Solutions

# York University

MATH 4939 – Midterm

Instructor: Georges Monette

February 14, 2020 – 9:30 am to 10:20 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:**

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 marks for each questions are shown at the end of the question. The sum of the marks 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. In R, the data frame `mtcars` has 32 rows and 11 variables of which `cyl` is a variable recording the number of cylinders in each type of car. Fix each of the following common data frame subsetting errors in R:

```
mtcars[mtcars$cyl = 4, ]

mtcars[-1:4, ]

mtcars[mtcars$cyl <= 5]

mtcars[mtcars$cyl == 4 | 6, ]
```

*(20 points)*

if selecting columns

```
mtcars[, names(mtcars) %in% c('cyl','disp')]
mtcars[, grep('cyl|disp', names(mtcars))]
```

2. The following is some output from a linear regression of life expectancy in a number of countries regressed on HE (health expenditures per capita per year in dollars US), smoke
(cigarettes per capita per year), hiv and special, that are two indicator variable to identify anomalous countries.

```
fit.hiv2 <- lm( LifeExp ~ log(HE) * (smoke + I(smoke^2)) + hiv+special, dd ,
                na.action = na.exclude)
summary(fit.hiv2)
```

```
|
|   Call:
|   lm(formula = LifeExp ~ log(HE) * (smoke + I(smoke^2)) + hiv +
|       special, data = dd, na.action = na.exclude)
|
|   Residuals:
|       Min      1Q  Median      3Q     Max
|   -9.0373 -2.3005  0.2043  2.0760  9.7344
|
|   Coefficients:
|                        Estimate Std. Error t value Pr(>|t|)
|   (Intercept)          3.283e+01  2.674e+00  12.280  < 2e-16 ***
|   log(HE)              6.091e+00  5.024e-01  12.124  < 2e-16 ***
|   smoke                3.642e-02  7.520e-03   4.844 3.31e-06 ***
|   I(smoke^2)          -1.518e-05  3.946e-06  -3.846 0.000181 ***
|   hiv                 -7.351e-01  7.593e-02  -9.681  < 2e-16 ***
|   special             -1.822e+01  2.137e+00  -8.526 2.11e-14 ***
|   log(HE):smoke       -4.878e-03  1.155e-03  -4.223 4.30e-05 ***
|   log(HE):I(smoke^2)   2.007e-06  5.726e-07   3.504 0.000614 ***
```

Work-Sellke adjustment

Construct a hypothesis matrix to estimate the predicted difference in life expectancy associated with an increase of 1,000 cigarettes per capita per year for a country with a level of health expenditures equal to 2,000 and cigarette consumption equal to 1,000.

*(10 points)*

Since the not linear in 'smoke' we need to take a difference

L for HE = 2000 & Smoke = 1000 :

$[1 \quad \log 2000 \quad 1000 \quad 1000^2 \quad hiv \quad special \quad \log 2000 \cdot 1000 \quad \log 2000 \cdot 1000^2]$

L for HE = 2000, Smoke = 2000 :

$[1 \quad \log 2000 \quad 2000 \quad 2000^2 \quad hiv \quad special \quad \log 2000 \cdot 2000 \quad \log 2000 \cdot 2000^2]$

Diff: $[0 \quad 0 \quad 1000 \quad 3{,}000{,}000 \quad 0 \quad 0 \quad \log 2000 * 1000 \quad \log 2000 \cdot 3000000]$

3. (continued from the previous question) Construct a hypothesis matrix to estimate the predicted difference in life expectancy associated with an increase of 1,000 dollars in health expenditures per capita per year for a country with a level of health expenditures equal to 2,000 and cigarette consumption equal to 1,000. *(10 points)*

L for HE = 2000 smoke = 1000: Same as above

L for HE = 3000 smoke = 1000

$[\ 1 \quad \log 3000 \quad 1000 \quad 1000^2$ hiv specific $\log 3000 \cdot 1000 \quad \log 3000 \cdot 1000^2\ ]$

Difference:

$[\ 0 \quad \log\frac{3}{2} \quad 0 \quad 0 \quad 0 \quad 0 \quad \log\frac{3}{2} \times 1000 \quad \log\frac{3}{2} \times 1000000\ ]$

4. Here are some fictitious data on the rate of complications for appendectomies performed at University Hospital, a large urban teaching and research hospital, and in County Hospital, a small-town hospital: at University Hospital there were 800 cases with 100 (12.5%) resulting in complication and at County Hospital there were 200 cases resulting in 5 (10%) complications. The p-value for a test of the hypothesis that there is no difference in the rate at the two hospitals is 0.0037. Suppose that appendectomies can be classified as high risk or low risk and that the high risk cases tend to be directed disproportionately to University Hospital instead of County Hospital. Construct two hypothetical tables, one for each level of risk, and draw a Paik diagram that shows how it is possible for both high- and low-risk patients to have a lower probability of complications at University Hospital than at County Hospital, although, overall the probability of complications is higher at University Hospital than at County Hospital. *(10 points)*

Given

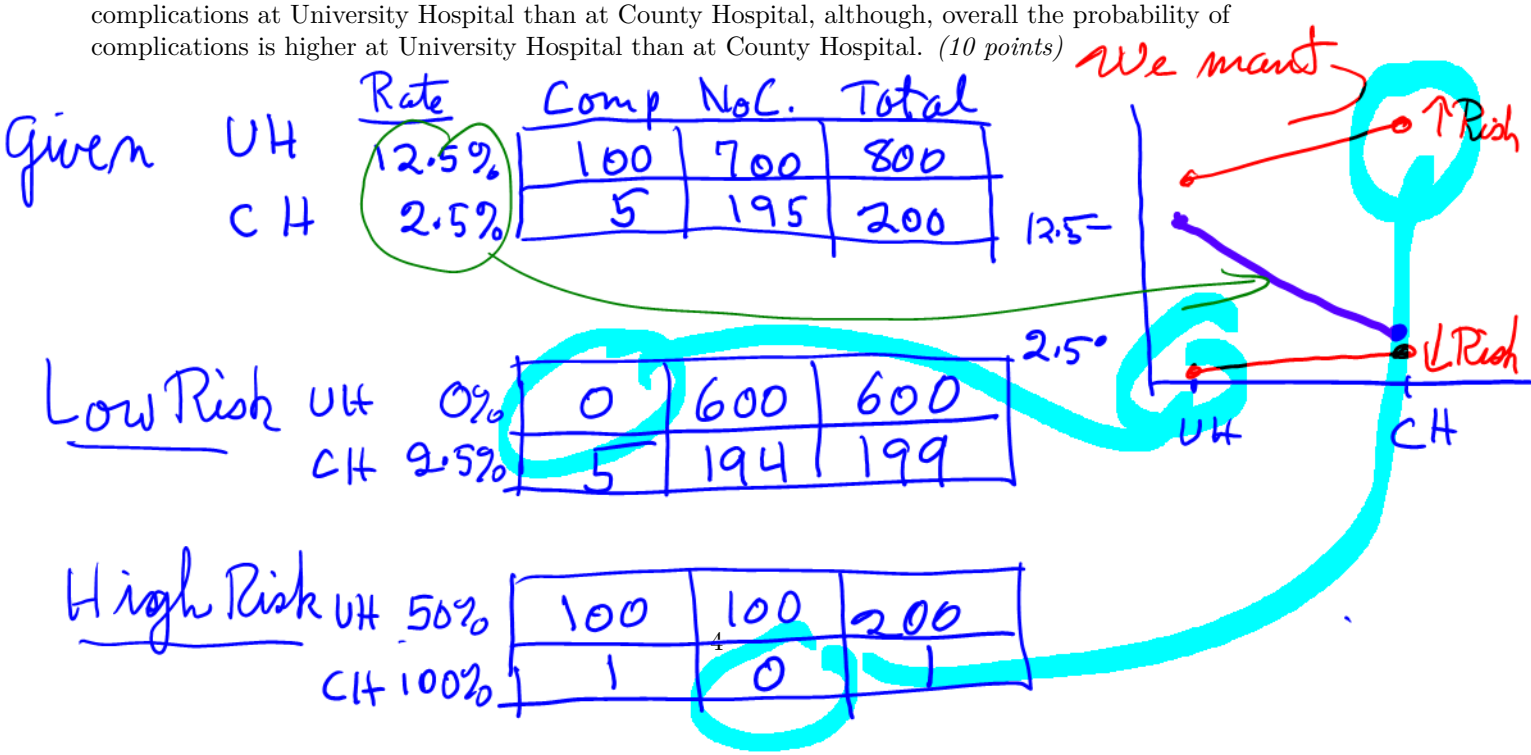
| | Rate | Comp | NoC. | Total |
|---|---|---|---|---|
| UH | 12.5% | 100 | 700 | 800 |
| CH | 2.5% | 5 | 195 | 200 |

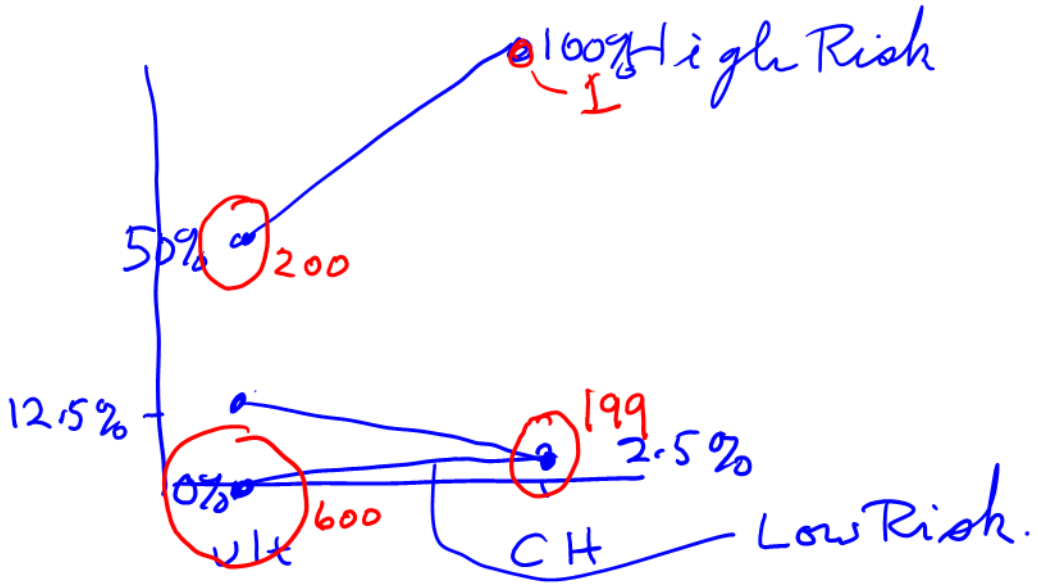
Low Risk

| | | | | |
|---|---|---|---|---|
| UH | 0% | 0 | 600 | 600 |
| CH | 2.5% | 5 | 194 | 199 |

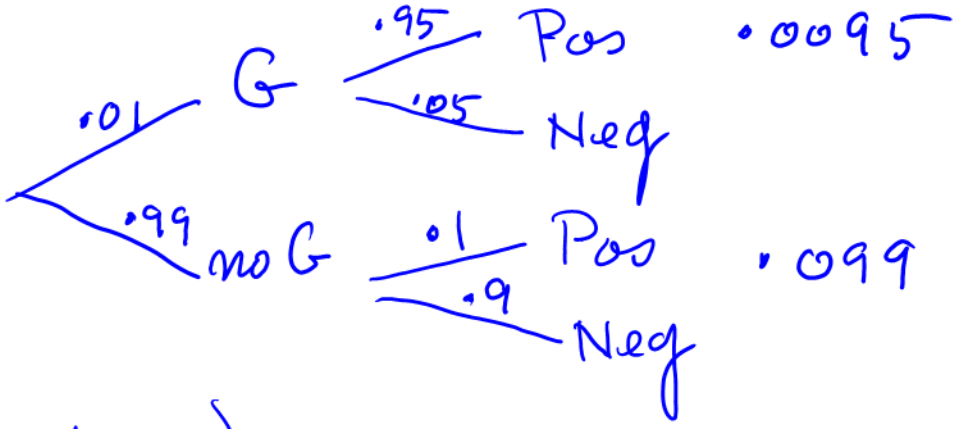
High Risk

| | | | | |
|---|---|---|---|---|
| UH | 50% | 100 | 100 | 200 |
| CH | 100% | 1 | 0 | 1 |




100% High Risk
1
50%
200
12.5%
199
2.5%
0%
600
VLT
CH
Low Risk.

5. Suppose a test for glaucoma has a sensitivity of .95 and a specificity of .90. You receive the test as a routine test on a regular visit to your optometrist. The prevalence of glaucoma in your age, ethnic and gender group among people who have not been previously diagnosed is 1 per 100. The test, alas, is positive. Use a natural frequency table to find the probability that you have glaucoma given the positive test result. *(10 points)*




$$\Pr(G \mid Pos) = \frac{.0095}{.0095 + .099} = 0.088$$

6. Suppose you were to read about a study based on a random survey of Ontario medical records that shows that smokers have twice as high a risk of kidney disease as non-smokers. Is it reasonable to conclude that smoking causes a higher risk of kidney disease? Why or why not? *(10 points)*

No because

a) Treatment clearly not randomized since it's a random survey of records so we can't rely on random assignment to infer a causal effect of smoking.

b) The risk comparison compares all smokers with all non-smokers so there is no control for

possible confounding factors, which would need to be controlled for to justify the inference of a causal relationship.

7. When performing a regression, 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)*

If the regression is performed for causal inference with observational data, it is important to

a) block confounding back-door paths by controlling for some confounding factors along each path even if the variable is not significant

b) Omit mediating factors even if - particularly if - a factor is significant.

Example: If ↑ stress causes ↑ coffee consumption and an ↑ in heart damage then the confounding effect of stress would need be controlled to estimate the causal effect of coffee.

If T coffee consumption causes heart damage by increasing some hormone, say adrenalin, whose higher levels in turn cause heart damage, then it is important to omit adrenalin levels in a regression to estimate the causal effect of coffee.

8. Let

```
d1 <- data.frame(id = c('a', 'a', 'b', 'c'), grade = c(1, 2, 1, 3))
d2 <- data.frame(id = c('a', 'c', 'c', 'd'), year = c(3, 1, 3, 4))
```

Describe the differences between the outputs of the following commands:

a. `merge(d1, d2)`

b. `merge(d1, d2, all.x = TRUE)` *(10 points)*

Since 'id' is the only variable in both d1 and d2, it is the key variable that links the two data frames for merging. The command in (a) take every combination for levels of "id" that occur in both d1 & d2

The result of (a) is:

| id | grade | year |
|---|---|---|
| a | 1 | 3 |
| a | 2 | 3 |
| c | 3 | 1 |
| c | 3 | 3 |

The result of (b) takes every row of d1 whether matched in d2 or not. On addition to the rows above the result also includes

| id | grade | year |
|---|---|---|
| b | 1 | NA |

9. Write a function in R that takes a character string and collapses multiple adjoining blanks between words to a single blank and remove leading and trailing blanks. *(10 points)*

```
collapse_blanks <-
  function(x) {
    x <- gsub('__*', '_', x)
    x <- gsub('^_*', '', x)
    x <- gsub('_*$', '', x)
    x
  }
```

Note: _ represents a blank

**END**