to maximize the chance of a successful surgery? *(15 points)*

6. Suppose you wish to estimate the relationship between income, Y, and education, X. Because of heteroscedasticity and curvature in the relationship you choose to fit a linear model using the log of Y:

```
fit <- lm( log(Y) ~ X, data)
```

Write the R code you would use to plot the estimated increase in income associated with an extra year of education as a function of years of education. It is not necessary to include error bars in the plot. *(15 points)*

7. Which of the following R expressions result in the following output?

```
[1]  8
```

(Write 'Y' for yes, 'N' for no, and 'D' or blank for 'do not know'. +1 for a correct answer, -1 for a wrong answer and 0 for 'D')

_____ `"+"(5,3)`

_____ `"/"(16,2)`

_____ `2^3`

_____ `4 + 4`

_____ `"^"(3,2)` *(5 points)*

8. Discuss whether you should use the aggregate (marginal) or the segregated (conditional) data to attempt to determine the true effect in the following situation: In a study of a group of male 50 to 55-year-old long-time smokers, researchers compared a group of heavy smokers with matched group (same age range, sex and similar socioeconomic and environmental backgrounds) of light smokers. It was found that lung function was worse in the group of heavy smokers than in the group of light smokers. The researchers also measured the amount of tar deposit in the lungs of the subjects and classified subjects as having heavy or light tar deposits. Would you get a better indication of the effect of smoking by comparing the aggregated data for the two groups or by comparing the tar level specific data? *(15 points)*

9. Discuss situations when a) it would be important to include a variable that is not significant and b) it would be important to exclude a variable that is highly significant. Give examples of each situation. *(15 points)*