MATH 4939 Sample Final Exam
Duration: 120 minutes

Question 1: (30 marks in 6 parts)

The following analysis uses the 'long' version or the familiar 'TBI' data set, with the following subset of the variables:

```
car::some( subset(dl, group == 'control'))
```

```
      HPC_L_TOT  id      Age   group years_pi
c01.2      3273 c01 35.08830 control        0
c02.4      3086 c02 35.79740 control        0
c11.2      4555 c11 46.95414 control        0
c12.2      3448 c12 18.46680 control        0
c24.2      4338 c24 26.61191 control        0
c27.2      3377 c27 32.33402 control        0
c40.2      4652 c40 30.25325 control        0
c44.2      4631 c44 42.46680 control        0
c46.2      4286 c46 49.34428 control        0
c51.2      4537 c51 20.27105 control        0
```

```
car::some( subset(dl, group == 'patient'))
```

```
      HPC_L_TOT  id      Age   group  years_pi
322.4      3129 322 45.12252 patient 2.3080082
338.2      4435 338 57.94114 patient 0.4052019
338.4      4151 338 61.11431 patient 3.5783710
349.4      3754 349 53.58248 patient 2.6748802
389.3      3851 389 49.45654 patient 0.9993155
393.2      3519 393 21.93566 patient 0.4736482
399.3      3260 399 20.10951 patient 0.9007529
438.2      4697 438 56.76112 patient 0.4298426
448.2      3842 448 54.83094 patient 0.3668720
448.3      3776 448 55.61396 patient 1.1498973
```

The data consists of measurements of the volume of brain structures obtained from MRIs. Members of the 'patient' group were victims of traumatic brain injuries (TBIs), the group 'control' consisted of uninjured participants used to estimate a 'normal' rate of change in volume associated with aging.

The following is a description of the variables:

- HPC_L_TOT: measure of the volume of the left hippocampus
- group: 'patient' or 'control'
- id: numeric id for each participant
- Age: age in years at the time of assessment

- years_pi: number of years post-injury at the time of assessment, members of the 'control' group are assigned a value of 0 for 'years_pi'.

Most members of the 'control' groups were assessed on two occasions. Members of the 'patient' group received a varying number of assessments, between 1 and 4. The first assessment occurred approximately five months post-injury, the second assessment approximately 12 months post-injury. Further assessments take place up to 4 years post-injury.

Consider the following model to estimate volume trajectories as a function of age and time post-injury:

```
fit <- lme(HPC_L_TOT ~ group * Age +
      years_pi + I(years_pi^2), dl, random = ~ 1 + years_pi | id)
summary(fit)
```

```
Linear mixed-effects model fit by REML
  Data: dl
       AIC      BIC    logLik
  4070.849 4107.548 -2025.424

Random effects:
 Formula: ~1 + years_pi | id
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 631.16197 (Intr)
years_pi     88.29700 -0.094
Residual     64.99817

Fixed effects:  HPC_L_TOT ~ group * Age + years_pi + I(years_pi^2)
                    Value Std.Error  DF   t-value p-value
(Intercept)      4318.126  334.2495 171 12.918871  0.0000
grouppatient     -362.747  380.7475 119 -0.952722  0.3427
Age                -2.326    7.5347 171 -0.308648  0.7580
years_pi         -142.482   20.5733 171 -6.925580  0.0000
I(years_pi^2)      24.672    5.4017 171  4.567459  0.0000
grouppatient:Age   -1.293    8.5996 171 -0.150378  0.8806
 Correlation:
                  (Intr) grpptn Age    yers_p I(_^2)
grouppatient      -0.878
Age               -0.925  0.812
years_pi           0.000  0.061  0.000
I(years_pi^2)      0.000  0.021  0.000 -0.822
grouppatient:Age   0.811 -0.927 -0.876 -0.098  0.001

Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max
-3.068953807 -0.242077116  0.006358397  0.218228579  3.100757301

Number of Observations: 296
Number of Groups: 121
```

a) [5 marks] Consider the trajectory for the expected volume of the left hippocampus for a TBI patient who sustains an injury at age 30 years. Sketch the expected trajectory as a function of age from 0.5 years post injury to 3 years post injury. Indicate the numerical value of the height of the trajectory at

a minimum of three selected points.

b) [5 marks] In the same sketch show the expected trajectory for an uninjured control from age 30 to age 33. Indicate the numerical value of the height of the trajectory at a minimum of two selected points.

c) [5 marks] What is the expected rate of change in the volume of the left hippocampus for TBI patient, sustaining an injury at age 30:
   1) at 1 year post-injury, and
   2) at 2 years post-injury?

d) [5 marks] Construct a hypothesis matrix that you could use to perform a Wald test to simultaneously test whether the rate of change in the volume of the left hippocampus of a TBI patient sustaining an injury at age 30 is the same at 1 year post injury and at 2 years post injury as the rate of change for an uninjured control of the same age.

e) [5 marks] What is the purpose of the quadratic term in `years_pi`? What are the advantages or disadvantages of using a quadratic term for this purpose? Discuss at least two alternative approaches that you could use to achieve this purpose?

f) [5 marks] A collaborator notices that the two coefficients that involve the effect of 'group' are not significant. They suggest that these two coefficients should be dropped since they are not significant. What is an appropriate course of action and how would you explain it to your collaborator?

Question 2: (10 marks)

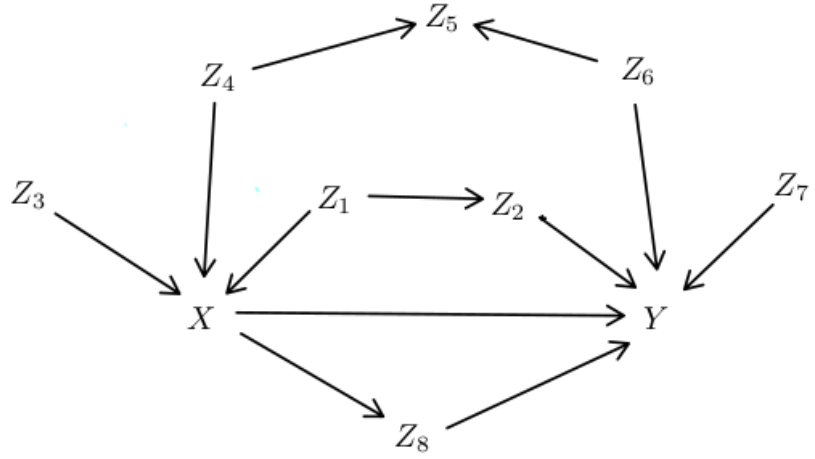
Consider the output for the G matrix in the previous question.

```
getG(fit)
```

```
Random effects variance covariance matrix
            (Intercept) years_pi
(Intercept)    398370.0  -5232.2
years_pi        -5232.2   7796.4
  Standard Deviations: 631.16 88.297
```

Interpret the parameters of this matrix. Calculate and interpret the point of minimum variance.

Question 3: (20 marks in 2 parts)



Consider the linear causal DAG above and the following models:

1. `Y ~ X`
2. `Y ~ X + Z8`
3. `Y ~ X + Z1 + Z5`
4. `Y ~ X + Z1`
5. `Y ~ X + Z2`
6. `Y ~ X + Z2 + Z7`
7. `Y ~ X + Z2 + Z3`
8. `Y ~ X + Z2 + Z8`
9. `Y ~ X + Z1 + Z5 + Z6`

a) [10 marks] For each of these models discuss briefly **whether and why** fitting the model would produce, or not, an unbiased estimate of the causal effect of `X`.

b) [10 marks] List the the models that provide an unbiased estimate of the causal effect of `X` in the same order as they appear in the list above. Compare each model with the next model in the list, explaining, for each pair, whether you can determine which model provides the smaller expected standard error of $\hat{\beta}_X$, and, if so, which model does so and why.

Question 4: (10 marks)

Write an essay on variable selection strategies in statistical analyses. What are the major relevant factors? How do they influence the choice of strategy?

Question 5: (10 marks)

Explain clearly, with an appropriate sketch, the relationship between the 'within effect', the 'between effect' and the 'contextual effect' when working with clustered data.

Question 6: (10 marks)

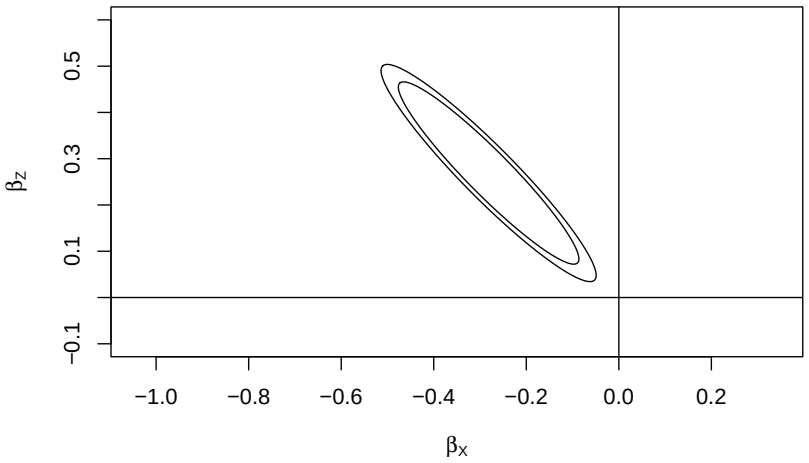
Discuss the taxonomy of outliers in regression presented in class. How do you identify the three archetypal kinds of outliers? What are the consequences of including a point of each type when the point is not, in fact, an observation from the target population?

Question 7: (20 points)

Consider the following confidence ellipses for a linear model regressing Y on X and Z. Consider three possible models for a least-squares regression of Y on X and Z:

1. $E(Y) = \beta_0 + \beta_X X + \beta_Z Z$
2. $E(Y) = \gamma_{02} + \gamma_X X$
3. $E(Y) = \gamma_{03} + \gamma_Z Z$

The following are confidence ellipses for model 1. The outer ellipse is a joint 95% confidence ellipse for the vector (β_X, β_Z) and the inner ellipse is scaled so that its orthogonal projections onto the axes produces 95% confidence intervals.



Can you determine the outcome of the following tests? If so what would be the outcome of 5% tests? Discuss briefly why. (The alternative in each case is the negation of H_0). Show the basis of your reasoning using a diagram or other explanation.

a) $H_0 : \beta_X = \beta_Z = 0$

b) $H_0 : \beta_X = 0$

c) $H_0 : \beta_Z = 0$

d) $H_0 : \gamma_X = 0$

e) $H_0 : \gamma_Z = 0$

f) $H_0 : \beta_X = \beta_Z$

g) $H_0 : \beta_X + \beta_Z = 0$

Question 8: (10 marks)

State and explain the principle of marginality. Discuss how it is an example of a principle of invariance.

Question 9: (10 marks)

Suppose a test for mononucleosis (a disease) has a specificity and a sensitivity of 95%.

a. Does this mean that the test will be wrong 5% of the time? Prove or disprove.
b. If you take the test and the result is positive, does this mean that the probability that you have mononucleosis is 95%. Prove or disprove.

Question 10: (10 marks)

Explain clearly with a suitable sketch the relationship between the 'within effect', the 'between effect' and the 'contextual effect' when working with clustered data.

Question 11: (10 marks)

The following are vocabulary scores obtained in samples of U.S. residents during the years 1974 to 2016, categorized by binary gender (Male and Female) and education (in years).

To make the coefficients easier to manipulate, 'year' has been changed to 'decade' relative to the year 2000 and the vocabulary rating has been multiplied by 100. The estimated coefficients are rounded to one decimal place.

```
library(car)
```

```
    Loading required package: carData
```

```
Vocab <- within(
  Vocab,
  {
    v100 <- vocabulary * 100
    decade <- (year - 2000)/10
  }
)
head(Vocab)
```

```
             year    sex education vocabulary decade v100
    19740001 1974   Male        14          9   -2.6  900
    19740002 1974   Male        16          9   -2.6  900
    19740003 1974 Female        10          9   -2.6  900
    19740004 1974 Female        10          5   -2.6  500
    19740005 1974 Female        12          8   -2.6  800
    19740006 1974   Male        16          8   -2.6  800
```

```
fit <- lm(v100 ~ sex * education * decade, Vocab)
smry <- summary(fit)
smry$coefficients[,"Estimate"] <- round(smry$coefficients[,"Estimate"],1)
smry$coefficients
```

```
                           Estimate Std. Error    t value      Pr(>|t|)
    (Intercept)                170.1  7.1209832 23.8861214 6.092696e-125
    sexMale                    -21.0 10.4757949 -2.0014787  4.534971e-02
    education                   33.2  0.5223064 63.5065527  0.000000e+00
    decade                      19.4  4.9964652  3.8841883  1.028900e-04
    sexMale:education            0.4  0.7622630  0.5438925  5.865195e-01
    sexMale:decade             -11.6  7.2237215 -1.6084923  1.077378e-01
```

```
education:decade             -2.7  0.3798654 -7.1622069  8.119444e-13
sexMale:education:decade      1.3  0.5432446  2.4382176  1.476558e-02
```

a) (10 points) Using this model, what is the estimated 'gender gap' (Female - Male) in v100' in the year 2000 for individuals with 20 years of education?

b) (10 points) Using this model is the gender gap in the year 1990 for individuals with 20 years of education getting narrower or getting wider? By how much per decade?

Question 12: (10 points)

Suppose you are investigating the relationships between a variable Y and two possible predictors X and Z. Is it feasible for an observation to have relatively low leverage in each of the regressions on X and on Z, but to have high leverage in the multiple regression of Y on both X and Z? Using what you know about leverage and influence discuss either why this is not feasible, or, if it is feasible, under what conditions would it be expected to happen.

Question 13: (10 points)

Consider the following (now very familiar) model regressing income (in 1,000s of dollars) on years of education in three types of occupations: bc: blue collar, wc: white collar, and prof: professional.

The coefficients have been rounded for ease of calculation.

```
library(car)
head(Prestige)
```

```
                    education income women prestige census type
gov.administrators      13.11  12351 11.16     68.8   1113 prof
general.managers        12.26  25879  4.02     69.1   1130 prof
accountants             12.77   9271 15.70     63.4   1171 prof
purchasing.officers     11.42   8865  9.11     56.8   1175 prof
chemists                14.62   8403 11.68     73.5   2111 prof
physicists              15.64  11030  5.13     77.6   2113 prof
```

```
d <- na.omit(Prestige)
d$type <- factor(d$type)
d$inc <- d$income/1000 # income in 1,000s of dollars
table(d$type)
```

```
 bc prof   wc
 44   31   23
```

```
fit <- lm(inc ~ type * education + I(education^2), d)
out <- summary(fit)
out$coefficients <- round(out$coefficients)   # to make things easier
out$coefficients
```

```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)               36         16       2        0
typeprof                  58         26       2        0
typewc                    28         13       2        0
education                 -8          4      -2        0
I(education^2)             1          0       2        0
typeprof:education        -5          2      -2        0
typewc:education          -3          1      -2        0
```

The three types of occupations are 'blue collar' (bc), 'white collar' (wc), and professional (prof).

You are a statistical consultant discussing this analysis with a client who tells you that your results don't make sense.

The negative coefficient for education says that predicted income is lower as education increases and the negative coefficient for 'typeprof:education' says that the change in income associated with additional education is lower for professional occupations than it is for blue collar occupations.

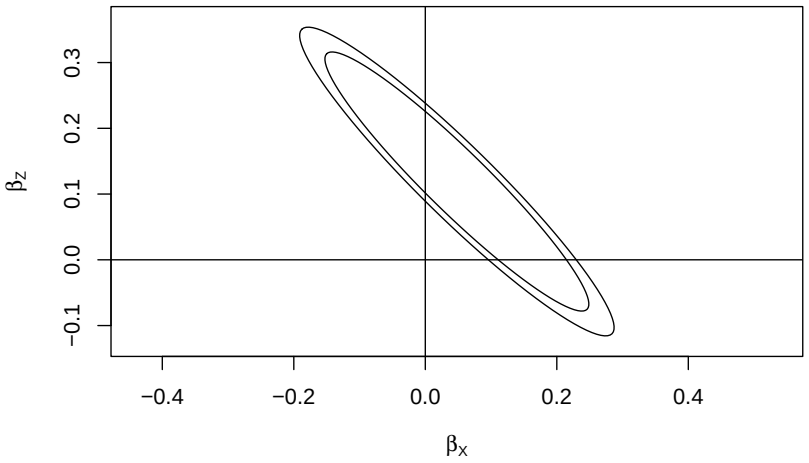
Clearly explain the interpretation of this output for your client. Take into account that the average years of education required for professional occupations is greater than for 'white collar' and 'blue collar' occupations. (Continue your answer on the back of this page.)

Question 14: (20 points)

Consider the following confidence ellipses for a linear model regressing Y on X and Z. Consider three possible models for a least-squares regression of Y on X and Z:

1. $E(Y) = \beta_0 + \beta_X X + \beta_Z Z$
2. $E(Y) = \gamma_{02} + \gamma_X X$
3. $E(Y) = \gamma_{03} + \gamma_Z Z$

The following are confidence ellipses for model 1. The outer ellipse is a joint 95% confidence ellipse for the vector (β_X, β_Z) and the inner ellipse is scaled so that its orthogonal projections onto the axes produces 95% confidence intervals.



Can you determine the outcome of the following tests? If so what would be the outcome of 5% tests? Discuss briefly why. (The alternative in each case is the negation of H_0). Show the basis of your reasoning using a diagram or other explanation.

a) $H_0 : \beta_X = \beta_Z = 0$

b) $H_0 : \beta_X = 0$

c) $H_0 : \beta_Z = 0$

d) $H_0 : \gamma_X = 0$

e) $H_0 : \gamma_Z = 0$

f) $H_0 : \beta_X = \beta_Z$

g) $H_0 : \beta_X + \beta_Z = 0$

Question 15: (10 points)

Discuss the following statement: "To choose variables in a multiple regression, you can start by testing one variable at a time and only add the variables that are significant."

Question 16: (10 points)

Consider a model regressing `Y` (e.g. math achievement) on `X` (e.g. SES) in J schools identified by a categorical variable `G`. Let `Xg` be a 'contextual variable' that is the mean of `X` within each school and let `Xd` be the 'centered-within-groups' version of `X`, i.e. `Xd = X - Xg`.

Consider the following two models:

1)
$$E(\texttt{Y}) = \beta_0 + \beta_1 \texttt{X} + \beta_2 \texttt{Xg}$$

2)
$$E(\texttt{Y}) = \psi_0 + \psi_1 \texttt{Xd} + \psi_2 \texttt{Xg}$$

Show that these models are equivalent.

Question 17: (10 points)

With reference to the previous question, derive the relationship between the parameters β_1, β_2 and the parameters ψ_1, ψ_2.

Question 18: (10 points)

With reference to the previous question, discuss what considerations would lead you to choose to fit one model versus the other?