

MATH 4939 Sample Questions for Final Exam

April 2026 (3 hours)

Total marks on this sample exam: 1,200

Total marks on final exam: approx. 180

Fill in your name on the front page but

DO NOT OPEN

until instructed to do so

Instructions:

- Aids allowed: non-programmable calculator, straight edge.
- Show your work in the empty space following each problem and on the back of the page, if necessary. You can use both the front and the back of each page.
- **Show all your work and justify your reasoning.**
- Marks sum to 20.

Question 1: (10 marks)

You are working as a consulting statistician on a project that includes a survey of 500 Canadian families. The survey 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 twice as large.

Your collaborators ask your opinion about this discrepancy, which, they claim, shows that either the sample is not representative and/or that respondents tend to dramatically under-report their equity.

How would you respond?

Question 2: (10 marks)

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 3: (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 4: (10 marks)

Is the following statement correct? Discuss. Use the geometry of confidence ellipses and data ellipses, if appropriate, in your discussion.

If you drop a variable that is not significant in a multiple linear regression, the other estimated coefficients should not change very much, nor should their significance change very much.

Question 5: (10 marks in 2 parts)

In a multiple regression of Y on two variables X_1 and X_2 , you find that the p -value for $\hat{\beta}_2$ is 0.56 and you are considering dropping X_2 from the model.

Comment on the following two statements.

- a) “Since X_2 is not significant, dropping it should not have much impact on inference for X_1 so, if our main interest concerns X_1 , there’s no problem dropping X_2 .” Discuss, referring to the geometry of confidence regions, where appropriate.
- b) (**Continuation**) Would it ever be important to keep X_2 in the model although its associated p -value does not achieve significance? Hint: Don’t forget to consider the purpose of the analysis.

Question 6: (10 marks)

Discuss the following statement: “To choose variables in a multiple regression, you can start by including all variables then drop the variables that are not significant one at a time.”

Question 7: (10 marks)

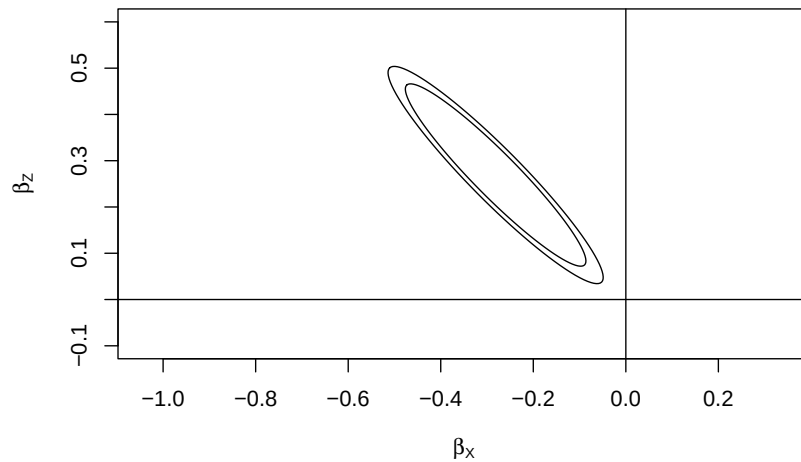
Discuss the following statement: “To choose variables in a multiple regression, you can start by including all variables, then drop the variables that are not significant.”

Question 8: (10 marks)

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 individual 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). State the basis of your reasoning by referring to the graph above or otherwise.

- 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 9: (10 marks)

Consider a multiple regression model of the form

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

where $\epsilon \sim N(0, \sigma^2 I)$ and X_1 and X_2 are matrices containing blocks of variables such that the matrix $[X_1 X_2]$ is of full column rank.

Prove that the least-squares coefficients for the regression of the (residuals of Y regressed on X_2) on the (residuals of X_1 regressed on X_2) are the same as the least-squares coefficients corresponding to X_1 in the multiple regression of Y on both $[X_1 X_2]$. (This is part of the ‘Frisch-Waugh-Lovell Theorem’, which could also be called the ‘Added-Variable-Plot Theorem’)

Question 10: (10 marks)

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 11: (10 marks)

Describe the taxonomy of outliers discussed in class. **Explain** how the inclusion of each type of outlier would influence the p-value for a test that the true slope of a linear regression is equal to zero. Is it possible for an outlier to have an effect on the p-value for a slope without having a meaningful effect on the slope itself?

Question 12: (10 marks)

Suppose you have observed a vector $\mathbf{x}$ of n observations that you wish to model with each component having an identical and independent Poisson distribution with mean θ , that is:

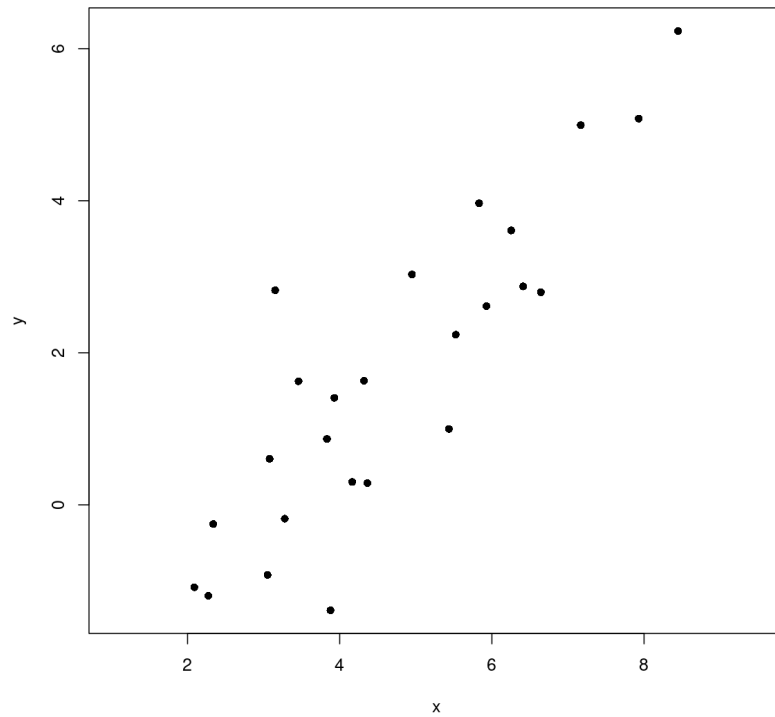
$$p(x_i; \lambda) = \frac{1}{x!} \lambda^x e^{-\lambda} \quad \lambda > 0, x = 0, 1, 2, \dots$$

.

Write a **function** in R that takes the vector of observations, $\mathbf{x}$, as input and plots the log-likelihood over a range of values for λ .

You may determine the range of values for λ from $\mathbf{x}$, or, for simplicity, use a range of values from 0.01 to 10. The function in R that computes the factorial of x , is `factorial(x)`.

Question 13: (10 marks)



Consider the scatterplot above showing a sample of 25 points from a population for two variables, x and y.

Obtain an approximate 95% confidence interval for the slope of a linear regression of y on x (obtain a value for the midpoint of the interval and for its half-width). Show your work including relevant sketches on the graph. State whether the data would lead to rejection of the null hypothesis that there is no relationship between x and y with a 5% test.

Question 14: (10 marks in 2 parts)

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 15: (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 16: (20 marks in 2 parts)

a) R uses 3-valued logic: what is the result of each of the following statements?

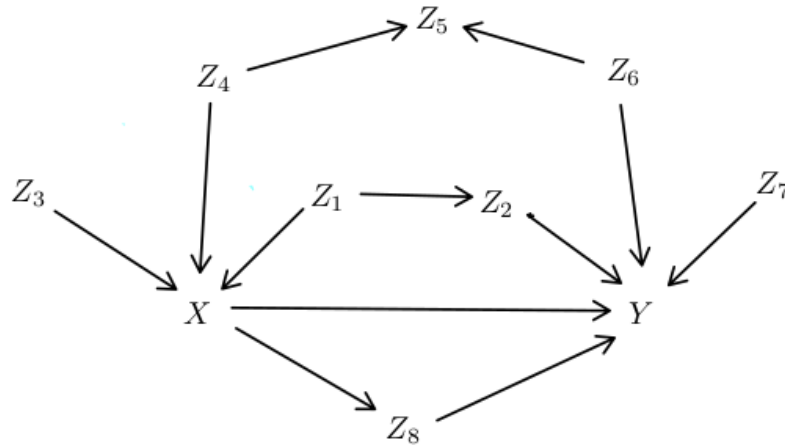
- 1) `TRUE | NA`
- 2) `TRUE & NA`
- 3) `FALSE & NA`
- 4) `NA | FALSE`

Explain briefly the 'logic' behind the results.

b) Describe the difference between the following two statements in R:

- 1) `isTRUE(x)`
- 2) `x == TRUE`

Question 17: (20 marks in 2 parts)



a) [10 marks] Consider the linear causal DAG above and the following models. For each model state briefly whether and why the model would, or would not, produce an unbiased estimate of the causal effect of X .

1. $Y \sim X$
2. $Y \sim X + Z_1 + Z_2$
3. $Y \sim X + Z_1 + Z_3 + Z_4$
4. $Y \sim X + Z_2$
5. $Y \sim X + Z_5 + Z_1 + Z_6$
6. $Y \sim X + Z_2 + Z_8$
7. $Y \sim X + Z_4$
8. $Y \sim X + Z_2 + Z_7$
9. $Y \sim X + Z_1 + Z_5 + Z_6$

b) [10 marks] Choose any two models that produce unbiased estimates of the causal effect of X and discuss

in detail whether one can be determined to have a lower standard error for its estimate of β_X .

Question 18: (20 marks in 2 parts)

In a multiple regression of Y on two variables X_1 and X_2 , you find that the p -value for $\hat{\beta}_2$ is 0.45 and you are considering dropping X_2 from the model.

Comment on the following two statements.

- a) “Since X_2 is not significant, dropping it should not have much impact on inference for X_1 so, if our main interest concerns X_1 , there’s no problem dropping X_2 .” Discuss, referring to the geometry of confidence regions, if appropriate.
- b) Are there situations where it would be important to keep X_2 in the model although its associated p -value does not achieve significance? Discuss, referring to the geometry of confidence regions and/or principles of causality, if appropriate.

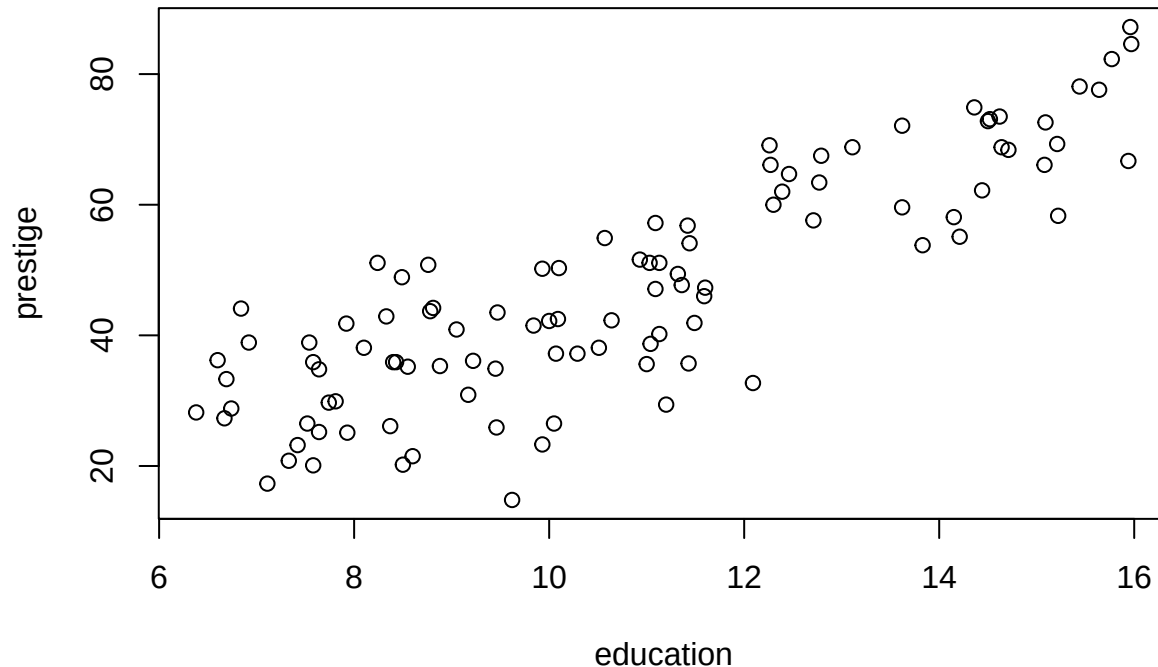
Question 19: (20 marks in 2 parts)

- a) Describe in words what each of the following statements in R does:
- a) `sum(x%%3 == 0)`
 - b) If `x` is a matrix: `apply(x, 2, function(x) sum(x%%2))`
 - c) `sum(grepl('John', x))`
 - d) `sum(grepl('John *Smith', x))`
- b) Let `mat` be a matrix, generated by Piazza, whose row names are student names and whose columns are the titles of Piazza posts. The content of the matrix is the number of posts with each title. For example, if the student named 'John' contributed 3 times to a post with the title 'Paradoxes', then the matrix would have the number 3 in the row labelled 'John' and the column labelled 'Paradoxes'. The titles used for each assignment post have the form 'A5 8.1', 'A5 8.2', etc. where 'A5' denotes assignment 5 and '8.1', etc., denote the questions within assignment 5. Write a one-liner that takes the matrix `mat` and returns a vector with the number of posts for assignment 5 posted by each student.

Question 20: (20 marks in 3 parts)

The following is a scatterplot of 'prestige' versus 'education' in 100 occupations.

```
library(car)
library(lattice)
with(Prestige, plot(education, prestige))
```



```
summary(Prestige[,c('education', 'prestige')])
```

```
##      education      prestige
##  Min.   : 6.380   Min.   :14.80
##  1st Qu.: 8.445   1st Qu.:35.23
##  Median :10.540   Median :43.60
##  Mean   :10.738   Mean   :46.83
##  3rd Qu.:12.648   3rd Qu.:59.27
##  Max.   :15.970   Max.   :87.20
```

```
round(sapply(Prestige[,c('education', 'prestige')], mean))
```

```
## education prestige
##       11         47
```

```
round(sapply(Prestige[,c('education', 'prestige')], sd))
```

```
## education prestige
##       3         17
```

- [10 marks] Visually derive a 95% confidence interval for the slope of a linear regression of 'prestige' on 'education'. Justify your answers with appropriate diagrams.
- [5 marks] What is your visual estimate of the correlation? Explain how you obtained it?

- c) [5 marks] Would a test that the slope of the linear regression is 0 achieve significance at the 0.05 level? Explain the rationale for your answer.

Question 21: (20 marks in 2 parts)

The following questions are based on the ‘Vocab’ data set used in an assignment in the course. The data set consists of 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)
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 2010 for individuals with 20 years of education?

- b) (10 points) Construct an “L matrix” to use in a Wald test to test whether there is evidence that the gender gap above is changing over time.

Question 22: (10 marks)

Explain why a random intercept model, `lme(Y ~ 1, random = ~ 1 | id)`, could also be fitted using a ‘compound symmetry’ error structure, i.e., no random effect but the within ‘id’ covariance has the form:

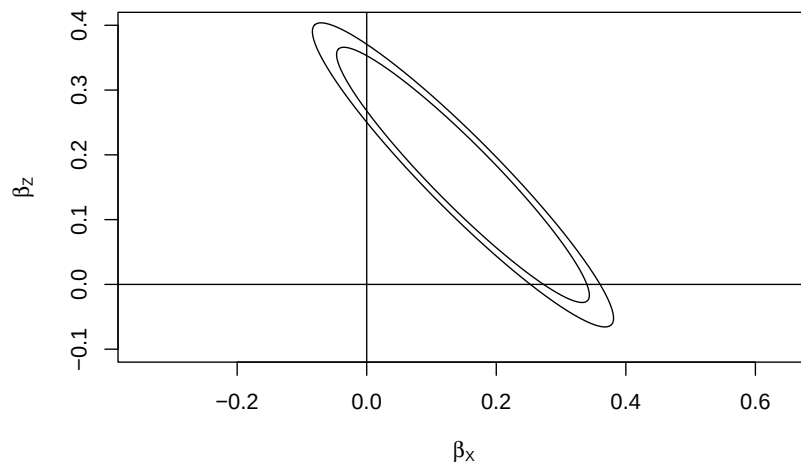
$$\text{Var}(\varepsilon_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

Question 23: (10 marks)

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. $E(Y) = \gamma_{03} + \gamma_Z Z$

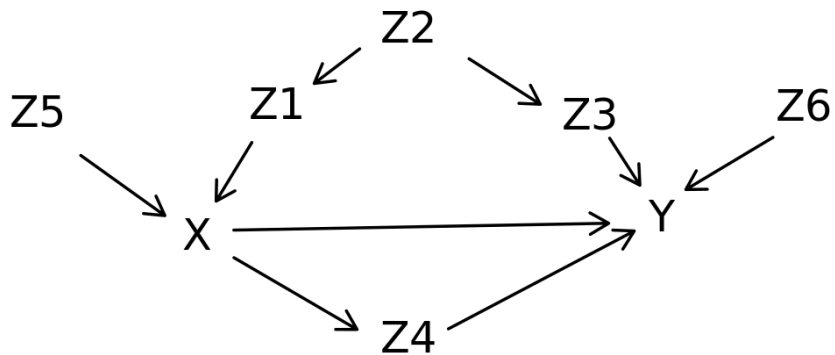
The following are confidence ellipses for model 1. The outer ellipse is a joint 95% confidence ellipse for the vector (β_X, β_Z) and the inner ellipse is scaled so that its orthogonal projections onto the axes produces 95% confidence intervals.



Provide answers to the following questions on the next page. You may also use the backs of these pages.

1. [5] 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)
 - 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) $H_0 : \gamma_Z = 0$
2. [5] Discuss the following statement: “In a multiple regression, dropping a non-significant term should have very little impact on the significance of the remaining terms.”

Question 24: (10 marks)



For the linear causal DAG above, identify one model that provides an unbiased estimate of the causal effect of **X** on **Y** that would yield a relatively small standard error for the estimated causal effect, and another model that also yields an unbiased estimate of the causal effect of **X** on **Y** but has a very large standard error. Discuss briefly why.

Question 25: (10 marks in 2 parts)

In a multiple regression of Y on two variables X_1 and X_2 , you find that the p -value for $\hat{\beta}_2$ is 0.56 and you are considering dropping X_2 from the model.

Comment on the following two statements.

- a) [5 marks] “Since X_2 is not significant, dropping it should not have much impact on inference for X_1 so, if our main interest concerns X_1 , there’s no problem dropping X_2 .” Discuss, referring to the geometry of confidence regions, where appropriate.
- b) [5 marks] Would it ever be important to keep X_2 in the model although its associated p -value does not achieve significance? Hint: Don’t forget to consider the purpose of the analysis.

Question 26: (10 marks)

Is the following statement correct? Discuss. Use the geometry of confidence ellipses and data ellipses, if appropriate, in your discussion.

If you drop a variable that is not significant in a multiple linear regression, the other estimated coefficients should not change very much, nor should their significance change very much.

Question 27: (10 marks in 2 parts)

In a multiple regression of Y on two variables X_1 and X_2 , you find that the p -value for $\hat{\beta}_2$ is 0.45 and you are considering dropping X_2 from the model.

Comment on the following two statements.

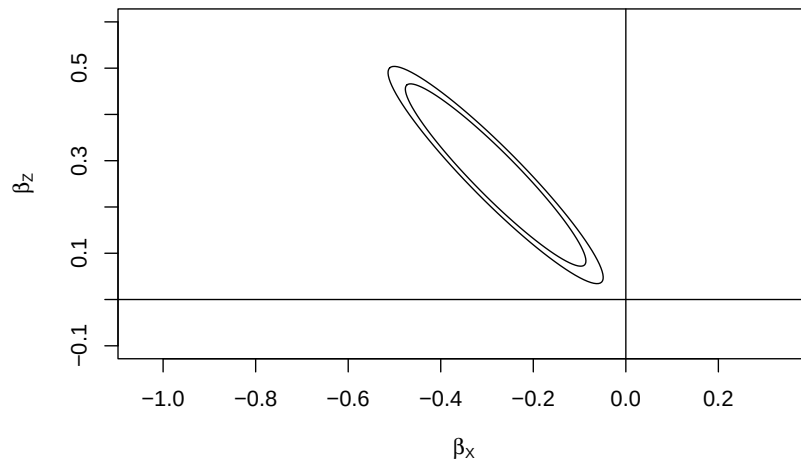
- a) “Since X_2 is not significant, dropping it should not have much impact on inference for X_1 so, if our main interest concerns X_1 , there’s no problem dropping X_2 .” Discuss, referring to the geometry of confidence regions, if appropriate.
- b) Are there situations where it would be important to keep X_2 in the model although its associated p -value does not achieve significance? Discuss, referring to the geometry of confidence regions and/or principles of causality, if appropriate.

Question 28: (10 marks)

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 individual 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). State the basis of your reasoning by referring to the graph above or otherwise.

- 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 29: (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 30: (10 marks)

Consider the following random slope models:

1. `lme(Y ~ 1 + X, data = dd, random = ~ 1 + X | id)`
2. `lme(Y ~ 1 + X, data = dd, random = ~ 1 + I(X - 10) | id)`
3. `lme(Y ~ 1 + X, data = dd, random = ~ 1 + dvar(X, id) | id)`

Discuss what features of your data would lead you to prefer one of these models over the others.

Question 31: (10 marks)

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 32: (10 marks)

[Note: All regressions in this question are least-squares regressions that include an intercept term] Consider observations on three numerical variables, Y , X and Z . Let $X_{.Z}$ denote the residuals of the least-squares regression of X regressed on Z .

Show that the least-squares coefficient for X in the multiple linear regression of Y on X and Z (i.e., $\text{lm}(Y \sim X + Z)$) is equal to the least-squares coefficient obtained by regressing Y on $X_{.Z}$. You may use the Frisch-Waugh-Lovell theorem.

Question 33: (10 marks)

[Note: All regressions in this question are least-squares regressions that include an intercept term] Consider observations on three numerical variables, Y , X and Z . Let $X_{.Z}$ denote the residuals of the least-squares regression of X regressed on Z .

Show that the least-squares coefficient for X in the multiple linear regression of Y on X and Z (i.e., $\text{lm}(Y \sim X + Z)$) is equal to the least-squares coefficient obtained by regressing Y on $X_{.Z}$. You may use the Frisch-Waugh-Lovell theorem.

Question 34: (10 marks)

Consider a multiple regression model of the form

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

where $\epsilon \sim N(0, \sigma^2 I)$ and X_1 and X_2 are matrices containing blocks of variables such that the matrix $[X_1 X_2]$ is of full column rank.

Prove that the least-squares coefficients for the regression of the (residuals of Y regressed on X_2) on the (residuals of X_1 regressed on X_2) are the same as the least-squares coefficients corresponding to X_1 in the multiple regression of Y on both $[X_1 X_2]$. (This is part of the ‘Frisch-Waugh-Lovell Theorem’, which could also be called the ‘Added-Variable-Plot Theorem’)

Question 35: (10 marks)

Consider a random coefficients model for a model with time as a predictor and with a random intercept and a random slope:

```
lme(Y ~ 1 + Time, data = dd, random = ~ 1 + Time | id)
```

Let

$$E(Y) = \gamma_0 + \gamma_1 \times Time$$

Let the G matrix have the decomposition

$$G = AA' \text{ with } A = \begin{pmatrix} a_{00} & a_{10} \\ 0 & a_{11} \end{pmatrix}$$

Draw graphs in “data space” and in “beta space” showing the interpretation of the parameters γ_0 , γ_1 , a_{00} , a_{10} and a_{11} .

Question 36: (10 marks)

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 37: (10 marks in 2 parts)

a) R uses 3-valued logic: what is the result of each of the following statements?

- 1) `TRUE | NA`
- 2) `TRUE & NA`
- 3) `FALSE & NA`
- 4) `NA | FALSE`

Explain briefly the 'logic' behind the results.

b) Describe the difference between the following two statements in R:

- 1) `isTRUE(x)`
- 2) `x == TRUE`

Question 38: (10 marks)

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 39: (10 marks)

Suppose you have observed a vector $\mathbf{x}$ of n observations that you wish to model with each component having an identical and independent exponential distribution with mean θ , that is:

$$f(x_i; \theta) = \frac{1}{\theta} \exp -x/\theta \quad x_i, \theta > 0$$

.

Write a function in R that takes the vector of observations, $\mathbf{x}$, as input and plots the log-likelihood over a range of values for θ .

You may determine the range of values for θ from $\mathbf{x}$, or, for simplicity, use a range of values from 0.01 to 10.

Question 40: (10 marks)

Consider fitting the following model:

```
fit <- lm(Y ~ sp(X), data = dd)
```

where `sp()` is a function defined as

```
sp <- function(x) {  
  spida2::gsp(x,  
    knots = c(1,2,3),  
    degree = c(0,1,1,0),  
    smooth = c(-1,0,0))  
}
```

The summary output is the following (with estimates rounded):

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.00000	0.09020	55.836	<2e-16 ***
sp(x)C(1).0	-1.00000	0.19217	-0.401	0.688
sp(x)C(1).1	2.00000	0.24143	0.821	0.412
sp(x)C(2).1	-1.00000	0.36409	-0.915	0.361

Draw a sketch of the predicted values over the interval of values of 'x' from 0 to 4. Indicate where and how the estimated parameters appear in the sketch.

Question 41: (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 42: (10 marks)

Describe the taxonomy of outliers discussed in class. **Explain** how the inclusion of each type of outlier would influence the p-value for a test that the true slope of a linear regression is equal to zero. Is it possible for an outlier to have an effect on the p-value for a slope without having a meaningful effect on the slope itself?

Question 43: (20 marks)

Describe the taxonomy of outliers discussed in class. **Explain** how the inclusion of each type of outlier would influence confidence intervals for the slope of a linear regression. (Note that confidence intervals have a *position* and a *width*)

Question 44: (10 marks)

Consider a random coefficients model for a model with time as a single predictor using a random intercept and a random slope, i.e., $\text{lm}(Y \sim T, \text{data} = \text{dd}, \text{random} = \sim 1 + T \mid \text{id})$.

Let the G matrix have the decomposition $G = AA'$, where

$$A = \begin{pmatrix} a_{00} & a_{10} \\ 0 & a_{11} \end{pmatrix}$$

Show that a_{00} is the estimated standard deviation of the heights of cluster true regression lines over the value of T at which these lines have minimum variance.

Question 45: (10 marks)

Let Σ be a symmetric matrix. Show that Σ is positive definite if and only if there exists a non-singular matrix A such that $\Sigma = AA'$.

You may use the Spectral Decomposition Theorem: Σ is symmetric if and only if there exists an orthogonal matrix Γ and a diagonal matrix Λ such that $\Sigma = \Gamma\Lambda\Gamma'$.

Question 46: (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 47: (15 marks in 2 parts)

Consider a random coefficients model for a model with time as a predictor and with a random intercept and a random slope:

```
lme(Y ~ 1 + Time, data = dd, random = ~ 1 + Time | id)
```

Let the G matrix have the decomposition

$$G = AA' \text{ with } A = \begin{pmatrix} a_{00} & a_{10} \\ 0 & a_{11} \end{pmatrix}$$

1. [10] Explain and justify the interpretation of the fixed effects and random effects parameters (i.e. a_{00} , a_{10} , a_{11} , and σ^2) of this model. Draw a sketch to illustrate your answer.
2. [5] Discuss the circumstances in which it could be preferable to use

```
random = ~ 1 + dvar(Time,id) | id
```

instead of

```
random = ~ 1 + Time | id
```

for the random portion of the mixed model.

Question 48: (15 marks in 3 parts)

Write a brief discussion with illustrations (using ellipses in data space or beta space if you wish) explaining the following issues concerning the interpretation of p-values in the output produced by the ‘summary’ method applied to a mixed model for longitudinal data analysis.

1. [5] Why should p-values for main effects that are marginal to an interaction be interpreted with caution. Are there any situations in which it is legitimate to use them?
2. [5] If two main effects that are not marginal to any interaction have non-significant p-values, is it appropriate to drop both terms?
3. [5] If a time-varying variable is included as a raw variable and as a contextual variable, what is the possible consequence of dropping the contextual variable if its p-value is not significant?

Question 49: (10 marks)

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

Consider the following two models:

1)

$$E(Y) = \beta_0 + \beta_1 X + \beta_2 X_g$$

2)

$$E(Y) = \psi_0 + \psi_1 X_d + \psi_2 X_g$$

Show that these models are equivalent.

Question 50: (10 marks)

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

Question 51: (20 marks in 2 parts)

Consider a random intercept model: `lme(Y ~ 1, random = ~ 1 | id)`.

1. [10] Explain how such a model could also be fitted using a ‘compound symmetry’ error structure, e.g., a `gls` model with no random effects but with a within ‘`id`’ covariance structure of the form:

$$\text{Var}(\varepsilon_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}, -1 < \rho < 1$$

2. [10] Explain whether or not the converse is valid, and why. If not, what kinds of data situations would be appropriate for a compound symmetry model but not for a random intercept model.

Question 52: (20 marks)

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

Consider the following two models:

1)

$$E(Y) = \beta_0 + \beta_1 X + \beta_2 X_g$$

2)

$$E(Y) = \psi_0 + \psi_1 X_d + \psi_2 X_g$$

Show that these models yield the same predicted value for Y , that $\hat{\beta}_1 = \hat{\psi}_1$ and that both of these estimates have the same standard error.

Question 53: (10 marks)

Consider an OLS model regressing Y (e.g. math achievement) on X (e.g. SES) in J schools identified by a categorical variable S . Let X_s be a ‘contextual variable’ that is the mean of X within each school and let X_d be the ‘centered-within-groups’ version of X , i.e. $X_d = X - X_s$.

Consider the following two models:

1)

$$E(Y) = \beta_0 + \beta_1 X + \beta_2 X_s$$

2)

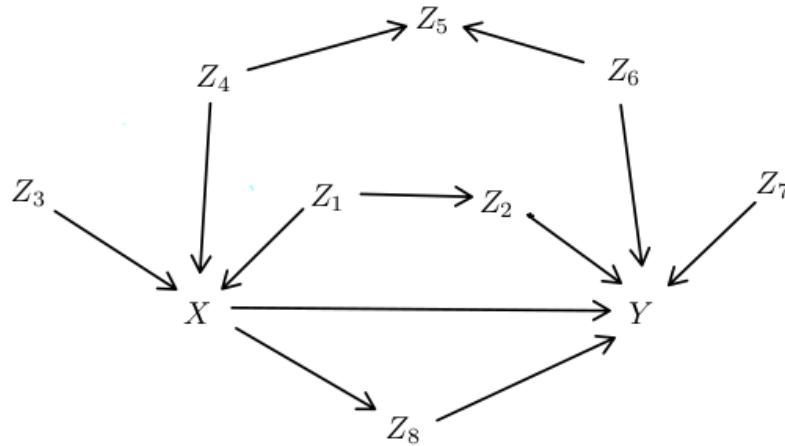
$$E(Y) = \psi_0 + \psi_1 X_d + \psi_2 X_s$$

Show that these models yield the same predicted value for Y if fitting an OLS model.

Question 54: (10 marks)

With reference to the previous question, derive the relationship between the parameters β_1, β_2 and the parameters ψ_1, ψ_2 , assuming that the two models define equal functions over possible values of $\mathbf{X}$, $\mathbf{X}_s$ and $\mathbf{X}_d$.

Question 55: (20 marks in 2 parts)

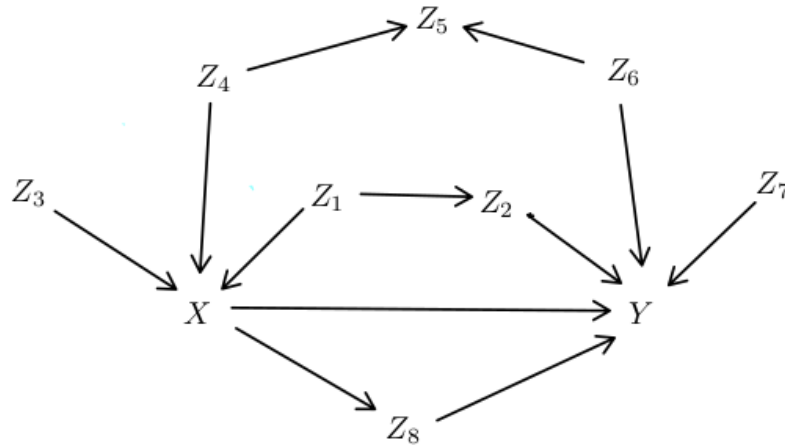


Consider the linear causal DAG above and the following models:

1. $Y \sim X$
2. $Y \sim X + Z_1 + Z_5$
3. $Y \sim X + Z_1$
4. $Y \sim X + Z_8$
5. $Y \sim X + Z_2$
6. $Y \sim X + Z_2 + Z_6$
7. $Y \sim X + Z_2 + Z_8$
8. $Y \sim X + Z_2 + Z_7$
9. $Y \sim X + Z_1 + Z_5 + Z_6$

- 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] Choose any two models that produce unbiased estimates of the causal effect of X and discuss in detail whether one can be determined to have a lower standard error for its estimate of β_X and why.

Question 56: (20 marks in 2 parts)



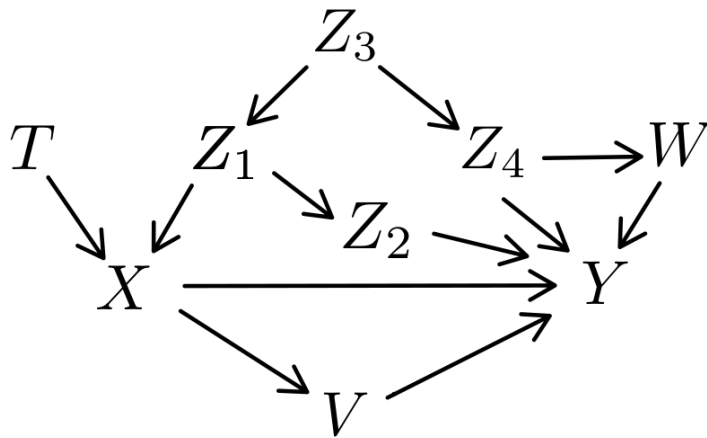
a) [10 marks] Consider the linear causal DAG above and the following models. For each model state briefly whether and why the model would, or would not, produce an unbiased estimate of the causal effect of X .

1. $Y \sim X$
2. $Y \sim X + Z_1 + Z_2$
3. $Y \sim X + Z_1 + Z_3 + Z_4$
4. $Y \sim X + Z_2$
5. $Y \sim X + Z_5 + Z_1 + Z_6$
6. $Y \sim X + Z_2 + Z_8$
7. $Y \sim X + Z_4$
8. $Y \sim X + Z_2 + Z_7$
9. $Y \sim X + Z_1 + Z_5 + Z_6$

b) [10 marks] Choose any two models that produce unbiased estimates of the causal effect of X and discuss

in detail whether one can be determined to have a lower standard error for its estimate of β_X .

Question 57: (20 marks in 2 parts)



a) [10 marks] Consider the linear causal DAG above and the following models. For each model state briefly whether and why the model would, or would not, produce an unbiased estimate of the causal effect of X .

1. $Y \sim X$

2. $Y \sim X + Z_1 + W$

3. $Y \sim X + Z_1$

4. $Y \sim X + Z_2 + W$

5. $Y \sim X + Z_1 + T$

6. $Y \sim X + Z_2 + Z_4$

7. $Y \sim X + Z_3 + Z_3$

8. $Y \sim X + Z_1 + V$

9. $Y \sim X + Z_2 + W$

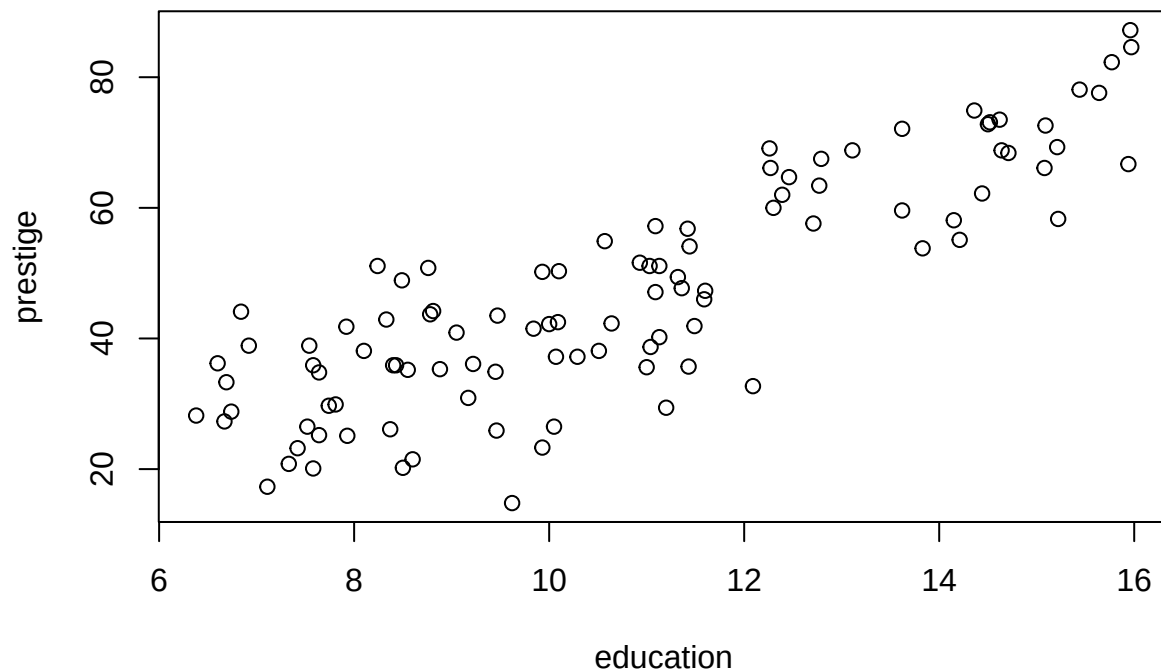
b) [10 marks] Choose any two models that produce unbiased estimates of the causal effect of X and discuss

in detail whether one can be determined to have a lower standard error for its estimate of β_X .

Question 58: (20 marks in 3 parts)

The following is a scatterplot of 'prestige' versus 'education' in 100 occupations.

```
library(car)
with(Prestige, plot(education, prestige))
```



```
round(sapply(Prestige[,c('education', 'prestige')], mean))
```

```
education  prestige
      11         47
```

```
round(sapply(Prestige[,c('education', 'prestige')], sd))
```

```
education  prestige
      3         17
```

- [10 marks] Visually derive a 95% confidence interval for the slope of a linear regression of 'prestige' on 'education'. Justify your answers with appropriate diagrams.
- [5 marks] What is your visual estimate of the correlation? Explain how you obtained it?

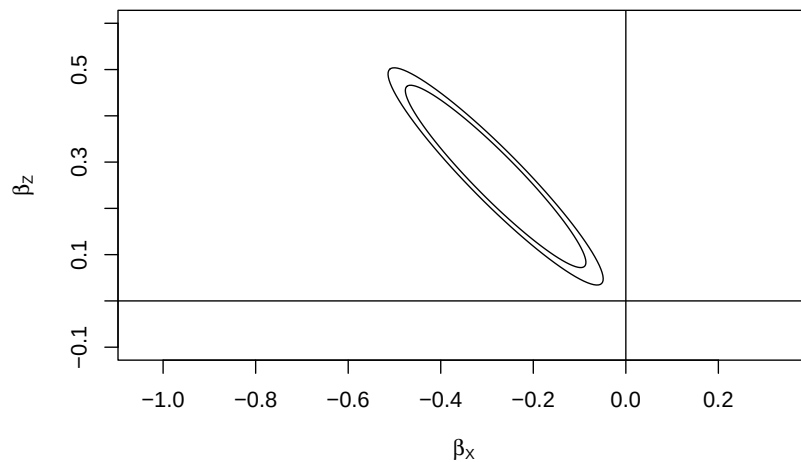
- c) [5 marks] Would a test that the slope of the linear regression is 0 achieve significance at the 0.05 level? Explain the rationale for your answer.

Question 59: (20 marks)

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.



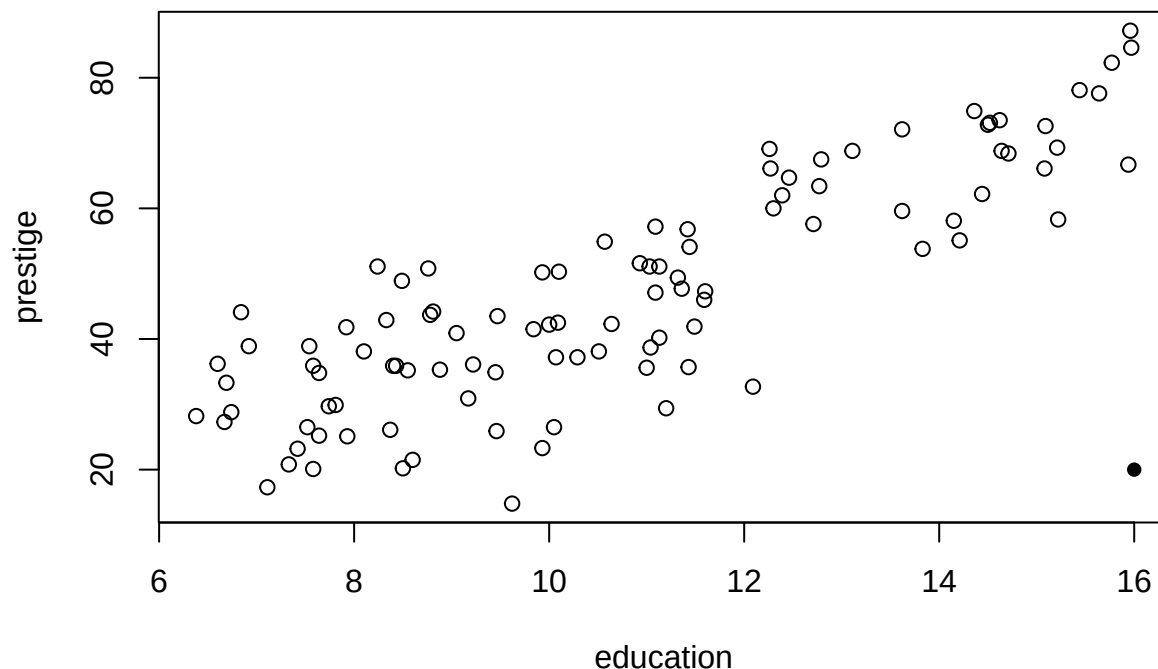
Can you determine the outcome of the following tests? If so what would be the outcome of 5% tests? Discuss briefly why. (The alternative in each case is the negation of H_0). Show the basis of your reasoning using a diagram or other explanation.

- 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 60: (20 marks in 2 parts)

The following is a scatterplot of 'prestige' versus 'education' in approximately 100 occupations, including one unusual occupation with an education level of 16 and a prestige of 20, show as a solid black circle. Consider the linear regression of prestige on education, with and without the unusual occupation.

```
library(car)
library(lattice)
with(Prestige, plot(education, prestige))
points(16,20, pch = 16)
```



```
summary(Prestige[,c('education', 'prestige')])
```

```
##      education      prestige
##  Min.   : 6.380   Min.   :14.80
## 1st Qu.: 8.445   1st Qu.:35.23
## Median :10.540   Median :43.60
## Mean   :10.738   Mean   :46.83
## 3rd Qu.:12.648   3rd Qu.:59.27
## Max.   :15.970   Max.   :87.20
```

```
round(sapply(Prestige[,c('education', 'prestige')], mean))
```

```
## education prestige
##        11        47
```

```
round(sapply(Prestige[,c('education', 'prestige')], sd))
```

```
## education prestige
##         3         17
```

a) What is the value of the leverage for the unusual point? Justify your answer.

- b) How much would the inclusion of the unusual occupation change the height of the fitted line at $x = 16$? Justify your answer.

Question 61: (20 marks)

Let Y and X be numerical variables and let G be a factor. Consider the following models. All but one will produce the same regression coefficient for X or $X_{\cdot G}$, where $X_{\cdot G}$ denotes the residuals of X regressed on G , but they will produce possibly different estimated standard errors.

Identify the model that produces a possibly different estimate of the coefficient.

Rank the others, where possible, according to the standard error of estimated coefficient for X stating which would be equal (if any) – assuming a very large n and ignoring the effect of slight differences in degrees of freedom for the error term.

Explain your reasoning briefly. Note that, following R notation, all the models below implicitly include an intercept term although in some models the fitted coefficient of the intercept be known to be zero.

- a) $Y \sim X + G$
- b) $Y \sim X$
- c) $Y_{\cdot G} \sim X_{\cdot G}$
- d) $Y \sim X_{\cdot G}$
- e) $Y \sim X + \hat{X}_G$ where $\hat{X}_G$ is the predictor of X in the regression of X on G
- f) $Y \sim X + \hat{X}_G + Z_G$ where Z_G is a “G-level” predictor, i.e. it has a constant value within each level of

G

.

Question 62: (20 marks)

Consider the following 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.

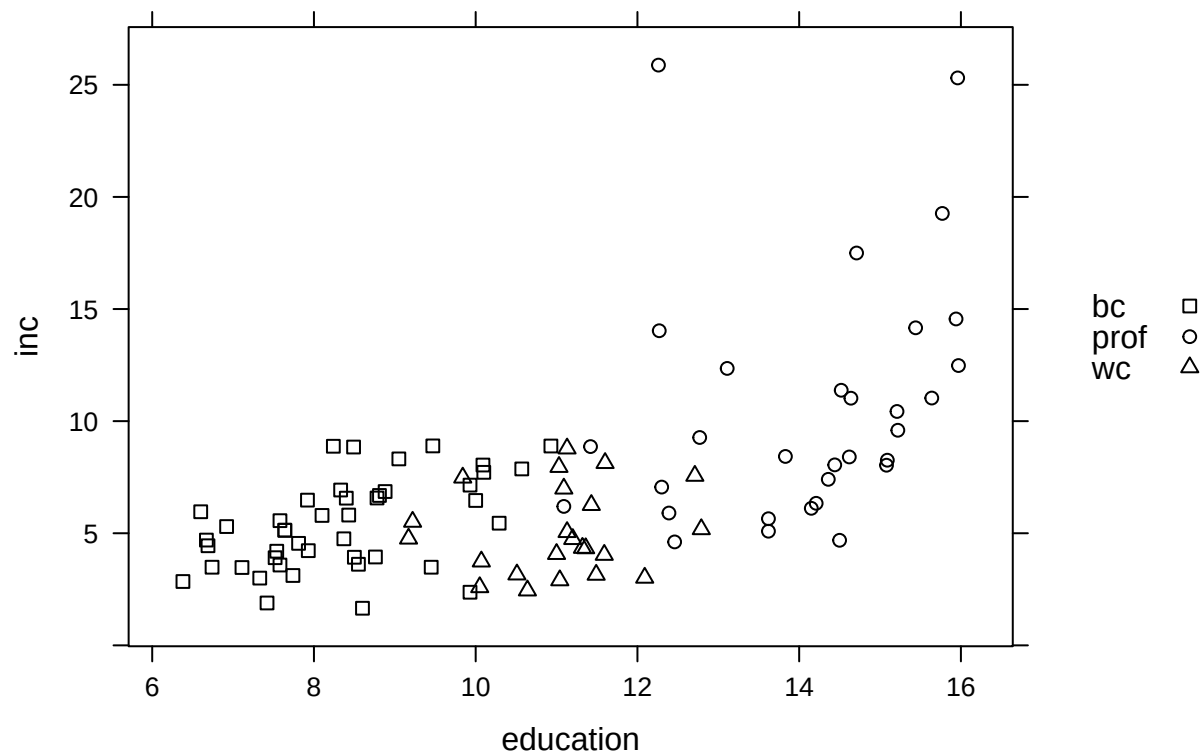
```
library(car)
library(lattice)
head(Prestige)
```

```
##               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
```

```
print(
  xyplot(inc ~ education, d, groups = type, auto.key = T,
    par.settings =
      list(superpose.symbol =
        list(pch = c(0,1,2), col = 'black'))))
```



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

Write a short essay explaining whether and how you should change your model to take these issues into account. If you believe that the model has a reasonable interpretation, write an explanation in response to your client's concerns. Write the explanation in terms your client would understand. Assume the client is a graduate student in a field other than mathematics or statistics.

Question 63: (20 marks in 2 parts)

Consider a random intercept model: `lme(Y ~ 1, random = ~ 1 | id)`.

1. [10] Explain how such a model could also be fitted using a ‘compound symmetry’ error structure, e.g., a `gls` model with no random effects but with a within ‘`id`’ covariance structure of the form:

$$\text{Var}(\varepsilon_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}, -1 < \rho < 1$$

2. [10] Explain whether or not the converse is valid, and why. If not, what kinds of data situations would be appropriate for a compound symmetry model but not for a random intercept model.

Question 64: (20 marks in 2 parts)

[Note: All regressions in this question are least-squares regressions that include an intercept term] Consider observations on three numerical variables, Y, X and Z . Let $X_{\cdot Z}$ denote the residuals of the least-squares regression of X regressed on Z , and let $\hat{X}_Z$ denote the least-squares predictor of X using Z

1. [10] Prove an extension of the Frisch-Waugh-Lovell Theorem that states that the least-squares coefficient for X in the multiple linear regression of Y on X and Z (i.e., $\text{lm}(Y \sim X + Z)$) is equal to the least-squares coefficient obtained by regressing Y on $X_{\cdot Z}$.
2. [10] Using the previous result, or otherwise, show that the regression coefficient for X in the regression of Y on X and $\hat{X}_Z$ is equal to the regression coefficient of X in the regression of Y on X and Z .

Question 65: (20 marks in 2 parts)

Consider a random coefficients model for a model with time as a predictor and with a random intercept and a random slope:

```
lme(Y ~ 1 + Time, data = dd, random = ~ 1 + Time | id)
```

Let the G matrix have the decomposition

$$G = AA' \text{ with } A = \begin{pmatrix} a_{00} & a_{10} \\ 0 & a_{11} \end{pmatrix}$$

1. [10] Explain and justify the interpretation of the fixed effects and random effects parameters (i.e. a_{00} , a_{10} , a_{11} , and σ^2) of this model. Draw a sketch to illustrate your answer.
2. [10] Discuss the circumstances in which it could be preferable to use:

```
random = ~ 1 + dvar(Time,id) | id
```

instead of:

```
random = ~ 1 + Time | id
```

for the random portion of the mixed model.

Would it ever be appropriate to use:

```
random = ~ 1 + cvar(Time,id) | id?
```

Question 66: (20 marks)

Consider the following model used to analyze the recovery of Verbal IQ (VIQ) after awakening from coma following a traumatic brain injury. The variable DCOMA records the duration of coma, and DAYSPC is the number of days following the end of coma on which VIQ was measured.

```
library(spida2)
library(nlme)
head(iq)

fit <- nlme(VIQ ~ b0 + b1*exp(-a*(DAYSPC-90)), data = iq,
            subset = DCOMA < 100,
            fixed = list(
              b0 ~ 1 + sqrt(DCOMA),
              b1 ~ 1 + sqrt(DCOMA),
              a ~ 1),
            random = list(
              ID = list(b0 ~ 1, b1 ~ 1)
            ),
            start = list(
              fixed = c(96, -1, -10, -1, .005)
            ),
            control = list(maxIter=1000,
                           msVerbose = TRUE, niterEM = 100,
                           pnlsMaxIter = 20,
                           msMaxIter = 1000, returnObject = T),
            verbose = TRUE
)
```

The following is a portion of the output from the model with some results rounded for convenience:

Nonlinear mixed-effects model fit by maximum likelihood

Model: $VIQ \sim b_0 + b_1 * \exp(-a * (DAYSPC - 90))$

Data: iq

Subset: DCOMA < 100

AIC	BIC	logLik
2462.861	2496.888	-1222.431

Random effects:

Formula: list(b0 ~ 1, b1 ~ 1)

Level: ID

Structure: General positive-definite, Log-Cholesky parametrization

	StdDev	Corr
b0.(Intercept)	14.000578	b0.(I)
b1.(Intercept)	1.752522	-1
Residual	5.407284	

Fixed effects: list(b0 ~ 1 + sqrt(DCOMA),
b1 ~ 1 + sqrt(DCOMA),
a ~ 1)

	Value	Std.Error	DF	t-value	p-value
b0.(Intercept)	100.0000	1.9780269	196	50.43970	0.0000

```

b0.sqrt(DCOMA)  -0.5000  0.5562611  123  -0.86979  0.3861
b1.(Intercept)  -3.0000  1.2254038  123  -2.57951  0.0111
b1.sqrt(DCOMA)  - 1.0000  0.3404744  123  -2.10055  0.0377
a               0.0100  0.0022118  123   5.70374  0.0000

```

Interpret the parameters of this model and draw a sketch showing a predicted recovery curve over a period of two years for a subject with a very short duration of coma approaching 0 days, and for a subject with a duration of coma of 25 days.

Question 67: (20 marks)

Suppose you have data on a continuous variable, Y , and two variables X and Z that are dichotomous with values 0 and 1.

The following table shows the values of $\bar{Y}$ for all the combinations of values of X and Z :

	X=0	X=1
Z = 0	8	6
Z = 1	3	2

The number of observations in each cell is:

	X=0	X=1
Z = 0	30	60
Z = 1	120	20

1. Draw the Paik-Agresti diagram for this data.
2. What is(are) the value(s) of the ‘conditional effect(s) of X’?
3. What is the value of the ‘marginal effect of X’?
4. If you were to fit the model $Y \sim X*Z$ in R, i.e. the model

$$\hat{Y} = \hat{\beta}_0 + X\hat{\beta}_1 + Z\hat{\beta}_2 + XZ\hat{\beta}_3$$

what would the values of the estimated regression coefficients be?

5. How would you get the ‘conditional effects of X’ from the regression coefficients? Express your answer in the form of a ‘hypothesis matrix’ multiplying the vector of fitted values, i.e. a matrix L so that your answer would have the form $L\hat{\beta}$.
6. What is the interpretation of the interaction coefficient, $\hat{\beta}_3$?
7. Why is the use of the word ‘effects’ problematic in this context? Can you think of a better word?
8. “Challenging – or tedious – so not on the quiz?” How would you get the marginal effect of X from the regression coefficients? What is the connection with the chain rule in multivariate calculus?

Question 68: (20 marks in 2 parts)

- a) Describe in words what each of the following statements in R does:

- a) `sum(x%%3 == 0)`
- b) If `x` is a matrix: `apply(x, 2, function(x) sum(x%%2))`
- c) `sum(grepl('John', x))`
- d) `sum(grepl('John *Smith', x))`

- b) Let `mat` be a matrix, generated by Piazza, whose row names are student names and whose columns are the titles of Piazza posts. The content of the matrix is the number of posts with each title. For example, if the student named ‘John’ contributed 3 times to a post with the title ‘Paradoxes’, then the matrix would have the number 3 in the row labelled ‘John’ and the column labelled ‘Paradoxes’. The titles used for each assignment post have the form ‘A5 8.1’, ‘A5 8.2’, etc. where ‘A5’ denotes assignment 5 and ‘8.1’, etc., denote the questions within assignment 5.

Write a one-liner that takes the matrix `mat` and returns a vector with the number of posts for assignment 5 posted by each student.

Question 69: (20 marks in 2 parts)

Data is obtained on the effects of two treatments (A and B). Data was recorded on symptoms weekly throughout 10 weeks of active treatment and then for 8 more weeks following termination of the treatment. A linear spline model is fit to the data and the following results were obtained [some numbers were rounded to facilitate arithmetic]:

```
sp <- function( x ) {
  gsp( x,
    knots = c( 10 ),      # 1 knots => 2 intervals
    degree = c( 1 , 1 ) , # linear in each interval
    smooth = c( 0 )       # continuous at the knot
  )
}
```

```
> fit <- lme( Symptom ~ sp(Weeks) * tx, random = ~ 1 + Weeks | id, data = d)
> summary(fit)
Linear mixed-effects model fit by REML
Data: d
      AIC      BIC    logLik
15099.44 15153.18 -7539.719

Random effects:
  Formula: ~1 + Weeks | id
Structure: General positive-definite, Log-Cholesky parametrization
              StdDev    Corr
(Intercept) 37.921194 (Intr)
Weeks        2.395329 -0.218
Residual     23.285420

Fixed effects: Symptom ~ sp(Weeks) * tx
              Value Std.Error   DF   t-value p-value
(Intercept)    131.0  6.450470 1516   20.43957  0.0000
sp(Weeks)D1(0)   -11.0  0.508022 1516  -22.08346  0.0000
sp(Weeks)C(10).1  18.0  0.571104 1516   31.94835  0.0000
txB              -2.5  9.122343   78   -0.28679  0.7750
sp(Weeks)D1(0):txB  1.5  0.718452 1516    2.15765  0.0311
sp(Weeks)C(10).1:txB -5.0  0.807664 1516   -5.99001  0.0000
. . .

Number of Observations: 1600
Number of Groups: 80
```

1. [10] Give a precise interpretation of the fixed effects regression coefficients in the model.
2. [10] Sketch the predicted trajectories for each treatment. Label the trajectories and indicate the heights of the curves at critical points.

Question 70: (20 marks in 2 parts)

Consider a random coefficients model for a model with time as a single predictor using a random intercept and a random slope, i.e., `lm(Y ~ T, data = dd, random = ~ 1 + T | id)`.

Let the G matrix have the decomposition $G = AA'$, where

$$A = \begin{pmatrix} a_{00} & a_{10} \\ 0 & a_{11} \end{pmatrix}$$

1. [10] Show that a_{00} is the estimated standard deviation of the heights of cluster true regression lines over the value of T at which these lines have minimum variance.
2. [10] Describe and prove the interpretation for a_{10} and for a_{11} .

Question 71: (20 marks in 2 parts)

The following questions are based on the 'Vocab' data set used in an assignment in the course. The data set consists of 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)
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 2010 for individuals with 20 years of education?

- b) (10 points) Construct an “L matrix” to use in a Wald test to test whether there is evidence that the gender gap above is changing over time.

Question 72: (20 marks in 4 parts)

Consider a random coefficients model for a model with time as a single predictor using a random intercept and a random slope, i.e., `lm(Y ~ 1 + T, data = dd, random = ~ 1 + T | id)`.

Let the G matrix have the decomposition $G = AA'$, where

$$A = \begin{pmatrix} a_{00} & a_{10} \\ 0 & a_{11} \end{pmatrix}$$

1. [5] Derive the x-coordinate of the focal point (the value of T over which random individual regression lines have minimum standard deviation).
2. [5] Derive the value of the standard deviation of individual regression lines above and below the focal point.
3. [5] Suppose the x-coordinate of the focal point is estimated to be equal to 10 and the model takes many iterations to converge. How could you write an equivalent model that might improve convergence?
4. [10] If the mean value of T varies from subject to subject, discuss other models you should consider with this data?

Question 73: (10 marks)

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

Consider the following two models:

1)

$$E(Y) = \beta_0 + \beta_1 X + \beta_2 X_g$$

2)

$$E(Y) = \psi_0 + \psi_1 X_d + \psi_2 X_g$$

Show that these models are equivalent.

Question 74: (10 marks)

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

Question 75: (10 marks)

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

Question 76: (30 marks in 6 parts)

Consider a multilevel longitudinal random slopes model where the model for the i th subject is given by:

$$Y_i = X_i\gamma + X_iu_i + \varepsilon_i$$

where X_i is a $n_i \times 2$ matrix whose first column is a column of 1's and whose second column consists of the times when the i th person was observed. Furthermore:

$$u_i \sim N(0, G), \varepsilon_i \sim N(0, \sigma^2 I), u_i \perp \varepsilon_i, i = 1, \dots, M$$

Let $\beta_i = \gamma + u_i$.

In the following questions, treat the parameters γ , G , and σ as if they were fixed and known.

- a. [5] Give an expression for the BLUE of β_i .
- b. [5] Give an expression for the BLUP of β_i as a function of the BLUE and other non-random quantities.
- c. [5] Describe how the BLUP is a shrinkage estimator based on the BLUE.
- d. [5] For large n_i and a relatively fixed variance for the values of X , Σ_X , how will the BLUP behave relative to the BLUE?
- e. [5] What are the implications for the BLUP if G is highly concentrated in some direction, i.e. G is nearly singular of rank 'close' to 1?
- f. [5] Under what conditions (sketching data to provide an example) would you expect $\hat{\gamma}_X$ to have a different sign than any of the individual $\hat{\beta}_{X_i}$'s?

Question 77: (30 marks in 3 parts)

Data is obtained on the effects of two treatments (A and B). Data was recorded on symptoms weekly throughout 10 weeks starting with week 1. The following results were obtained [some numbers are rounded to facilitate arithmetic]:

```
> fit <- lme( Symptom ~ Weeks * tx, random = ~ 1 + Weeks | id, data = d)
> summary(fit)
```

Linear mixed-effects model fit by REML

Data: d

	AIC	BIC	logLik
	15099.44	15153.18	-7539.719

Random effects:

Formula: ~1 + Weeks | id

Structure: General positive-definite, Log-Cholesky parametrization

	StdDev	Corr
(Intercept)	37.921194	(Intr)
Weeks	2.395329	-0.218
Residual	23.285420	

Fixed effects: Symptom ~ Weeks * tx

	Value	Std.Error	DF	t-value	p-value
(Intercept)	131.0	6.450470	1516	20.43957	0.0000
Weeks	-11.0	0.508022	1516	-22.08346	0.0000
txB	-2.5	9.122343	78	-0.28679	0.7750
Weeks:txB	1.5	0.718452	1516	2.15765	0.0311

. . .

Number of Observations: 1600

Number of Groups: 80

1. [10] Give a precise interpretation of the fixed effects regression coefficients in the model.
2. [10] Sketch the predicted trajectories for each treatment. Label the trajectories and indicate the heights of the curves at critical points.
3. [10] Interpret the estimated parameters for the random effects, including any implications for the variance of the response conditional on Weeks.

Question 78: (30 marks in 3 parts)

Data is obtained on the effects of two treatments (A and B). Data was recorded on symptoms weekly throughout 10 weeks of active treatment and then for 8 more weeks following termination of the treatment. A linear spline model is fit to the data and the following results were obtained [some numbers were rounded to facilitate arithmetic]:

```
sp <- function( x ) {  
  gsp( x,  
    knots = c( 10 ),      # 1 knots => 2 intervals  
    degree = c( 1 , 1 ) , # linear in each interval  
    smooth = c( 0 )       # continuous at the knot  
  )  
}
```

```
> fit <- lme( Symptom ~ sp(Weeks) * tx, random = ~ 1 + Weeks | id, data = d)  
> summary(fit)  
Linear mixed-effects model fit by REML  
Data: d  
      AIC      BIC    logLik  
15099.44 15153.18 -7539.719  
  
Random effects:  
  Formula: ~1 + Weeks | id  
Structure: General positive-definite, Log-Cholesky parametrization  
              StdDev    Corr  
(Intercept) 37.921194 (Intr)  
Weeks        2.395329 -0.218  
Residual     23.285420  
  
Fixed effects: Symptom ~ sp(Weeks) * tx  
              Value Std.Error   DF   t-value p-value  
(Intercept)    131.0   6.450470 1516   20.43957  0.0000  
sp(Weeks)D1(0)   -11.0   0.508022 1516  -22.08346  0.0000  
sp(Weeks)C(10).1    18.0   0.571104 1516   31.94835  0.0000  
txB              -2.5   9.122343   78   -0.28679  0.7750  
sp(Weeks)D1(0):txB    1.5   0.718452 1516    2.15765  0.0311  
sp(Weeks)C(10).1:txB -5.0   0.807664 1516   -5.99001  0.0000  
  
. . . .
```

Number of Observations: 1600

Number of Groups: 80

1. [10] Give a precise interpretation of the fixed effects regression coefficients in the model.
2. [10] Sketch the predicted trajectories for each treatment. Label the trajectories and indicate the heights of the curves at critical points.
3. [10] Interpret the estimated parameters for the random effects.

Question 79: (40 marks in 3 parts)

A group of 60 subjects diagnosed with schizophrenia were followed for six years. Each year, each subject's treating physician prescribed one of three types of drugs: Typical, Atypical or Clozapine, based on their clinical judgment. The data records the drug used in the previous year and the level of symptoms exhibited by the subject in a variable called `neg`. Note that this is the data set studied in Lab 2.

The main goal of this analysis is to attempt to assess the effectiveness of "Clozapine" compared with "Typical" drugs.

The output below shows abbreviated results for six models used for this purpose. Each model seems to suggest a different estimate for the 'effect' of "Clozapine" compared with "Typical" drugs.

```
head( Drugs, 8)
```

	age1	yr sill1	pos	neg	gen	total	year	drug	status	Sex	Subj
17	29.1	10	16	8	33	57	1	Typical	Married	F	F36
72	29.1	10	21	23	60	104	2	Atypical	Married	F	F36
127	29.1	10	9	7	20	36	3	Clozapine	Married	F	F36
182	29.1	10	8	9	25	42	4	Clozapine	Married	F	F36
237	29.1	10	8	7	25	40	5	Clozapine	Married	F	F36
292	29.1	10	12	7	26	45	6	Clozapine	Married	F	F36
38	36.4	16	12	8	29	49	1	Typical	Separated/Divorced	F	F41
93	36.4	16	13	10	31	54	2	Atypical	Separated/Divorced	F	F41

```
dim( Drugs)
```

```
[1] 318 11
```

```
Drugs$drug <- relevel( Drugs$drug, 'Typical')
fit1 <- lm( neg ~ drug, Drugs)
pr( fit1)
```

	numDF	denDF	F-value	p-value	
	3	315	694.4892	<.00001	
	Estimate	Std.Error	DF	t-value	p-value
(Intercept)	15.232704	0.481526	315	31.634242	<.00001
drugAtypical	0.045073	0.862500	315	0.052259	0.95836
drugClozapine	1.066146	0.809706	315	1.316707	0.18889

```
fit2 <- lme( neg ~ drug, Drugs, random = ~ 1 | Subj)
pr( fit2)
```

	numDF	denDF	F-value	p-value	
	3	263	148.3786	<.00001	
	Estimate	Std.Error	DF	t-value	p-value
(Intercept)	16.297836	0.788330	263	20.673872	<.00001
drugAtypical	-0.815021	0.614065	263	-1.327255	0.18558
drugClozapine	-2.115292	0.682395	263	-3.099806	0.00215

```
fit3 <- lme( neg ~ drug + cvar( drug, Subj), Drugs, random = ~ 1 | Subj)
pr( fit3)
```

	numDF	denDF	F-value	p-value
	5	50	99.01414	<.00001

	Estimate	Std.Error	DF	t-value	p-value
(Intercept)	14.119202	1.273724	263	11.084980	<.00001
drugAtypical	-0.896756	0.627306	263	-1.429535	0.15404
drugClozapine	-2.678858	0.715431	263	-3.744397	0.00022
cvar(drug, Subj)Atypical	2.616211	2.891131	50	0.904909	0.36985
cvar(drug, Subj)Clozapine	6.429352	2.347188	50	2.739172	0.00851

```
fit4 <- lme(neg ~ drug + cvar(drug, Subj) + year, Drugs, random = ~ 1 | Subj)
pr(fit4)
```

numDF	denDF	F-value	p-value
6	50	84.67485	<.00001

	Estimate	Std.Error	DF	t-value	p-value
(Intercept)	15.791551	1.359658	262	11.614352	<.00001
drugAtypical	0.117037	0.678516	262	0.172489	0.86319
drugClozapine	-1.036161	0.842019	262	-1.230567	0.21959
cvar(drug, Subj)Atypical	1.602418	2.902672	50	0.552049	0.58337
cvar(drug, Subj)Clozapine	4.786655	2.388817	50	2.003776	0.05053
year	-0.477814	0.135917	262	-3.515489	0.00052

```
fit5 <- lme(neg ~ dvar(drug, Subj) + cvar(drug, Subj) + year, Drugs, random = ~ 1 | Subj)
pr(fit5)
```

numDF	denDF	F-value	p-value
6	50	84.67485	<.00001

	Estimate	Std.Error	DF	t-value	p-value
(Intercept)	15.791551	1.359658	262	11.614352	<.00001
dvar(drug, Subj)Atypical	0.117037	0.678516	262	0.172489	0.86319
dvar(drug, Subj)Clozapine	-1.036161	0.842019	262	-1.230567	0.21959
cvar(drug, Subj)Atypical	1.719455	2.822255	50	0.609249	0.54512
cvar(drug, Subj)Clozapine	3.750494	2.235498	50	1.677700	0.09965
year	-0.477814	0.135917	262	-3.515489	0.00052

```
fit6 <- lme(neg ~ dvar(drug, Subj) + cvar(drug, Subj) + year,
            correlation = corAR1(form = ~ 1 | Subj), Drugs, random = ~ 1 | Subj)
pr(fit6)
```

numDF	denDF	F-value	p-value
6	50	84.33179	<.00001

	Estimate	Std.Error	DF	t-value	p-value
(Intercept)	15.815210	1.370365	262	11.540874	<.00001
dvar(drug, Subj)Atypical	0.241801	0.687552	262	0.351684	0.72536
dvar(drug, Subj)Clozapine	-0.957126	0.900737	262	-1.062602	0.28894
cvar(drug, Subj)Atypical	1.736376	2.806682	50	0.618658	0.53895
cvar(drug, Subj)Clozapine	3.702072	2.223157	50	1.665232	0.10212
year	-0.485736	0.150367	262	-3.230327	0.00139

1. [20] Almost every model produces a different estimate of the ‘effect’ of Clozapine compared with ‘Typical’ drugs. Consider the different estimates provided by each model and discuss plausible reasons for the differences or lack of differences. You only need to compare models sequentially, i.e. `fit1` with `fit2`, `fit2` with `fit3`, and so on.

2. [10] The models `fit4` and `fit5` produce different estimates for the coefficient associated with `cvar(drug, Subj)Clozapine`. Discuss the reason for and interpretation of this difference.
3. [10] Discuss the strengths and limitations of this kind of study to assess the effectiveness of a new drug, Clozapine.

Question 80: (40 marks in 8 parts)

Suppose you have data on a continuous variable, Y , and two variables X and Z that are dichotomous with values 0 and 1.

The following table shows the values of $\bar{Y}$ for all the combinations of values of X and Z :

	X=0	X=1
Z = 0	8	6
Z = 1	3	2

The number of observations in each cell is:

	X=0	X=1
Z = 0	30	60
Z = 1	120	20

1. [5 marks] Draw the Paik-Agresti diagram for this data including the line for the marginal effect of X .
[Hint: You can wait until having answered part 3 below before drawing in the marginal effect.]

2. [5 marks] What is(are) the value(s) of the ‘conditional effect(s) of X ’?

3. [5 marks] What is the value of the ‘marginal effect of X’?
4. [5 marks] If you were to fit the model $Y \sim X*Z$ in R, i.e. the model

$$\hat{Y} = \hat{\beta}_0 + X\hat{\beta}_1 + Z\hat{\beta}_2 + XZ\hat{\beta}_3$$

what would the values of the estimated regression coefficients be?

5. [5 marks] How would you get the ‘conditional effects of X’ from the regression coefficients? Express your answer in the form of a ‘hypothesis matrix’ multiplying the vector of fitted values, i.e. a matrix L so that your answer would have the form $L\hat{\beta}$.
6. [5 marks] What is the interpretation of the interaction coefficient, $\hat{\beta}_3$?

7. [5 marks] Why is the use of the word ‘effects’ problematic in this context? Can you think of a better word?

8. [10 marks] How would you get the marginal effect of X from the regression coefficients? Hint: Think of using a version of the chain rule in multivariate calculus?

Cumgrade: 1200