MATH 4939

Some sample questions for Quiz 3 on January 29, 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.

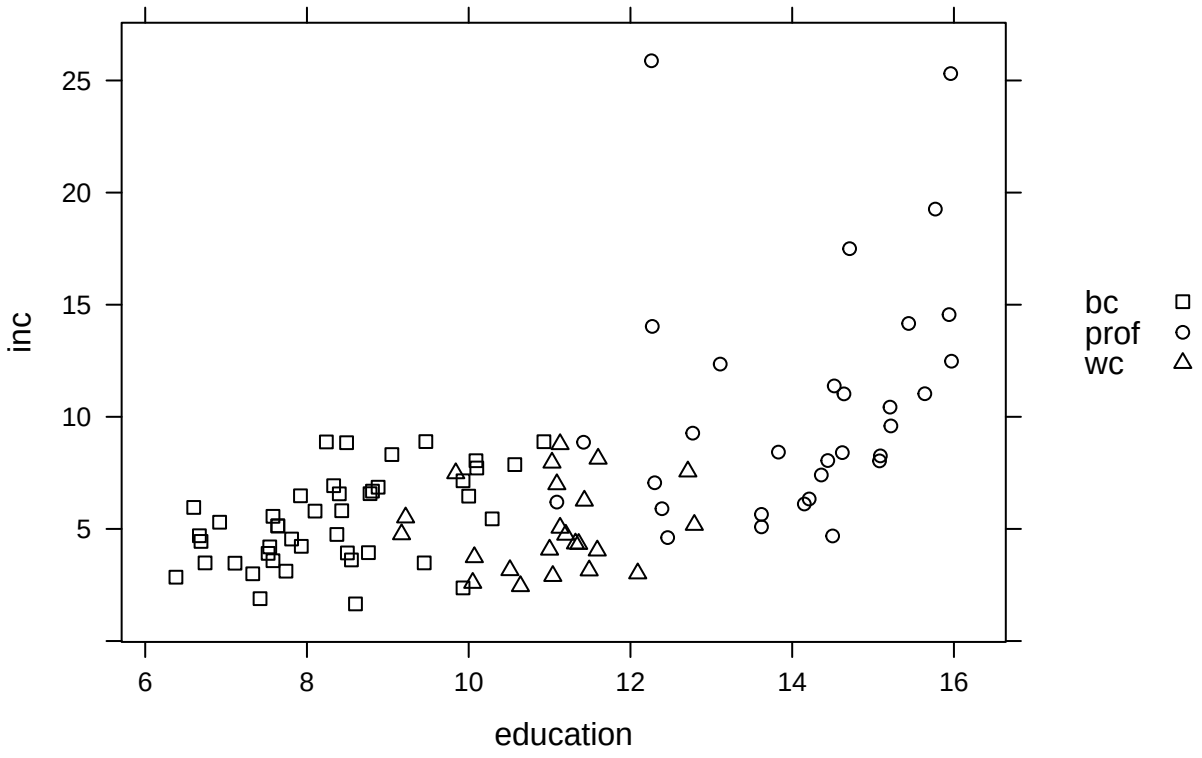
```
library(car, quietly = TRUE, warn = FALSE)
library(lattice)
head(Prestige)
```

```
                    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.

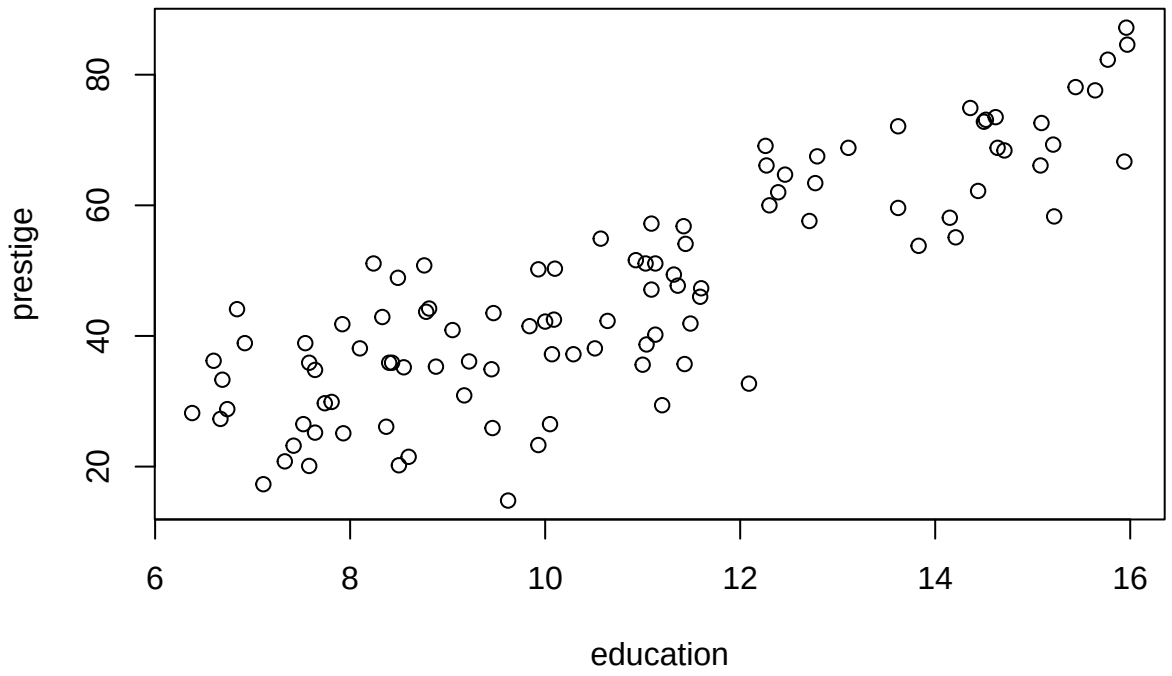
The negative coefficient for education says that predicted income is lower as education increases and the negative coefficient for 'typeprof:education' says that the change in income associated with additional education is lower for professional occupations than it is for blue collar occupations.

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.

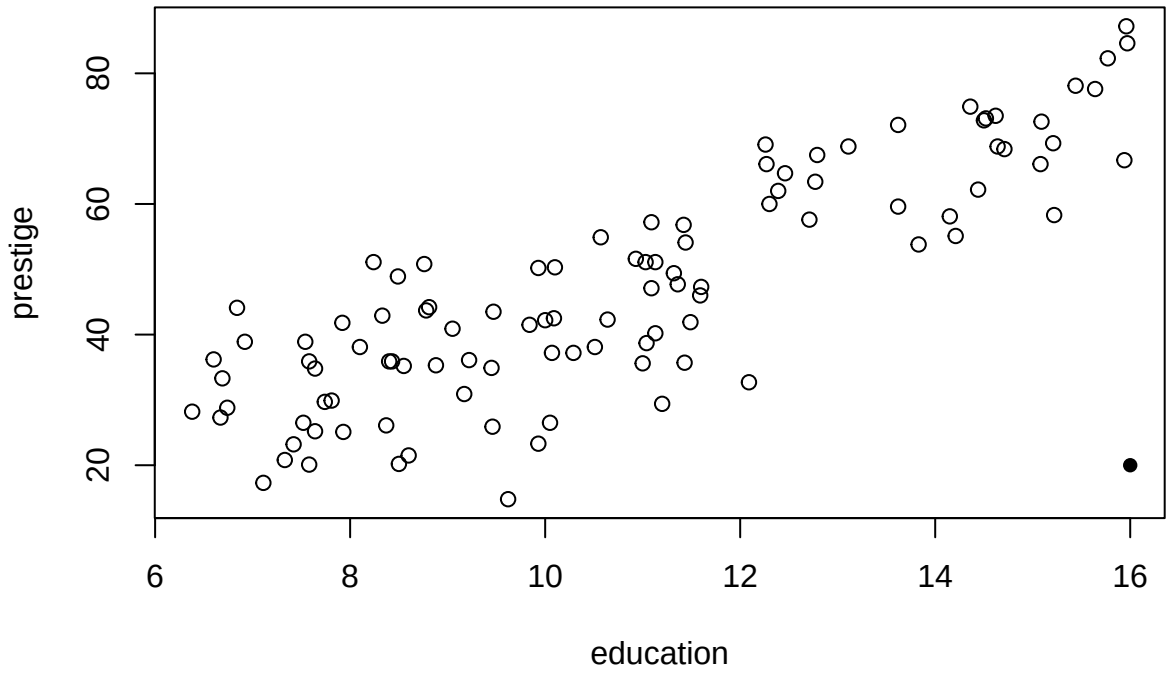
b) [5 marks] What is the your visual estimate of the correlation? Explain how you obtained it?

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.

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`