MATH 4939 Mid-term Test

Duration: 50 minutes

February 14, 2024 9:30 am

Instructions:

- Aids allowed: non-programmable calculator.
- Answer all questions in this booklet. You can use both the front and the back of each page.
- The marks sum to 80.

Question 1: (20 points in 2 parts)

The following analysis uses the familiar 'Vocab' data set, consisting or vocabulary scores obtained in samples of U.S. residents during the years 1974 to 2016, categorized by binary gender (Male and Female) and education (in years).

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

```
library(car)
```

```
## Loading required package: carData
```

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

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

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

```
##                           Estimate Std. Error    t value      Pr(>|t|)
## (Intercept)                  170.1  7.1209832 23.8861214 6.092696e-125
## sexMale                      -21.0 10.4757949 -2.0014787  4.534971e-02
## education                     33.2  0.5223064 63.5065527  0.000000e+00
## decade                        19.4  4.9964652  3.8841883  1.028900e-04
## sexMale:education              0.4  0.7622630  0.5438925  5.865195e-01
## sexMale:decade               -11.6  7.2237215 -1.6084923  1.077378e-01
## education:decade              -2.7  0.3798654 -7.1622069  8.119444e-13
## sexMale:education:decade       1.3  0.5432446  2.4382176  1.476558e-02
```

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

	Estimate
(Intercept)	170.1
sexMale	-21.0
education	33.2
decade	19.4
sexMale:education	0.4
sexMale:decade	-11.6
education:decade	-2.7
sexMale:education:decade	1.3

Decade = 0

$$\frac{\partial E(Y)}{\partial \text{sexMale}} = -21 + 0.4 \times \text{Educ} - 11.6 \times \text{decade} + 1.3 \times \text{Educ} \times \text{decade}$$

So $\frac{\partial E(Y)}{\partial \text{sexMale}} \Big|_{\text{Decade}=0,\ \text{Educ}=21} = -21 + 0.4 \times 20$

$= -21 + 8 = -13$

But we are asked for 'Female - Male'
so the answer is $-(-13) = 13$

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

We need the gender gap in the year 1990 to see whether it's + or −,
Using the same formula.

Gap (Male - Female)

$$= -21 + 0.4 \times \text{Educ} - 11.6 \times \text{decade} + 1.3 \times \text{Educ} \times \text{decade}$$

$$= -21 + 0.4 \times 20 - 11.6 \times (-1) + 1.3 \times 20 \times (-1)$$

$$= -21 + 8 + 11.6 - 26$$

$$= -27.4$$

How is it changing? Take derivative w.r.t. time:

$$\frac{\partial^2 E(Y)}{\partial \text{SexMale}\, \partial \text{Decade}} = -11.6 + 1.3 \times \text{Educ} = -11.6 + 1.3 \times 20$$

$$= -11.6 + 26 = 14.4$$

Cont'd

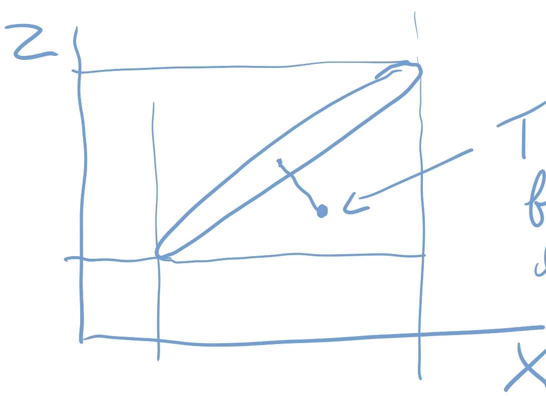
So it is getting less negative

$\therefore$ narrowing.

Question 2: (10 points)

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

Leverage is monotone in Mahalanobis distance, Δ. A point in X, Z space would have low leverage if it has relatively small Δ w.r.t. X & Z separately. It can have a large Δ w.r.t (X, Z) together if X & Z are highly correlated, e.g.



This point has $\Delta < 1$ for X & Z separately but $\Delta \gtrsim 3$ for X, Z together.

You may continue your answer on this side.

Question 3: (10 points)

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

The coefficients have been rounded for ease of calculation.

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

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

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

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

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

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

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

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

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

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

The predicted income for professional occupation is

$$36 + 58 - 8\,\text{educ} + 1\,\text{educ}^2 - 5 \times \text{education}$$

and for a blue collar occupation

$$36 - 8\,\text{educ} + 1 \times \text{educ}^2$$

An extra year of education is associated with a change of

$-8 + 2 \times \text{educ} - 5$ for "prof"

and $-8 + 2\,\text{educ}$ for bc

You may continue your answer on this side.

but if typical educ for bc is 10 and 15 for "prof", then the value of an extra year would be

$-8 \times 2 \times 10 - 5 = 7$ for 'bc'

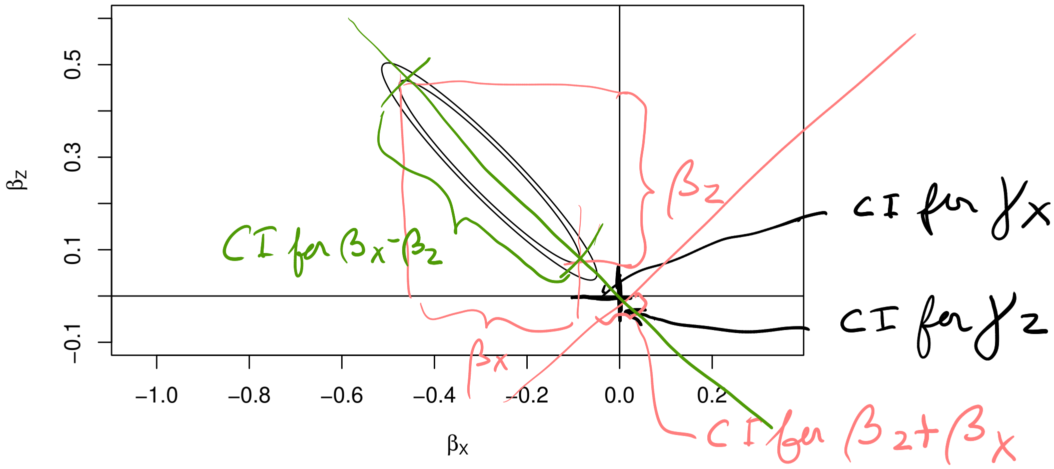
and $-8 \times 2 \times 15 = 22$ for 'prof.'

Question 4: (20 points)

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

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

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



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

a) $H_0 : \beta_X = \beta_Z = 0$ Reject : $(0,0) \notin$ Outer ellipse

b) $H_0 : \beta_X = 0$ Reject . $0 \notin$ CI for β_X

c) $H_0 : \beta_Z = 0$ Reject $0 \notin$ CI for β_Z

d) $H_0 : \gamma_X = 0$ Accept $0 \in$ CI for γ_X

e) $H_0 : \gamma_Z = 0$ accept $0 \in$ CI for γ_Z

f) $H_0 : \beta_X = \beta_Z$ Reject $0 \notin$ CI for $\beta_X - \beta_Z$

g) $H_0 : \beta_X + \beta_Z = 0$ accept $0 \in$ CI for $\beta_X + \beta_Z$

You may continue your answer on this side.

Question 5: (10 points)

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

The above is a counterexample showing how a model that is significant for β_x, β_z and (β_x, β_z) jointly may not show a significant relationship with X or Z individually, so that a forward stepwise procedure would stop with a model containing only the intercept.

Other issues:

- If a model includes interactions and polynomial powers, stepwise selection will not respect the PCM.
- If the purpose is causal inference, stepwise selection is likely to include mediators and fail to include needed confounders.

You may continue your answer on this side.

Question 6: (10 points)

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

Consider the following two models:

1)

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

2)

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

Show that these models are equivalent.

The models are equivalent because each column of each X matrix is a linear combination of columns of the other matrix.

- The intercept appears in both models
- For the second column we have

$$\underbrace{X}_{\text{model 1}} = \underbrace{X_g + X_d}_{\text{model 2}}$$

$$\underbrace{X_d}_{\text{model 2}} = \underbrace{X - X_g}_{\text{model 1}}$$

- The third column is the same in both model

$\therefore$ models are equivalent

You may continue your answer on this side.