

MATH 4939 Sample Midterm Questions

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 XXX The exam is graded out of XXX, so there are 10 bonus marks.

If needed, continue your answer on this page.

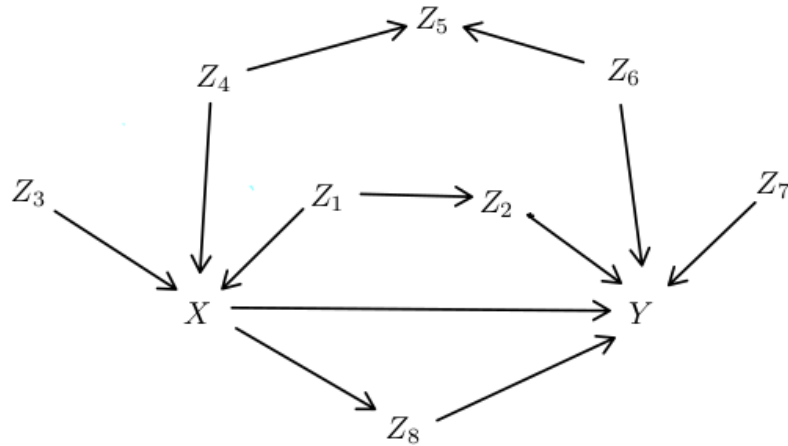
Question 1: (10 marks)

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 $\Sigma = \Gamma\Lambda\Gamma'$.

If needed, continue your answer on this page.

Question 2: (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. $Y \sim X + Z_2 + Z_7$

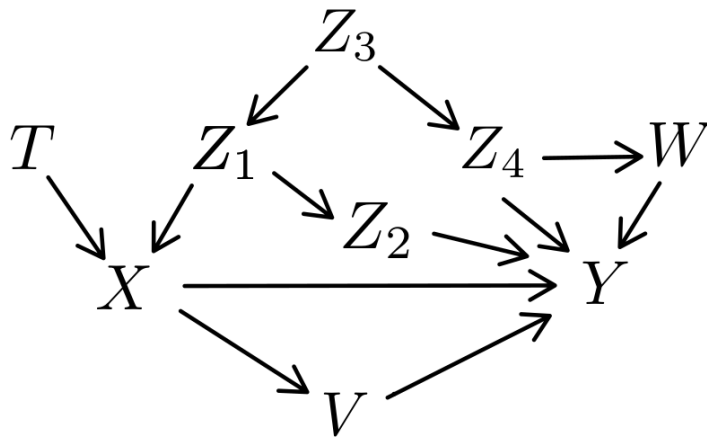
9. $Y \sim X + Z_1 + Z_5 + Z_6$

If needed, continue your answer on this page.

- b) [10 marks] Choose any two models that produce unbiased estimates of the causal effect of $\mathbf{X}$ and discuss in detail whether one can be determined to have a lower standard error for its estimate of β_X .

If needed, continue your answer on this page.

Question 3: (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. $Y \sim X + Z_2 + W$

If needed, continue your answer on this page.

- b) [10 marks] Choose any two models that produce unbiased estimates of the causal effect of $\mathbf{X}$ and discuss in detail whether one can be determined to have a lower standard error for its estimate of β_X .

If needed, continue your answer on this page.

Question 4: (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’)

If needed, continue your answer on this page.

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 0.45 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.

If needed, continue your answer on this page.

Question 6: (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.

If needed, continue your answer on this page.

Question 7: (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?

If needed, continue your answer on this page.

Question 8: (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 $X_{\cdot Z}$.

If needed, continue your answer on this page.

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 .

If needed, continue your answer on this page.

Question 9: (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.

If needed, continue your answer on this page.

- 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.

If needed, continue your answer on this page.

Question 10: (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’?

If needed, continue your answer on this page.

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 $L\hat{\beta}$.

If needed, continue your answer on this page.

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?

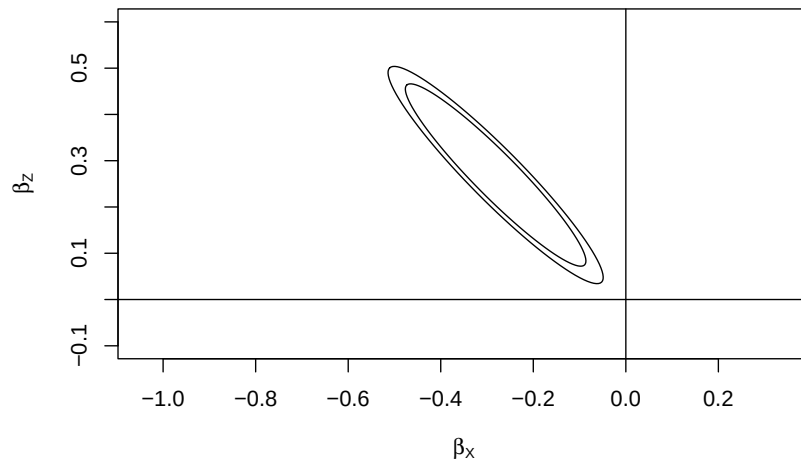
If needed, continue your answer on this page.

Question 11: (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) $H_0 : \beta_X - \beta_Z = 0$

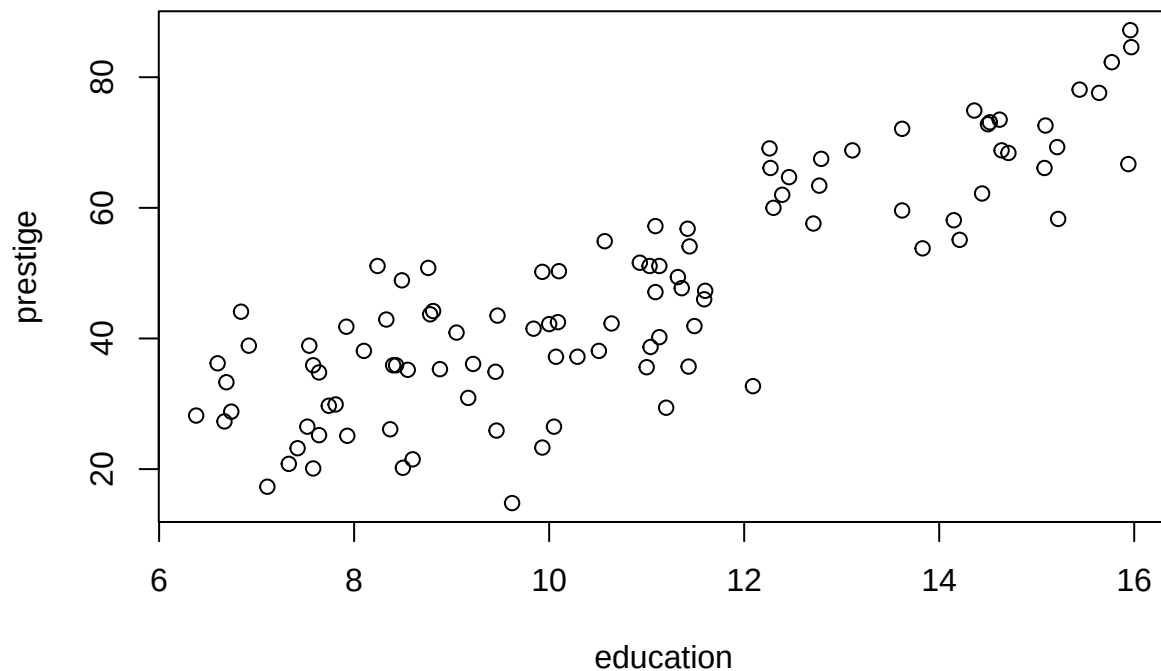
If needed, continue your answer on this page.

If needed, continue your answer on this page.

Question 12: (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
```

- 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.

If needed, continue your answer on this page.

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.

If needed, continue your answer on this page.

Question 13: (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.

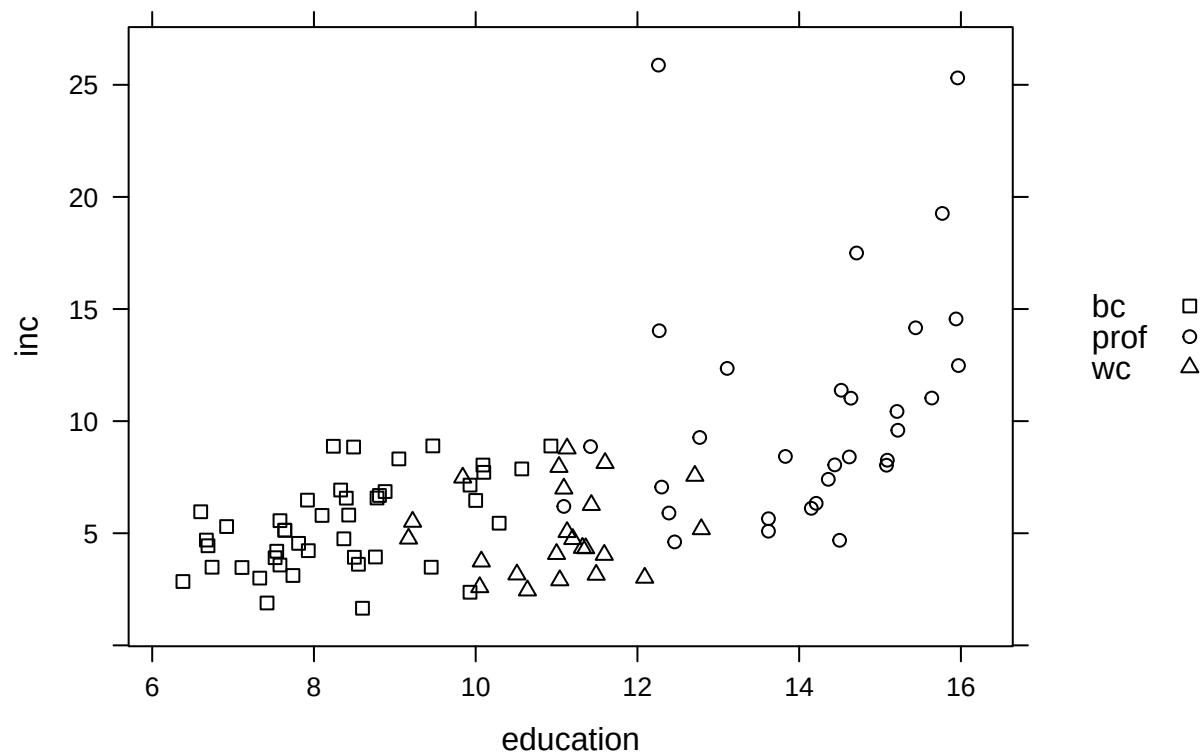
```
library(car)
library(lattice)
head(Prestige)
```

```
##               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.

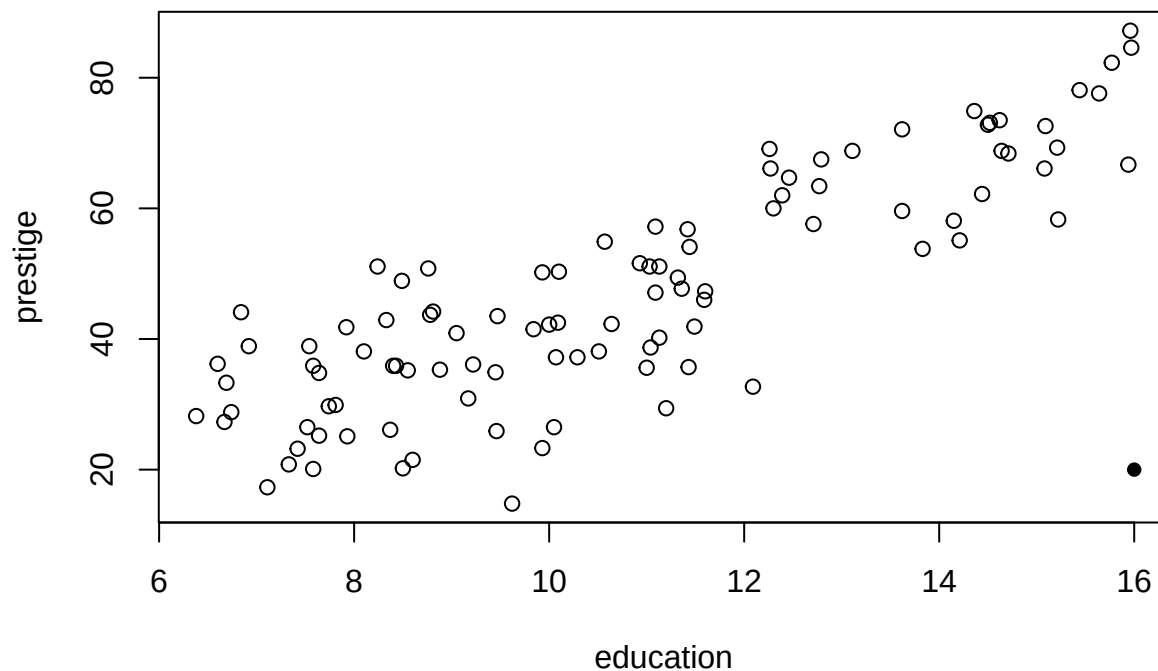
Write a short essay explaining whether and how you should change your model to take these issues into account. If you believe that the model has a reasonable interpretation, write an explanation in response to your client's concerns. Write the explanation in terms your client would understand. Assume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

If needed, continue your answer on this page.

Question 14: (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.