York University

MATH 4939 – Final Exam

Professor Georges Monette

April 17, 2018 – 9 am to 11 am (120 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 12 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 155. The exam will be graded out of 140 so that you can potentially earn 15 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. The output below uses the schizophrenia data in which patients were observed at years 1, 2, 3, 4, 5, and 6 taking one of three drugs: Atypical, Clozapine, Typical each year.

```
fitgl <- lme( gen ~ drug + cvar(drug,id) + year, dd, random = ~ 1 | id)
summary( fitgl )
```

```
|   Linear mixed-effects model fit by REML
|    Data: dd
|          AIC      BIC   logLik
|     2148.02 2177.964 -1066.01
|
|   Random effects:
|    Formula: ~1 | id
|           (Intercept) Residual
|   StdDev:    5.803759 6.097606
|
|   Fixed effects: gen ~ drug + cvar(drug, id) + year
|                                Value Std.Error  DF   t-value p-value
|   (Intercept)               33.64712  2.732709 262 12.312733  0.0000
|   drugClozapine             -1.55705  1.533425 262 -1.015409  0.3108
|   drugTypical                2.11299  1.244117 262  1.698387  0.0906
|   cvar(drug, id)Clozapine    8.58797  3.688862  50  2.328082  0.0240
|   cvar(drug, id)Typical     -1.29235  3.662873  50 -0.352825  0.7257
|   year                      -1.02063  0.249215 262 -4.095356  0.0001
```

Sketch the predicted response as a function of time (with years ranging from 0 to 6), for a patient who took Atypical drugs 1/3 of the time and Clozapine 2/3 of the time if they were on Atypical drugs and if they were on Clozapine. On the graph, indicate the numerical value of the intercept and of the slope of the line. Repeat for the same patient taking Clozapine. *(15 points)*

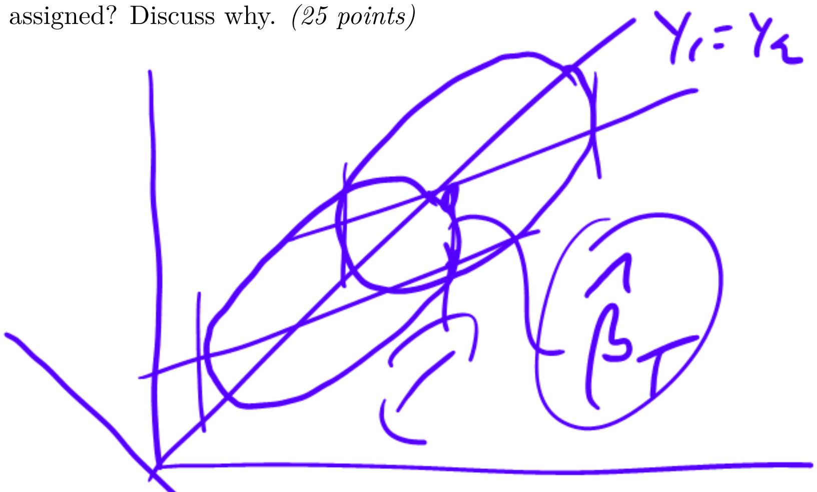
$$Y = \boxed{X\beta} + \varepsilon$$

fit <- lm(. . .)

$\hat{Y}$ predict(fit)

2. 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 treatments used on kidney stones: Treatment A and Treatment B. Doctors are more likely to use Treatment A on large (and therefore, more severe) stones and more likely to use Treatment B on small stones. Should a patient who doesn't know the size of his or her stone examine the general population data, or the stone size-specific data when determining which treatment will be more effective? Why? Draw a Paik diagram if it helps to make your point more clearly. *(15 points)*

3. Discuss how Lord's Paradox illustrates the usefulness of the gain score to test the difference between two treatments in which the response of interest has been measured before and after the application of the treatments but in which treatment assignment has not been randomized. Compare (a) regression of the gain score on the treatment variable with (b) regression of the post-test on the treatment variable using the pre-test as a covariate. Which of the two methods would be better if treatments had been randomly assigned? Discuss why. *(25 points)*



Y_1 & Y_2 is symm in each treatment

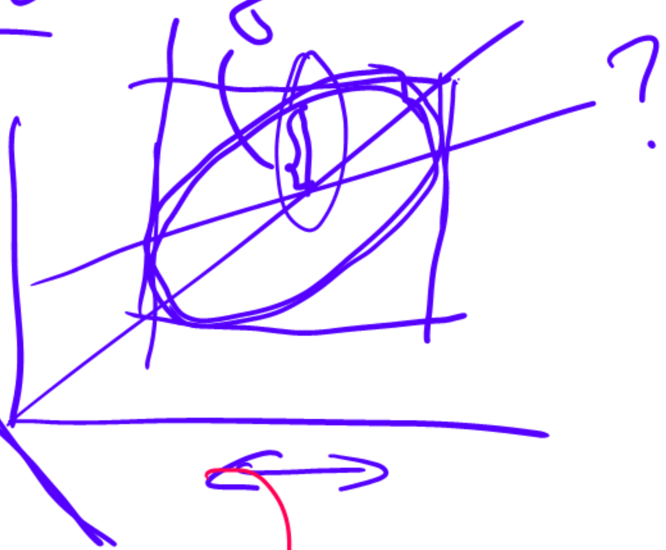
$E(G) = E(Y_1 - Y_2) = 0$

R A after Y_1

Est of TE

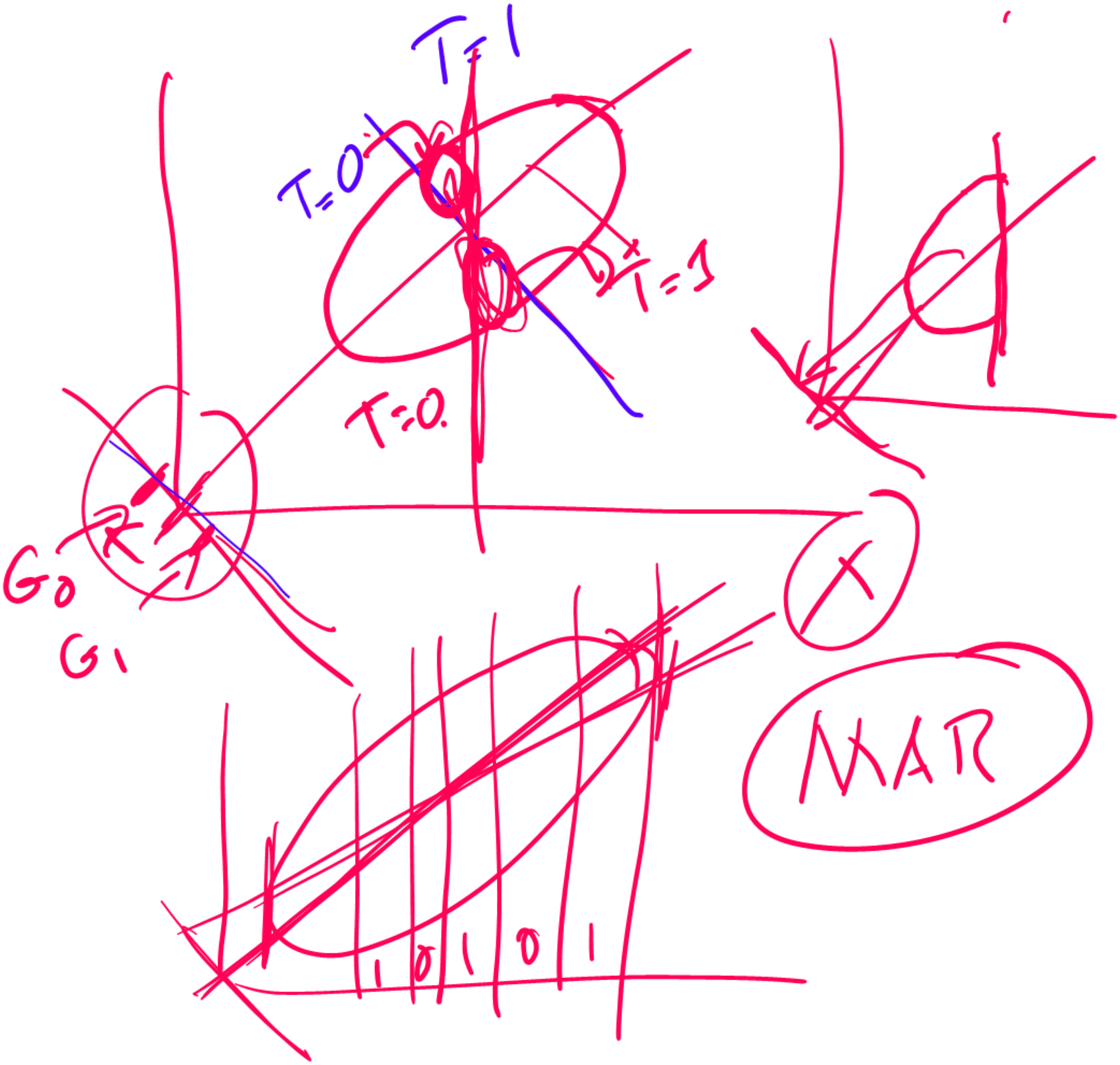
$Var(\hat{\beta}_T)$ smaller than $Var(G)$

$Y \sim T + Y_1$



Using Covariate is more efficient

$$SE(\hat{\beta}_T) = \boxed{\frac{\sigma}{\sqrt{n}} \frac{1}{\sigma_{T|X}}}$$


T=1
T=0
T=1
T=0.
G0
G1
1
0
1
0
1
MAR

4. Let `mat` be a matrix of integers in R. Write a function to find how many rows have exactly two instances of the number 7. *(5 points)*

5. Consider a mixed model of the form `lme( Y ~ X, data, random = ~ 1 + X | id)` in which there are two observations per cluster and the predictor, `X`, has the same two values, 0 and 1, in each cluster. Determine whether the variance parametrization of the model is identifiable. *(20 points)*

$$Z = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \end{bmatrix}$$

$$Z_i = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$$

$$Z_i G Z_i' = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} g_{00} & g_{01} \\ g_{10} & g_{11} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} g_{00} & g_{01} \\ g_{00}+g_{10} & g_{01}+g_{11} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

$$= \begin{bmatrix} g_{00} & g_{01}+g_{00} \\ g_{01}+g_{00} & g_{00}+2g_{10}+g_{11} \end{bmatrix}$$

$\underline{G}$ $\underline{R}$

$$\begin{bmatrix} \sigma_Y^2 & \sigma_{12} \\ \sigma_{12} & \sigma_{Y_2}^2 \end{bmatrix} = \begin{bmatrix} & '' & \end{bmatrix} + \underline{\sigma^2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Solving for 4 parameters with 3 equations.

∴ not identifiable.

6. Let Y and X be a numerical variables and let G be a factor. Consider the following models. All but one of these models will produce the same regression coefficient for X or Xr but they will produce different standard errors. Identify the model that produces a different coefficient. Rank the others where you can according to the standard error of the estimated coefficient for X stating which would be equal if any (assume a very large n and ignore the effect of slight differences in degrees of freedom for the error term). Explain your reasoning briefly.

   (A) `Y ~ X + G`
   (B) `Y ~ X`
   (C) `Yr ~ Xr` where Yr is the residual of Y regressed on G and Xr is the same for X
   (D) `Y ~ Xr`
   (E) `Y ~ X + Xh` where Xh is the predictor of X in the regression of X on G
   (F) `Y ~ X + Xh + Zg` where Zg is a 'G-level' numerical variable, i.e. it has the same value for all observations within any value of G.
(25 points)

Thm: Projection Theorem

$Y = Xb + e \qquad e \perp \mathcal{L}(X)$

& X is of full row rank

Then $b = \hat{\beta} = (X'X)^{-1}X'Y$

e is resid of LS fit

$e'e = |e|^2 =$ Resid SS. $= SSE$

Theorem: AVP = Frisch-Waugh Lovell

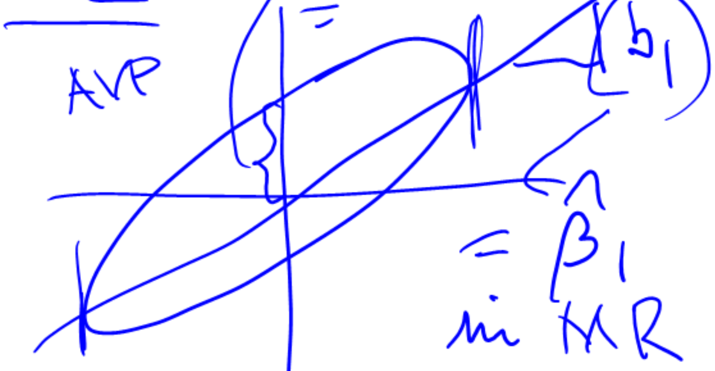
mult Y on X_1 & X_2

AVP

Resid of Y on X_2

Resid SS in AVP

= " " " MultReg



$Y \longrightarrow$ resid of Y on X_2 ?

$X_1 \longrightarrow$ resid of X_1 on X_2

$$E = Y - \hat{Y} = Y - HY \qquad H = X(X'X)^{-1}X'$$

$$= \underbrace{(I - H)}_{Q \sim \text{orthog proj.}} Y \qquad \left[\underline{Q^2 = Q} \quad \underline{Q = Q'} \right]$$

Wrt X_2

$$Q_2 = I - X_2(X_2'X_2)^{-1}X_2'$$

$$Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + e \qquad e \perp \underline{\mathcal{L}(X_1, X_2)}$$

$$\underbrace{Q_2 Y}_{\text{AVP vert}} = \underbrace{Q_2 X_1}_{\text{AVP hori. axis}}\hat{\beta}_1 + \underbrace{\boxed{Q_2 X_2 \hat{\beta}_2}}_{\overset{\parallel}{0}} + \underbrace{Q_2 e}_{e}$$

$$\boxed{\underbrace{Q_2 Y}_{\text{vert}} = \underbrace{Q_2 X_1}_{\text{horiz.}}\hat{\beta}_1 + e}$$

? Is $e \perp \mathcal{L}(Q_2 X_1)$?

$$\underline{e' Q_2 X_1} = e' X_1 = 0$$

$\boxed{\text{QED}}$

Linear Propensity Score

instead of all of X_2

Reg Y on X_1 & $\hat{X}_1|X_2$

$\hat{X}_1|X_2 \qquad \hat{X}_1 = X_2 c_{21} + d$

$\hat{X}_1|X_2 = (H_2 X_1) \qquad H_2 = X_2(X_2'X_2)^{-1}X_2'$

("H_2")

True model $\quad Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + e$

new $\quad Y = X_1 c_1 + \underbrace{H_2X_1}\, c_2 + f \quad$ Propensity model

AVP & PM

$$Q_3 = I - H_2X_1(X_1'\underset{H_2}{\boxed{H_2'H_2}}X_1)^{-1}X_1'\underset{H_2}{H_2'}$$

$$= I - H_2X_1(X_1'H_2X_1)^{-1}X_1'H_2$$

$$Q_3Y = Q_3X_1\hat{\beta}_1 + \boxed{Q_3X_2}\hat{\beta}_2 + Q_3e$$

vert axis

$$Q_3X_1 = (I - H_2X_1(\cancel{\cdots})^{-1}\cancel{X_1'H_2X_1})$$

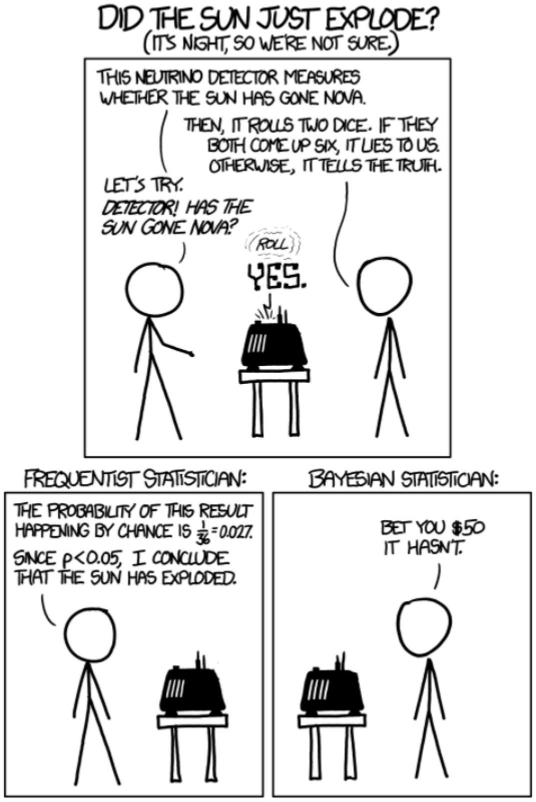
$$= X_1 - H_2X_1 = Q_2X_1$$

$$Q_3X_2 = (I - H_2X_1(\cdots)^{-1}X_1'H_2)X_2$$

$$= \circ \circ \circ$$

$$Q_3X_2\hat{\beta}_2 = \underset{\sim}{d}$$

7. The following XKCD cartoon shows two statisticians interpreting the same data: one who uses a frequentist approach unquestioningly and one who uses a Bayesian approach.


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY. DETECTOR! HAS THE SUN GONE NOVA?
(ROLL)
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36}=0.027$. SINCE $p<0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

Make some reasonable assumptions, stating them explicitly, and calculate a reasonable value for the Bayesian statistician's posterior probability that the sun has exploded. Discuss why there is a difference between the 'p-value' of 0.027 and the Bayesian posterior probability. Under what circumstances would you expect a p-value to be close to a posterior probability? *(15 points)*

8. Write a function in R that takes a character string and collapses multiple adjoining blanks to a single blank. *(5 points)*

9. In R, If `x` is a matrix, what does `x[] <- 0` do? How is it different from `x <- 0`? *(5 points)*

10. 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, ]
    ```

    (10 points)

11. Write a function in R that removes from a data frame every variable whose name starts with the letter ‘X’ and ends in a number. *(5 points)*

12. A survey of Canadian families yielded average ‘equity’ (i.e. total owned in real estate, bonds, stocks, etc. minus total owed) of $48,000. Aggregate government data of the total equity in the Canadian population shows that this figure must be much larger, in fact more than three times as large. Does this show that respondents must tend to dramatically underreport their equity or is there a probable explanation that is consistent with honest reporting by respondents? *(10 points)*

END