MATH 4939

Some sample questions for the mid-term test on February 12, 2025

Question 1: (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.

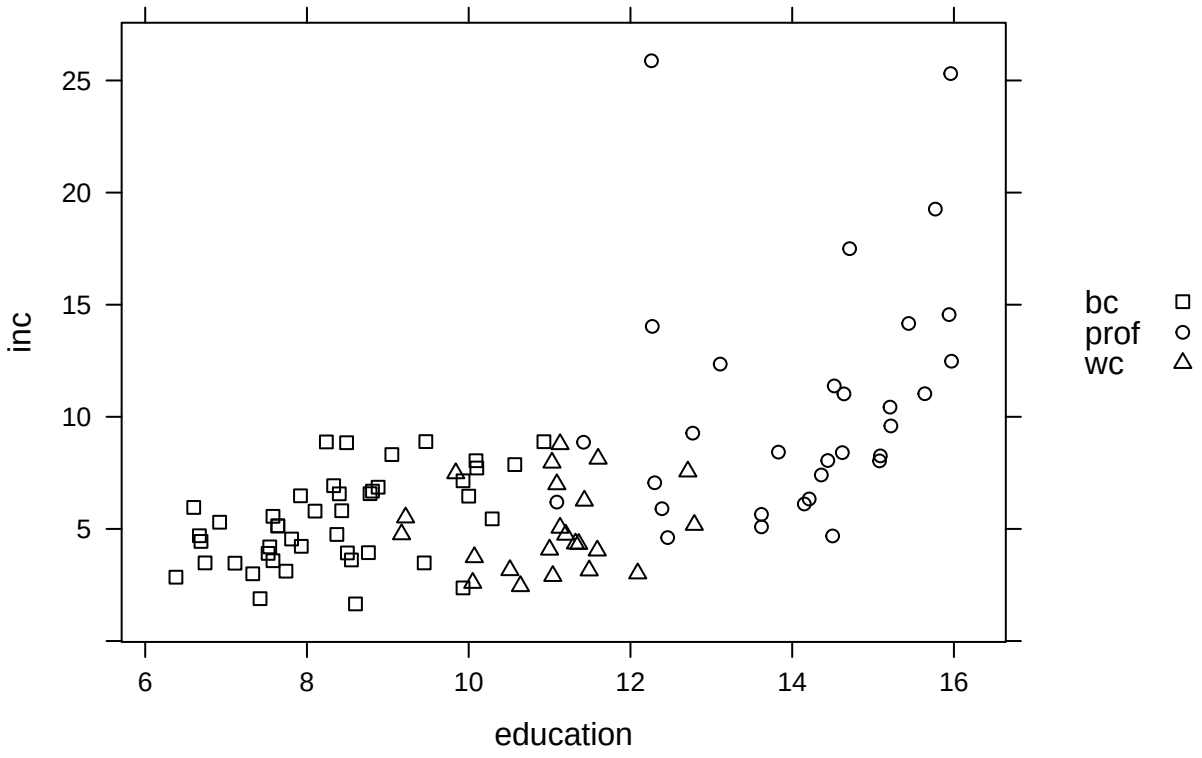
```
library(car, quietly = TRUE, warn = FALSE)
library(lattice)
head(Prestige)
```

```
                    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'))))
```



```r
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.

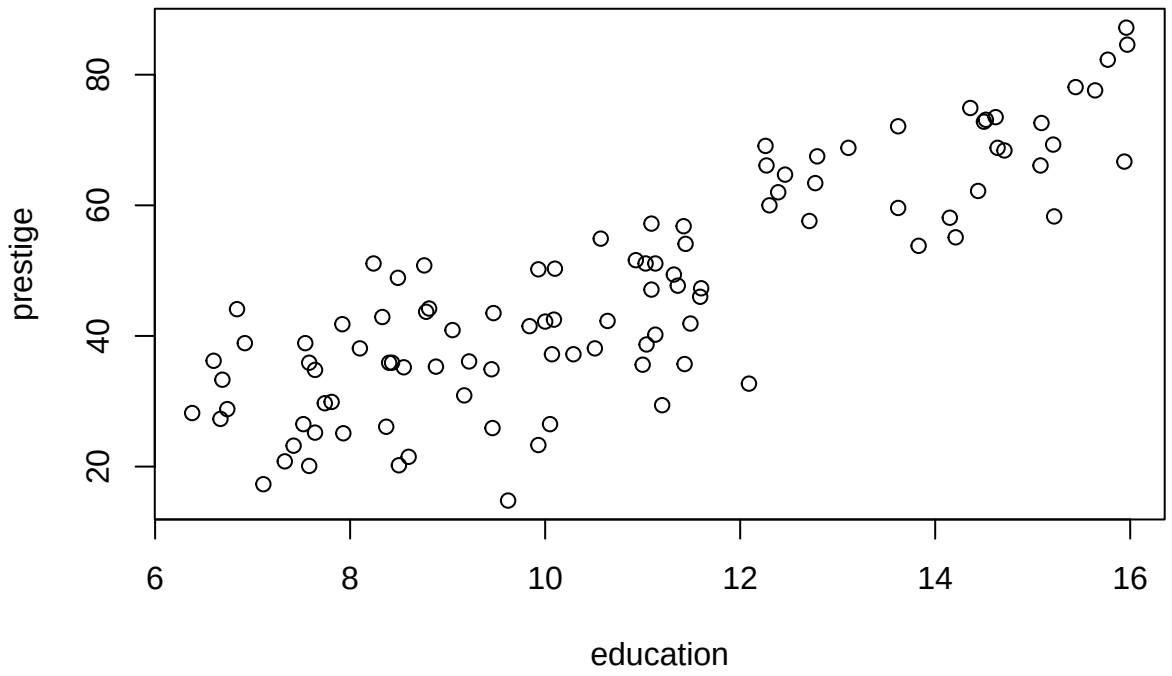
The negative coefficient for education says that predicted income is lower as education increases and the negative coefficient for 'typeprof:education' says that the change in income associated with additional education is lower for professional occupations than it is for blue collar occupations.

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 2: (20 marks in 3 parts)

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

```
library(car, quietly = TRUE, warn = FALSE)
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
```

a) [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.

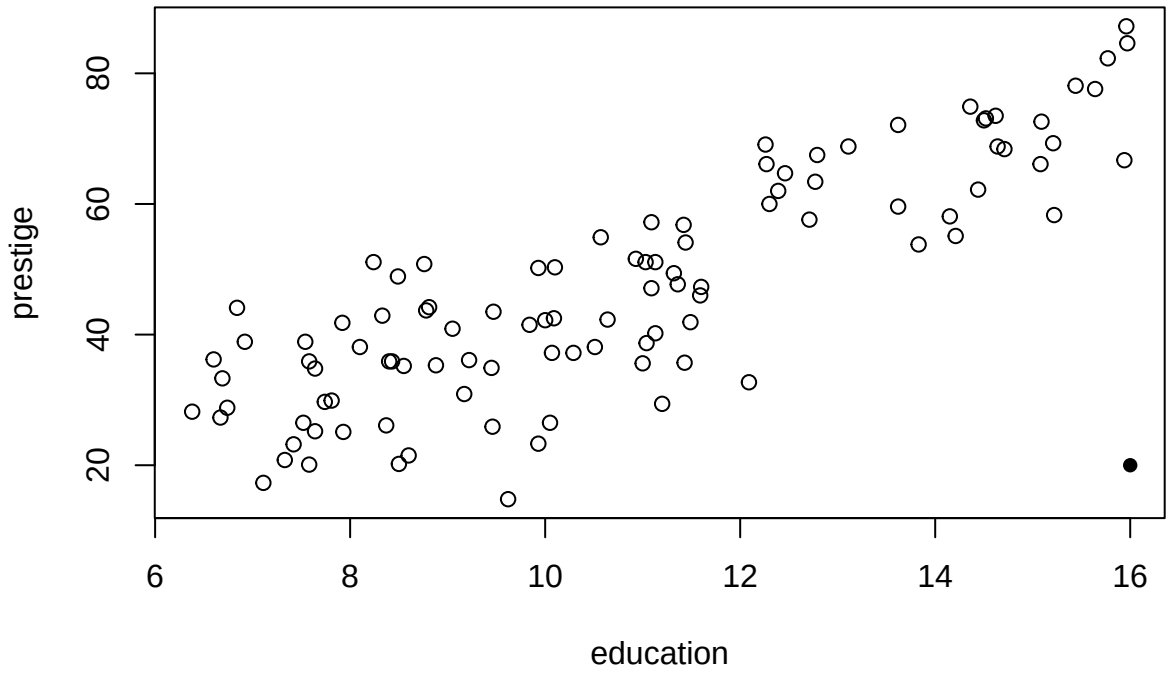
b) [5 marks] What is the your visual estimate of the correlation? Explain how you obtained it?

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 3: (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, quietly = TRUE, warn = FALSE)
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 4: (20 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 5: (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 6: (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 7: (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 8: (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 9: (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 10: (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:

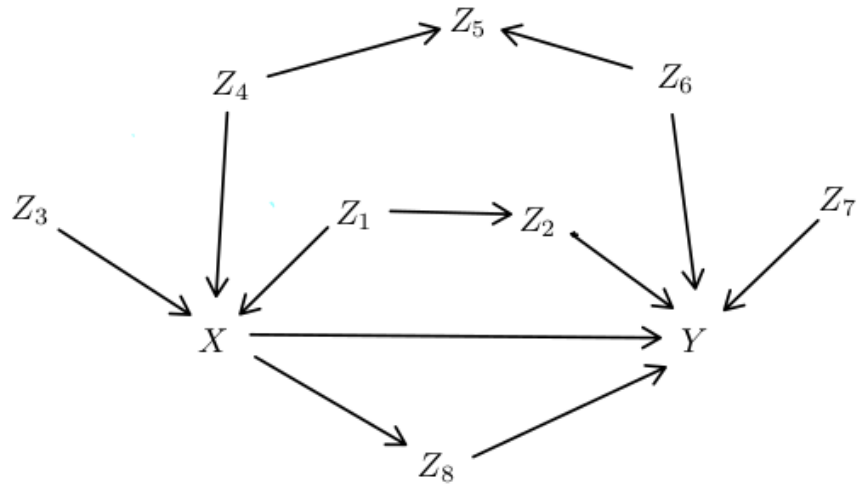
$$f(x_i;\theta) = \frac{1}{\theta} \exp -x/\theta \quad x_i, \theta > 0$$

.

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 11: (20 marks in 2 parts)



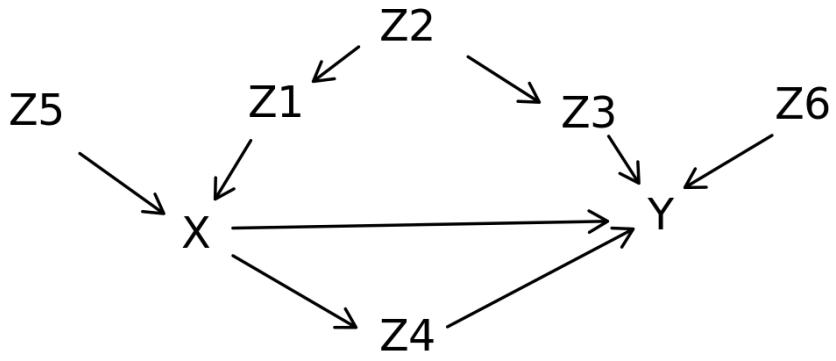
Consider the linear causal DAG above and the following models:

1. `Y ~ X`
2. `Y ~ X + Z1 + Z5`
3. `Y ~ X + Z1`
4. `Y ~ X + Z8`
5. `Y ~ X + Z2`
6. `Y ~ X + Z2 + Z6`
7. `Y ~ X + Z2 + Z8`
8. `Y ~ X + Z2 + Z7`
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] 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 12: (10 marks)



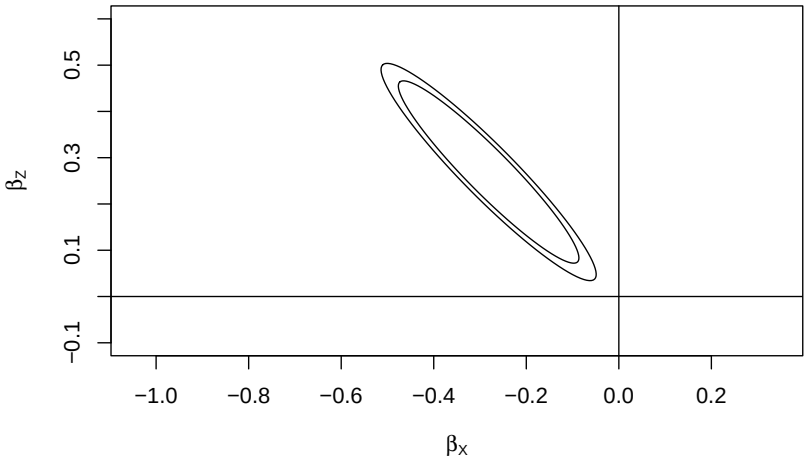
For the linear causal DAG above, identify one model that provides an unbiased estimate of the causal effect of `X` on `Y` that would yield a relatively small standard error for the estimated causal effect, and another model that also yields an unbiased estimate of the causal effect of `X` on `Y` but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

Question 13: (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 14: (10 marks)

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

Question 15: (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 16: (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 17: (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 18: (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 `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 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.