STA2101 Assignment 4*

Ahmed Nawaz Amanullah †

February 2021

1

1.1 Problem

In the United States, admission to university is based partly on high school marks and recommendations, and partly on applicants' performance on a standardized multiple choice test called the Scholastic Aptitude Test (SAT). The SAT has two sub-tests, Verbal and Math. A university administrator selected a random sample of 200 applicants, and recorded the Verbal SAT, the Math SAT and first-year university Grade Point Average (GPA) for each student. The data are available here. We seek to predict GPA from two test scores. Throughout, please use the usual $\alpha = 0.05$ significance level.

1.1.1 (a)

First, fit a model using just the Math score as a predictor. "Fit" means estimate the model parameters. Does there appear to be a relationship between Math score and grade point average?

i. Answer Yes or No.

†with help from the Overleaf team

ii. Fill in the blank. Students who did better on the Math test tended to have _____ first-year grade point average.

iii. Do you reject $H_0 : \beta_1 = 0$?

iv. Are the results statistically significant? Answer Yes or No.

v. What is the p-value? The answer can be found in *two* places on your printout.

vi. What proportion of the variation in first-year grade point average is explained by the score on the SAT Math test? The answer is a number from your printout.

vii. Give a predicted first-year grade point average for a student who got 700 on the Math SAT. The answer is a number you could get with a calculator from your printout.

1.1.2 (b)

Now fit a model with both the Math and Verbal sub-tests.

i. Give the test statistic, the degrees of freedom and the p-value for each of the following null hypotheses. The answers are numbers from your printout.

   A. $H_0 : \beta_1 = \beta_2 = 0$
   B. $H_0 : \beta_1 = 0$
   C. $H_0 : \beta_2 = 0$
   D. $H_0 : \beta_0 = 0$

ii. Controlling for Math score, is Verbal score related to first-year grade point average?

   A. Give the value of the test statistic. The answer is a number from your printout.
   B. Give the p-value. The answer is a number from your printout.

C. Do you reject the null hypothesis?

D. Are the results statistically significant? Answer Yes or No.

E. In plain, non-statistical language, what do you conclude? The answer is something about test scores and grade point average.

iii. Allowing for Verbal score, is Math score related to first-year grade point average?

A. Give the value of the test statistic. The answer is a number from your printout.

B. Give the p-value. The answer is a number from your printout.

C. Do you reject the null hypothesis?

D. Are the results statistically significant? Answer Yes or No.

E. In plain, non-statistical language, what do you conclude? The answer is something about test scores and grade point average.

iv. Give a predicted first-year grade point average for a student who got 650 on the Verbal and 700 on the Math SAT.

v. Let's do one more test. We want to know whether GPA increases faster as a function of the Verbal SAT, or the Math SAT. That is, we want to compare the regression coefficients, testing $H_0 : \beta_1 = \beta_2$.

A. Express the null hypothesis in matrix form as $\mathbf{L}\boldsymbol{\beta} = \mathbf{h}$.

B. Carry out an F test. Feel free to use my **ftest** function (from lecture) if you wish.

C. State your conclusion in plain, non-technical language. It's something about first-year grade point average.

1.2 Solution

1.2.1 (a)

Disclosed below is the code implemented in R to fit the required model and generate the required outputs for answering the questions.

```
# Question 1
# set working directory to the folder containing the data
wdir <- "/Users/amanull9/Documents/UofT MSc in Stats/Student/"
wdir <- paste(wdir, "Fall/STA2101 (Applied Stats I)/", sep = "")
wdir <- paste(wdir, "Assignments/Assn4", sep = "")
setwd(wdir)
# read data in
uniapp_data <- read.csv(file="uniapplicants.csv",header=T, sep=",")

# part (a)
# Ok, now fit a linear reg model with GPA as response and Math as
#       as explanatory
lm1 <- lm(GPA ~ MATH, data = uniapp_data)
# get summary
summary(lm1)
# predict the GPA forecasted for Math score of 700
gpa_hat <- predict(lm1, data.frame(MATH=700))
# print forecast
sprintf("Predicted GPA for SAT Math score of 700: %.2f", gpa_hat)
```

Looking at our **summary** output in the R console, there appears to be relationship between Math score and GPA.

i. Yes

ii. higher

iii. Yes

iv. Yes

v. 0.005885

vi. 0.03768 (i.e. Multiple R-squared, which measures amount of variation of response variable explained by the predictor variables).

vii. 2.72.

1.2.2 (b)

The code written in R for this section is disclosed below,

```
# part (b) (i)
# Fit full model (with both SAT scores)
lm2 <- lm(GPA ~ VERBAL+MATH, data = uniapp_data)
# get summary
summary(lm2)

# part (b) (ii)
# Control for Math, is Verbal score related to GPA?
# reduced model and full model already fit - use anova
anova(lm1, lm2)

# part (b) (iii)
# Allow for Verbal, is Math related to GPA?
# fit reduced model with just verbal (control group)
lm3 <- lm(GPA ~ VERBAL, data = uniapp_data)
# use anova to do the F test
anova(lm3, lm2)

# part (b) (iv)
# predict GPA for Verbal and Math scores of 650 and 700 respectively
gpa_hat <- predict(lm2, data.frame(VERBAL=650,MATH=700))
# print result
sprintf("Predicted GPA for Verbal 650 and Math 700: %.2f", gpa_hat)

# part (b) (v)
# retrieve ftest function written by Prof Brunner
source("http://www.utstat.utoronto.ca/~brunner/Rfunctions/ftest.txt")
# run ftest
ftest(lm2, rbind(c(0, 1, -1)))
```

By looking at the output in R console, we retrieve the following responses to the questions.

i. A. $F = 12.93$, $df = 2$ and 197, p-value $= 5.305e-06$

   B. $t = 4.178$, $df = 197$, p-value $= 4.41e-05$

C. $t = 1.636$, $df = 197$, p-value $= 0.103$

D. $t = 1.378$, $df = 197$, p-value $= 0.170$

ii. A. $t = 4.178$ or $F = 17.458$

B. p-value $= 4.41e - 05$

C. Yes

D. Yes

E. There is evidence to suggest that, controlling for Math score, higher Verbal score is associated with higher GPA (i.e. positive relationship)

iii. A. $t = 1.636$ or $F = 2.6779$

B. p-value $= 0.103$

C. No

D. No

E. There is insufficient evidence to suggest that, allowing for Verbal score, there is any relationship between Math score and GPA

iv. 2.81

v. A. $\begin{pmatrix} 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 0 \end{pmatrix}$

B. $F = 1.9909873$, $df = 1$ and 197, p-value $= 0.1598149$

C. There is insufficient evidence to suggest that GPA increases faster as a function for either Verbal or Math scores.

2

2.1 Problem

Let $X_1, ..., X_n$ be a random sample from a distribution with density

$$f(x) = \frac{\theta e^{\theta(x-\mu)}}{\left(1 + e^{\theta(x-\mu)}\right)^2} \tag{1}$$

for x real, where $-\infty < \mu < \infty$ and $\theta > 0$. Here are some numerical data:

4.82	3.66	4.39	1.66	3.80	4.69	1.73	4.50	9.29	4.05	4.50	-0.64	1.40
4.18	2.70	5.65	5.47	0.55	4.64	1.19	2.28	7.16	4.80	3.19	2.33	2.57
2.31	0.35	2.81	2.35	2.52	3.44	2.71	-1.43	7.61	0.93	2.52	6.86	6.14
4.37	3.79	5.04	4.50	1.92	3.25	-0.06	2.81	3.09	2.95	3.69		

You can read the data from
http://www.utstat.toronto.edu/~brunner/data/legal/mystery.data.txt.

2.1.1 (a)

Find the maximum likelihood estimates of μ and θ.

2.1.2 (b)

Obtain an approximate 95% confidence interval for θ.

2.1.3 (c)

Test $H_0 : \mu = 0$ at the $\alpha = 0.05$ significance level with a large-sample Z test.

2.2 Solution

2.2.1 (a)

Let's first derive the likelihood and log-likelihood functions,

$$L(\mu, \theta) = \prod_{i=1}^{n} f(x_i; \mu, \theta) \tag{2}$$

$$L(\mu, \theta) = \prod_{i=1}^{n} \frac{\theta e^{\theta(x_i-\mu)}}{\left(1 + e^{\theta(x_i-\mu)}\right)^2} \tag{3}$$

$$L(\mu, \theta) = \frac{\theta^n e^{\theta \sum_{i=1}^{n}(x_i-\mu)}}{\left(\prod_{i=1}^{n}\left(1 + e^{\theta(x_i-\mu)}\right)\right)^2} \tag{4}$$

$$l(\mu, \theta) = \log L(\mu, \theta) \tag{5}$$

$$l(\mu, \theta) = \log\left(\frac{\theta^n e^{\theta \sum_{i=1}^{n}(x_i-\mu)}}{\left(\prod_{i=1}^{n}\left(1 + e^{\theta(x_i-\mu)}\right)\right)^2}\right) \tag{6}$$

$$l(\mu, \theta) = n \log \theta + \theta \sum_{i=1}^{n} (x_i - \mu) - 2 \log \left(\prod_{i=1}^{n} \left(1 + e^{\theta(x_i - \mu)}\right) \right) \tag{7}$$

$$l(\mu, \theta) = n \log \theta + \theta \sum_{i=1}^{n} (x_i - \mu) - 2 \sum_{i=1}^{n} \log \left(1 + e^{\theta(x_i - \mu)}\right) \tag{8}$$

At a cursory glance, differentiating and solving to obtain the analytical maximum likelihood estimates for μ and θ appears to be very difficult. So, we will try to obtain them numerically via R using the **nlm** function. This will also give us the Hessian, which will prove useful to obtain the estimated asymptotic covariance matrix for the latter parts of this question.

Additionally, we would also want a good starting point for our numerical search. Can we use method of moments here? Let's check. First, we want expressions for population moments,

$$E(X) = \int_{-\infty}^{\infty} x \frac{\theta e^{\theta(x-\mu)}}{\left(1 + e^{\theta(x-\mu)}\right)^2} \, dx \tag{9}$$

Hmm... seems like a tough integral. Can we do it? Try substitution method first, with substitution $y = e^{\theta(x-\mu)}$. If we use this, note that $dy = \theta e^{\theta(x-\mu)} dx$ and $x = \mu + \frac{1}{\theta} \log y$. This gives us an integral of

$$E(X) = \int_{0}^{\infty} \frac{\mu + \frac{1}{\theta} \log y}{\left(1 + y\right)^2} \, dy \tag{10}$$

Hmm... there seems to be no openly obvious way to solve this. To begin, let's try integration by parts.

$$E(X) = \mu \int_{0}^{\infty} \frac{1}{(1+y)^2} \, dy + \frac{1}{\theta} \int_{0}^{\infty} \frac{\log y}{\left(1 + y\right)^2} \, dy \tag{11}$$

$$E(X) = \mu \left(-\frac{1}{(1+y)} \Big|_{0}^{\infty} \right) + \frac{1}{\theta} \int_{0}^{\infty} \frac{\log y}{\left(1 + y\right)^2} \, dy \tag{12}$$

$$E(X) = \mu + \frac{1}{\theta} \left(\left(-\frac{\log y}{1+y} + \int \frac{1}{y(1+y)} \, dy \right) \Big|_{0}^{\infty} \right) \tag{13}$$

To integrate $\int \frac{1}{y(1+y)} \, dy$, we need partial fractions,

$$\frac{1}{y(1+y)} = \frac{A}{y} + \frac{B}{1+y} \tag{14}$$

$$\frac{1}{y(1+y)} = \frac{A(1+y)+By}{y(1+y)} \tag{15}$$

$$\frac{1}{y(1+y)} = \frac{A+(A+B)y}{y(1+y)} \tag{16}$$

By inspection we note that $A = 1$ and $B = -1$. So, returning to where we were,

$$E(X) = \mu + \frac{1}{\theta}\left(\left(-\frac{\log y}{1+y} + \int\left(\frac{1}{y} - \frac{1}{1+y}\right)\,dy\right)\Bigg|_0^\infty\right) \tag{17}$$

$$E(X) = \mu + \frac{1}{\theta}\left(\left(-\frac{\log y}{1+y} + \log y - \log(1+y)\right)\Bigg|_0^\infty\right) \tag{18}$$

$$E(X) = \mu + \frac{1}{\theta}\left(\left(\left(1 - \frac{1}{1+y}\right)\log y - \log(1+y)\right)\Bigg|_0^\infty\right) \tag{19}$$

$$E(X) = \mu + \frac{1}{\theta}\left(\left(\frac{y}{1+y}\log y - \log(1+y)\right)\Bigg|_0^\infty\right) \tag{20}$$

$$E(X) = \mu + \frac{1}{\theta}\left(\left(\frac{y\log y - (1+y)\log(1+y)}{1+y}\right)\Bigg|_0^\infty\right) \tag{21}$$

$$E(X) = \mu + \frac{1}{\theta}\left(\left(\frac{y\log\left(\frac{y}{1+y}\right) - \log(1+y)}{1+y}\right)\Bigg|_0^\infty\right) \tag{22}$$

Now, this is where it seems it may get messy. Try to evaluate the obtained anti-derivative at the upper limit?

$$\lim_{y\to\infty}\left(\frac{y\log\left(\frac{y}{1+y}\right) - \log(1+y)}{1+y}\right) \tag{23}$$

$$= \lim_{y\to\infty}\left(\frac{y}{1+y}\log\left(\frac{y}{1+y}\right) - \frac{\log(1+y)}{1+y}\right) \tag{24}$$

$$= \lim_{y\to\infty}\left(\frac{1}{1+\frac{1}{y}}\log\left(\frac{1}{1+\frac{1}{y}}\right) - \frac{\log(1+y)}{1+y}\right) \tag{25}$$

$$= \frac{1}{1+0}\log\left(\frac{1}{1+0}\right) - \lim_{y\to\infty}\frac{\log(1+y)}{1+y} \tag{26}$$

$$= 0 - \lim_{y\to\infty}\frac{\frac{1}{1+y}}{1} = 0 - 0 = 0 \tag{27}$$

So, it seems we could evaluate the anti-derivative at the upper limit, getting a value of 0 (note that we applied L'Hopital's Rule at the final step above). Now, the lower limit, haha (pass me the cyanide pill plz),

$$\left.\frac{y\log\left(\frac{y}{1+y}\right) - \log(1+y)}{1+y}\right|_{y=0} \tag{28}$$

Now, $\log(1+y) = 0$ and $1+y = 1$ at $y = 0$ clearly. So, let's focus on the term $y\log\left(\frac{y}{1+y}\right)$,

$$\left. y\log\left(\frac{y}{1+y}\right)\right|_{y=0} \tag{29}$$

$$= \lim_{y\to 0}\frac{\log\left(\frac{y}{1+y}\right)}{\frac{1}{y}} \tag{30}$$

$$= \lim_{y\to 0}\frac{\frac{1}{y} - \frac{1}{1+y}}{-\frac{1}{y^2}} \tag{31}$$

$$= \lim_{y\to 0}\frac{\frac{1}{y(1+y)}}{-\frac{1}{y^2}} \tag{32}$$

$$= \lim_{y\to 0} -\frac{y^2}{y(1+y)} \tag{33}$$

$$= \lim_{y\to 0} -\frac{y}{1+y} = 0 \tag{34}$$

So $y\log\left(\frac{y}{1+y}\right) = 0$ at $y = 0$. Thus,

$$\left.\frac{y\log\left(\frac{y}{1+y}\right) - \log(1+y)}{1+y}\right|_{y=0} = \frac{0-0}{1} = 0 \tag{35}$$

So, going back to our integral,

$$E(X) = \mu + \frac{1}{\theta}\left(\left(\frac{y\log\left(\frac{y}{1+y}\right) - \log(1+y)}{1+y}\right)\Bigg|_0^\infty\right) \tag{36}$$

$$E(X) = \mu + \frac{1}{\theta}(0-0) = \mu \tag{37}$$

Ok, to complete Method of Moments, we want to get $E(X^2)$ too (yo, where's that cyanide pill?),

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \frac{\theta e^{\theta(x-\mu)}}{\left(1+e^{\theta(x-\mu)}\right)^2}\,dx \tag{38}$$

Let's make use of the substitution that worked oh so well earlier, $y = e^{\theta(x-\mu)}$,

$$E(X) = \int_0^\infty \frac{\left(\mu + \frac{1}{\theta}\log y\right)^2}{(1+y)^2}\,dy \tag{39}$$

$$= \int_0^\infty \frac{\mu^2 + 2\frac{\mu}{\theta}\log y + \frac{1}{\theta^2}(\log y)^2}{(1+y)^2}\,dy \tag{40}$$

$$= \mu^2\int_0^\infty \frac{1}{(1+y)^2}\,dy + 2\frac{\mu}{\theta}\int_0^\infty \frac{\log y}{(1+y)^2}\,dy + \frac{1}{\theta^2}\int_0^\infty \frac{(\log y)^2}{(1+y)^2}\,dy \tag{41}$$

$$= \mu^2(1) + 2\frac{\mu}{\theta}(0) + \frac{1}{\theta^2}\int_0^\infty \frac{(\log y)^2}{(1+y)^2}\,dy \tag{42}$$

$$= \mu^2 + \frac{1}{\theta^2}\int_0^\infty \frac{(\log y)^2}{(1+y)^2}\,dy \tag{43}$$

So, we need to see if we can evaluate,

$$\int_0^\infty \frac{(\log y)^2}{(1+y)^2}\,dy \tag{44}$$

Hmm... dunno if we can do this chief, let's try integration by parts to start,

$$\int_0^\infty \frac{(\log y)^2}{(1+y)^2}\,dy \tag{45}$$

$$= \left(\left(\frac{y}{1+y}\log y - \log(1+y)\right)\log y - \int \frac{1}{y}\left(\frac{y}{1+y}\log y - \log(1+y)\right)\,dy\right)\Bigg|_0^\infty \tag{46}$$

$$= \left(\left(\frac{y}{1+y}\log y - \log(1+y)\right)\log y - \int \left(\frac{1}{1+y}\log y - \frac{1}{y}\log(1+y)\right)\,dy\right)\Bigg|_0^\infty \tag{47}$$

Can't figure out anti-derivative, if we can express it in elementary form at all, try another way of parts

$$\int_0^\infty \frac{(\log y)^2}{(1+y)^2}\,dy \tag{48}$$

$$= \left(-\frac{(\log y)^2}{(1+y)} + \int 2\frac{\log y}{y}\frac{1}{1+y}\,dy\right)\Bigg|_0^\infty \tag{49}$$

Same here, can't figure out if I can express anti-derivative in elementary form. So, let's switch gears here. What about moment generating functions?

$$E(e^{tX}) = \int_{-\infty}^\infty e^{tx}\frac{\theta e^{\theta(x-\mu)}}{\left(1+e^{\theta(x-\mu)}\right)^2}\,dx \tag{50}$$

Let's make use of the substitution that worked oh so well the first time, $y = e^{\theta(x-\mu)}$,

$$E(e^{tX}) = \int_0^\infty e^{t\left(\mu+\frac{1}{\theta}\log y\right)}\frac{1}{(1+y)^2}\,dy \tag{51}$$

$$= \int_0^\infty e^{t\mu}e^{\frac{t}{\theta}\log y}\frac{1}{(1+y)^2}\,dy \tag{52}$$

$$= e^{t\mu}\int_0^\infty y^{\frac{t}{\theta}}\frac{1}{(1+y)^2}\,dy \tag{53}$$

$$= e^{t\mu}\left(-\frac{y^{\frac{t}{\theta}}}{1+y} + \int \frac{t}{\theta}y^{\frac{t}{\theta}-1}\frac{1}{1+y}\,dy\right)\Bigg|_0^\infty \tag{54}$$

$$= e^{t\mu}\left(-\frac{y^{\frac{t}{\theta}}}{1+y} + \frac{t}{\theta}y^{\frac{t}{\theta}-1}\log(1+y) - \frac{t}{\theta}\int\left(\frac{t}{\theta}-1\right)y^{\frac{t}{\theta}-2}\log(1+y)\,dy\right)\Bigg|_0^\infty \tag{55}$$

Consider now,

$$\int y^{\frac{t}{\theta}-2}\log(1+y)\,dy \tag{56}$$

$$= y^{\frac{t}{\theta}-2}((1+y)\log(1+y)-y) - \left(\frac{t}{\theta}-2\right)\int y^{\frac{t}{\theta}-3}((1+y)\log(1+y)-y)\,dy \tag{57}$$

Observe that this leads us to the recursive relationship,

$$\left(\frac{t}{\theta}-1\right)\int y^{\frac{t}{\theta}-2}\log(1+y)\,dy \tag{58}$$

$$= y^{\frac{t}{\theta}-2}((1+y)\log(1+y)-y) - \left(\frac{t}{\theta}-2\right)\int y^{\frac{t}{\theta}-3}(\log(1+y)-y)\,dy \tag{59}$$

$$= y^{\frac{t}{\theta}-2}((1+y)\log(1+y)-y) + \left(\frac{\frac{t}{\theta}-2}{\frac{t}{\theta}-1}\right)y^{\frac{t}{\theta}-1} - \left(\frac{t}{\theta}-2\right)\int y^{\frac{t}{\theta}-3}\log(1+y)\,dy \tag{60}$$

Hmm... seems like a lot to unpack. Cutting our losses here, moment generating function way seems too difficult. Let's be satisfied with what we got out of this whole exercise, $E(X) = \mu$, so at least we have a method of moments estimate for one of our parameters, $\tilde{\mu} = \bar{X}$. For θ, we will just use an arbitrary starting point (pick 1 for now, we could try different starting points if things go wrong or to see if we land on different optima).
Disclosed below is the code implemented in R for this question,

```
# Question 2
# first read the data
d_source <- "http://www.utstat.toronto.edu/~brunner/data/legal/"
d_source <- paste(d_source, "mystery.data.txt",sep="")

# Quick load method
```

```
m_data <- scan(d_source)

# part (a): For density function provided find MLE
# function to compute negative log likelihood
mdll <- function(theta, datta){
        # parameters
        mu <- theta[1]
        theta2 <- theta[2]
        # length of data
        n <- length(datta)
        # intermediate term in log-likelihood expression
        intermediate <- log(1+exp(theta2*(datta-mu)))
        # negative log likelihood (written in 2 lines)
        mdll <- n*(theta2*mu-log(theta2)-theta2*mean(datta))
        mdll <- mdll + 2*sum(intermediate)
        mdll
}

# mom + arbitrary initial ests
mu <- mean(m_data)
theta2 <- 1

# use nlm (non-linear minimization) of R
theta_search <- nlm(mdll, c(mu, theta2), hessian=T, datta=m_data)

# print output
theta_search

# check eigenvalues of of hessian of search
# If both eigenvalues positive, we have positive definiteness
# which implies a local minimum of negative ll as required
eigen(theta_search$hessian)$values

# plot theta and y's (slices across optimal value)
# mu
mu_dom <- seq(from=0.1, to=10, by=0.1)
y <- mu_dom - mu_dom
for(i in 1:length(mu_dom)) {
```

```
        y[i] <- mdll(c(mu_dom[i], 0.8760225),m_data)
}
plot(mu_dom, y, type='l')

# theta
t_dom <- seq(from=0.1, to=10, by=0.1)
y <- t_dom - t_dom
for(i in 1:length(t_dom)) {
        y[i] <- mdll(c(3.3392024, t_dom[i]),m_data)
}
plot(t_dom, y, type='l')
```

By looking at the output in R console, we have the MLE estimates of $\hat{\mu} = 3.3392004$ and $\hat{\theta} = 0.8760226$. By ascertaining that the eigenvalues of the Hessian are both positive, we know that the Hessian is positive definite, we know we have a local minimum. By plotting slices across the obtained optimum (see code above), we do a sanity check of no multiple valleys to worry about (at least in the range of the plot). We also note that our Method of Moments estimate ($\tilde{\mu} = \bar{X} = 3.3806$) and our arbitrary θ starting point of 1 are quite close to the MLE values obtained. So our starting point seems to be alright (MLE obtained also matches what we had when we first did this assignment, when we tried 2 or 3 initial estimates apparently, so we also have the sanity check of multiple starting points).

2.2.2 (b)

For this (and the following) section(s), we note from the lecture slides that, according to Wald, under regularity conditions, the distribution of the MLE vector is asymptotically multivariate normal, i.e. $N_k\left(\boldsymbol{\theta}, \frac{1}{n}\mathcal{I}(\boldsymbol{\theta})^{-1}\right)$, where $\mathcal{I}(\boldsymbol{\theta})$ is the Fisher Information in one observation. As covered in the slides, the asymptotic covariance matrix can be approximated with the observed Fisher information, which is the Hessian estimated by the **nlm** function evaluated at the MLE. Basically, the estimated asymptotic covariance matrix is this Hessian inverted, and the diagonal elements are the asymptotic standard errors squared.

Now, for this question, we want the 95% confidence interval for θ, that is, we want to compute $(\hat{\theta} - S_{\hat{\theta}} z_{0.025}, \hat{\theta} + S_{\hat{\theta}} z_{0.025})$. The R code implemented to carry out this computation is disclosed below.

```
# part (b): 95% confidence interval for theta
# first invert hessian
ts_hess <- theta_search$hessian
# asymptotic covariance matrix
v_hat <- solve(ts_hess)
# since theta is the 2nd element, we need the sqrt of the 2nd
# diagonal element
theta_se <- sqrt(v_hat[2,2])
# get confidence interval bounds
theta_hat <- theta_search$estimate[2]
theta_clb <- theta_hat - qnorm(0.975)*theta_se
theta_cub <- theta_hat + qnorm(0.975)*theta_se
# print output
sprintf("theta 0.95 CI: (%.3f, %.3f)", theta_clb, theta_cub)
```

Executing the code gives us a confidence interval of (0.672, 1.080).

2.2.3 (c)

For this question, we want to test $H_0 : \mu = 0$ at significance level $\alpha = 0.05$. So our test statistic is $Z = \frac{\hat{\mu}}{S_{\hat{\mu}}}$, which we would check to see if it falls in the critical region or not to ascertain whether to reject the hypothesis or not. The R code implemented to carry out this computation is disclosed below.

```
# part (c): test mu = 0 at 0.05 significance
# 1st diagonal element of inverted hessian will give mu's se
mu_se <- sqrt(v_hat[1,1])
# compute test statistic
z_t <- theta_search$estimate[1]/mu_se
# get critical value
z_c <- qnorm(0.975)
# see if hypothesis is rejected
# display computation and test results
reject <- 'no'
if(abs(z_t)>z_c){
        reject <- 'yes'
}
# print the test stat and the critical value
```

```
sprintf('z_t: %.2f, z_c: %.2f', z_t, z_c)
sprintf('Reject?: %s', reject)
```

Since the test statistic (11.97) falls in the critical region ($z_{0.025} = 1.96$) and the test statistic is positive, we reject the null hypothesis and say that there is evidence to suggest that the value of the μ parameter is greater than 0.

3

3.1 Problem

Looking at the expression for the multivariate normal likelihood on the formula sheet, how can you tell that for *any* fixed positive definite $\boldsymbol{\Sigma}$, the likelihood is greatest when $\boldsymbol{\mu} = \bar{\mathbf{y}}$?

3.2 Solution

From the formula sheet (or lecture slide), we have the following expression for the multivariate normal likelihood:

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}(2\pi)^{\frac{p}{2}}} e^{\left\{-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu})\right\}} \tag{61}$$

$$= |\boldsymbol{\Sigma}|^{-\frac{n}{2}}(2\pi)^{-\frac{np}{2}} e^{-\frac{n}{2}\left\{tr(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) + (\bar{\mathbf{y}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu})\right\}} \tag{62}$$

where $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T$ is the sample variance-covariance matrix.

Since we have $\boldsymbol{\Sigma}$ to be positive definite, for any non-zero $(\bar{\mathbf{y}} - \boldsymbol{\mu})$ we have $(\bar{\mathbf{y}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu}) > 0$ (note that $\boldsymbol{\Sigma}$ being positive definite makes $\boldsymbol{\Sigma}^{-1}$ positive definite (confirmed via Google)). Consequently, the term $(\bar{\mathbf{y}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu})$ is a minimum for $(\bar{\mathbf{y}} - \boldsymbol{\mu}) = 0$ or $\boldsymbol{\mu} = \bar{\mathbf{y}}$. Furthermore, since all other terms are fixed, the exponent is a maximum for $\boldsymbol{\mu} = \bar{\mathbf{y}}$, and thus likelihood is greatest when $\boldsymbol{\mu} = \bar{\mathbf{y}}$.

Alternate argument (put forward when assignment was previously done): From assignment 2 (requires $\boldsymbol{\Sigma}$ be square symmetric as well), we have $(\Sigma)^{-1} = (\Sigma)^{-1/2}(\Sigma)^{-1/2}$ for (Σ) positive definite. So,

$$(\bar{\mathbf{y}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu}) \tag{63}$$

$$= (\bar{\mathbf{y}} - \boldsymbol{\mu})^T(\Sigma)^{-1/2}(\Sigma)^{-1/2}(\bar{\mathbf{y}} - \boldsymbol{\mu}) \tag{64}$$

$$= ((\Sigma)^{-1/2}(\bar{\mathbf{y}} - \boldsymbol{\mu}))^T((\Sigma)^{-1/2}(\bar{\mathbf{y}} - \boldsymbol{\mu})) \tag{65}$$

$$= \mathbf{x}^T\mathbf{x} \geq 0 \tag{66}$$

with equality when $\mathbf{x} = 0$ or $\boldsymbol{\mu} = \bar{\mathbf{y}}$. Thus, as argued earlier, since all other terms are fixed, the exponent is a maximum for $\boldsymbol{\mu} = \bar{\mathbf{y}}$, and thus likelihood is greatest when $\boldsymbol{\mu} = \bar{\mathbf{y}}$.

N.B. Note that additional assumption of $\boldsymbol{\Sigma}$ being square symmetric is required for our arguments (as the necessary argument of a positive definite matrix being one with positive eigenvalues works if matrix is symmetric according to Google apparently).

4

4.1 Problem

Based on a random sample of size n from a p-dimensional multivariate normal distribution, derive a formula for the large-sample likelihood ratio test statistic G^2 for the null hypothesis that $\boldsymbol{\Sigma}$ is diagonal (all covariances between variables are zero). You may use the likelihood function on the formula sheet. You may also use without proof the fact that the unrestricted MLE is $\hat{\theta} = (\bar{\mathbf{y}}, \hat{\boldsymbol{\Sigma}})$**.**

Hint: Because zero covariance implies independence for the multivariate normal, the joint density is a product of marginals under H_0. To be direct, I am suggesting that you *not* use the likelihood function on the formula sheet to calculate the numerator of the likelihood ratio. You'll eventually get the right answer if you insist on doing it that way, but it's a lot more work.

4.2 Solution

We want a formula for $G^2 = -2\log\left(\frac{\max_{\theta\in\Theta_0} L(\theta)}{\max_{\theta\in\Theta} L(\theta)}\right) = 2\left(-l(\hat{\theta_0}) - \left[-l(\hat{\theta})\right]\right)$. Now, we are provided the unrestricted MLE and likelihood formula, $\hat{\theta} = (\bar{\mathbf{y}}, \hat{\boldsymbol{\Sigma}})$ and $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-\frac{n}{2}}(2\pi)^{-\frac{np}{2}}e^{-\frac{n}{2}\left\{tr(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) + (\bar{\mathbf{y}}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{y}}-\boldsymbol{\mu})\right\}}$ respectively. So, right away, we can get the log likelihood at the unrestricted MLE,

$$l(\hat{\theta}) = \log\left(|\hat{\boldsymbol{\Sigma}}|^{-\frac{n}{2}}(2\pi)^{-\frac{np}{2}}e^{-\frac{n}{2}\left\{tr(\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\Sigma}}^{-1}) + (\bar{\mathbf{y}}-\bar{\mathbf{y}})^T\hat{\boldsymbol{\Sigma}}^{-1}(\bar{\mathbf{y}}-\bar{\mathbf{y}})\right\}}\right) \tag{67}$$

$$= \log\left(|\hat{\boldsymbol{\Sigma}}|^{-\frac{n}{2}}(2\pi)^{-\frac{np}{2}}e^{-\frac{n}{2}\{tr(\mathbf{I})+0\}}\right) \tag{68}$$

$$= \log\left(|\hat{\boldsymbol{\Sigma}}|^{-\frac{n}{2}}(2\pi)^{-\frac{np}{2}}e^{-\frac{np}{2}}\right) \tag{69}$$

$$= -\frac{n}{2}\log|\hat{\boldsymbol{\Sigma}}| - \frac{np}{2}\log(2\pi e) \tag{70}$$

Now, we want to derive a likelihood expression under the null hypothesis, using the hint provided to us, that is zero covariance (which is the null hypothesis) implies independence for multivariate normal. So, as suggested, the joint density under the null hypothesis is $f(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \prod_{i=1}^{p}\frac{1}{\sqrt{2\pi\sigma_{ii}}}e^{-\frac{(y_i-\mu_i)^2}{2\sigma_{ii}}}$. Thus, we have the likelihood expression under the null hypothesis as $L(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \prod_{i=1}^{n}\prod_{j=1}^{p}\frac{1}{\sqrt{2\pi\sigma_{jj}}}e^{-\frac{(y_{ij}-\mu_j)^2}{2\sigma_{jj}}}$. Consequently, the log likelihood expression under the null hypothesis is $l(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\left(-\frac{1}{2}\log(2\pi\sigma_{jj}) - \frac{(y_{ij}-\mu_j)^2}{2\sigma_{jj}}\right)$. Now, we want to get the MLE under the null hypothesis. So, without further ado, let's get to differentiating, setting to 0 and solving for the MLEs!

$$\frac{\partial l}{\partial \mu_j} = \sum_{i=1}^{n}\left(\frac{(y_{ij}-\mu_j)}{\sigma_{jj}}\right) = 0 \tag{71}$$

$$n\mu_j = \sum_{i=1}^{n} y_{ij} \tag{72}$$

$$\mu_j = \bar{y}_j \tag{73}$$

$$\frac{\partial l}{\partial \sigma_{jj}} = \sum_{i=1}^{n}\left(-\frac{1}{2}\frac{2\pi}{2\pi\sigma_{jj}} + \frac{(y_{ij}-\mu_j)^2}{2\sigma_{jj}^2}\right) = 0 \tag{74}$$

$$\sum_{i=1}^{n}\left(-\frac{1}{2\sigma_{jj}} + \frac{(y_{ij}-\mu_j)^2}{2\sigma_{jj}^2}\right) = 0 \tag{75}$$

$$\sum_{i=1}^{n}\frac{(y_{ij}-\mu_j)^2 - \sigma_{jj}}{2\sigma_{jj}^2} = 0 \tag{76}$$

$$n\sigma_{jj} = \sum_{i=1}^{n}(y_{ij}-\mu_j)^2 \tag{77}$$

$$\sigma_{jj} = \frac{1}{n}\sum_{i=1}^{n}(y_{ij} - \mu_j)^2 \tag{78}$$

Next, plug in the derived MLEs to the likelihood expression.

$$l(\hat{\theta_0}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\left(-\frac{1}{2}\log\left(2\pi\frac{1}{n}\sum_{k=1}^{n}(y_{kj} - \bar{y}_j)^2\right) - \frac{(y_{ij} - \bar{y}_j)^2}{2\frac{1}{n}\sum_{k=1}^{n}(y_{kj} - \bar{y}_j)^2}\right) \tag{79}$$

$$= \sum_{j=1}^{p}\left(-\frac{n}{2}\log\left(2\pi\frac{1}{n}\sum_{k=1}^{n}(y_{kj} - \bar{y}_j)^2\right) - \sum_{i=1}^{n}\frac{(y_{ij} - \bar{y}_j)^2}{2\frac{1}{n}\sum_{k=1}^{n}(y_{kj} - \bar{y}_j)^2}\right) \tag{80}$$

$$= -\frac{np}{2} - \frac{n}{2}\sum_{j=1}^{p}\log\left(2\pi\frac{1}{n}\sum_{k=1}^{n}(y_{kj} - \bar{y}_j)^2\right) \tag{81}$$

$$= -\frac{np}{2} - \frac{n}{2}\sum_{j=1}^{p}\left(\log(2\pi) + \log\left(\frac{1}{n}\sum_{k=1}^{n}(y_{kj} - \bar{y}_j)^2\right)\right) \tag{82}$$

$$= -\frac{np}{2} - \frac{np}{2}\log(2\pi) - \frac{n}{2}\sum_{j=1}^{p}\log\left(\frac{1}{n}\sum_{k=1}^{n}(y_{kj} - \bar{y}_j)^2\right) \tag{83}$$

$$= -\frac{np}{2}\log(2\pi e) - \frac{n}{2}\sum_{j=1}^{p}\log\left(\frac{1}{n}\sum_{k=1}^{n}(y_{kj} - \bar{y}_j)^2\right) \tag{84}$$

$$= -\frac{np}{2}\log(2\pi e) - \frac{n}{2}tr(\log(\hat{\mathbf{\Sigma}})) \tag{85}$$

Finally, we can put all our work together for the required formula for G^2,

$$G^2 = 2\left(-l(\hat{\theta_0}) - \left[-l(\hat{\theta})\right]\right) \tag{86}$$

$$= 2\left(\frac{np}{2}\log(2\pi e) + \frac{n}{2}tr(\log(\hat{\mathbf{\Sigma}})) - \left[\frac{n}{2}\log|\hat{\mathbf{\Sigma}}| + \frac{np}{2}\log(2\pi e)\right]\right) \tag{87}$$

$$= n(tr(\log(\hat{\mathbf{\Sigma}})) - \log|\hat{\mathbf{\Sigma}}|) \tag{88}$$

Furthermore, according to the lecture slides, under the null hypothesis, G^2 has an an approximate chi-square distribution for large sample size, with degrees of freedom equal to number of non-redundant, linear equalities specified by the null hypothesis. In our case, under the null hypothesis, all covariances

between the variables are zero, so the degrees of freedom is equal to $\frac{p(p-1)}{2}$ (note the keyword *non-redundant*, meaning $\sigma_{ij} = 0$ and $\sigma_{ji} = 0$ count as one non-redundant equality, i.e. the degrees of freedom is equal to number of unique pairings of the p variables, that is not counting different arrangements of the same pairing). Thus, under the null hypothesis, G^2 is an approximate chi-square distribution for large sample size with $\frac{p(p-1)}{2}$ degrees of freedom.

N.B. In case it was not mentioned before, from the slides, $\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T$.

5

5.1 Problem

The file http://www.utstat.toronto.edu/~brunner/data/illegal/bp.data.txt has diastolic blood pressure, education, cholesterol, number of cigarettes per day and weight pounds for a sample of middle-aged men. There are missing values; **summary** on the data frame will tell you what they are.

Assuming multivariate normality and using R, carry out a large-sample likelihood ratio test to determine whether there are any non-zero covariances among just these three variables: education, number of cigarettes, and weight. Guided by the usual $\alpha = 0.05$ significance level, what do you conclude? Is there evidence that the three variables are related (non-independent)? Answer Yes or No. For this question, let's agree that we will base the sample covariance matrix only on *complete observations*. That is, there will be no missing values on any variables. Don't forget that $\hat{\mathbf{\Sigma}}$, like $\hat{\sigma}_j^2$, has n in the denominator and not $n-1$. What is n?

5.2 Solution

As mentioned, assume a three dimensional multivariate normality among the three variables of interest, education, number of cigarettes, and weight. We have already derived an expression for the test statistic in the previous question,

$$G^2 = n(tr(\log(\hat{\mathbf{\Sigma}})) - \log|\hat{\mathbf{\Sigma}}|) \tag{89}$$

where $\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T$ and n is the number of *complete observations* (since we were asked to base our sample covariance matrix (and thus the hypothesis test) on *complete observations*).

Furthermore, according to the lecture slides, under the null hypothesis, G^2 has an an approximate chi-square distribution for large sample size, with degrees of freedom equal to number of non-redundant, linear equalities specified by the null hypothesis. In our case, under the null hypothesis, all covariances between the three variables of interest are zero (that way if we reject it, then there is at least one non-zero covariance among the three variables, which we are testing for), so the degrees of freedom is equal to 3 (note the keyword *non-redundant*, meaning $\sigma_{ij} = 0$ and $\sigma_{ji} = 0$ count as one non-redundant equality, i.e. the degrees of freedom is equal to number of unique pairings of the 3 variables, that is not counting different arrangements of the same pairing). Provided below is the R code implemented to carry out this test.

```
# Question 5: Large sample likelihood ratio test for multi-normal
# Load the data:
d_source <- "http://www.utstat.toronto.edu/~brunner/data/illegal/"
d_source <- paste(d_source, "bp.data.txt", sep = "")
bp_data <- read.table(d_source, header = T)

# will work with question 4's result
# idea is to isolate out the subset of variables we are
#       interested in and then apply the likelihood ratio test stat
#       expression derived in Q4
# filter out very outlying values (observed using summary() and
#       boxplot())
# using summary() and boxplot(), found observations with education
#   and cig/day values of 99 and 999 respectively
# numbers seem too high, may be artificial, coded-in numbers
#   representing missing values, so filter them out:
bp_data_nm <- bp_data[bp_data$edu != 99,]
bp_data_nm <- bp_data_nm[bp_data_nm$cig != 999,]

# extract out the variables data we are interested in
X <- cbind(bp_data_nm$edu, bp_data_nm$cig, bp_data_nm$wt)
```

```
# next calculate the sample variance-covariance matrix
n <- dim(X)[1]
# note that var() in R divided by (n-1), so multiply by (n-1)/n to
#   ensure we are dividing by n as instructed
Sigma_hat <- var(X)*((n-1)/n)
#now the likelihood ratio
G2 <- n*(sum(log(diag(Sigma_hat))) - log(det(Sigma_hat)))
# df = number of non-redundant linear combinations
# s_12 = 0; s_13 = 0; s_23 = 0
# df = 3
# get p-value
pval <- 1-pchisq(G2, df = 3)

# print the test stat and the p-val
sprintf("G2: %.5f; pval: %.5f", G2, pval)
# see if we reject
# significance level
alpha <- 0.05
reject <- 'no'
if(pval<alpha){
        reject <- 'yes'
}
sprintf('Reject?: %s', reject)
```

Looking at the output in R console, we get a test statistic value of $G^2 = 2.93946$ and a p-value of 0.40105. Since p-value is greater than the significance level, $\alpha = 0.05$, we say that, **no**, there is insufficient evidence to suggest that there is any relation (i.e. correlation or dependence) between the three variables. Finally, n, or the number of *complete observations* in our test was 211 (checked in R console).

6

6.1 Problem

Here is a useful variation on Problem 4. Suppose n independent and identically (distributed?) data vectors $\mathbf{d}_i = \begin{pmatrix} \mathbf{x}_i \\ \mathbf{y}_i \end{pmatrix}$ are multivariate normal. The

notation means that $\mathbf{d}_i$ is $\mathbf{x_i}$ stacked on top of $\mathbf{y}_i$. For example, $\mathbf{x}_i$ could be physical measurements and $\mathbf{y}_i$ could be psychological measurements. Derive a likelihood ratio test to determine whether $\mathbf{x}_i$ and $\mathbf{y}_i$ are independent. Your answer is a formula for G^2 and a formula for the degrees of freedom. Part of the job here is to make up good, simple notation.

6.2 Solution

We want a formula for $G^2 = -2\log\left(\frac{\max_{\theta\in\Theta_0} L(\theta)}{\max_{\theta\in\Theta} L(\theta)}\right) = 2\left(-l(\hat{\theta_0}) - \left[-l(\hat{\theta})\right]\right)$. Now, we are provided the unrestricted MLE and likelihood formula, $\hat{\theta} = (\bar{\mathbf{d}}, \hat{\mathbf{\Sigma}})$ and $L(\boldsymbol{\mu}, \mathbf{\Sigma}) = |\mathbf{\Sigma}|^{-\frac{n}{2}}(2\pi)^{-\frac{np}{2}} e^{-\frac{n}{2}\left\{tr(\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}) + (\bar{\mathbf{d}}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\bar{\mathbf{d}}-\boldsymbol{\mu})\right\}}$ respectively (from problem 4 and the lecture slides). So, right away, we can get the log likelihood at the unrestricted MLE,

$$l(\hat{\theta}) = \log\left(|\hat{\mathbf{\Sigma}}|^{-\frac{n}{2}}(2\pi)^{-\frac{np}{2}} e^{-\frac{n}{2}\left\{tr(\hat{\mathbf{\Sigma}}\hat{\mathbf{\Sigma}}^{-1}) + (\bar{\mathbf{d}}-\bar{\mathbf{d}})^T\hat{\mathbf{\Sigma}}^{-1}(\bar{\mathbf{d}}-\bar{\mathbf{d}})\right\}}\right) \tag{90}$$

$$= \log\left(|\hat{\mathbf{\Sigma}}|^{-\frac{n}{2}}(2\pi)^{-\frac{np}{2}} e^{-\frac{n}{2}\{tr(\mathbf{I})+0\}}\right) \tag{91}$$

$$= \log\left(|\hat{\mathbf{\Sigma}}|^{-\frac{n}{2}}(2\pi)^{-\frac{np}{2}} e^{-\frac{np}{2}}\right) \tag{92}$$

$$= -\frac{n}{2}\log|\hat{\mathbf{\Sigma}}| - \frac{np}{2}\log(2\pi e) \tag{93}$$

Now, we want to derive a likelihood expression under the null hypothesis. Under the null hypothesis, $\mathbf{x}_i$ and $\mathbf{y}_i$ are independent. Thus, the likelihood expression under the null hypothesis is

$$L\left(\begin{pmatrix}\boldsymbol{\mu}_\mathbf{x}\\ \boldsymbol{\mu}_\mathbf{y}\end{pmatrix}, \begin{pmatrix}\mathbf{\Sigma}_\mathbf{x} & \mathbf{0}\\ \mathbf{0} & \mathbf{\Sigma}_\mathbf{y}\end{pmatrix}\right) = \prod_{i=1}^{n} f_\mathbf{x}(\mathbf{x}_i; \boldsymbol{\mu}_\mathbf{x}, \mathbf{\Sigma}_\mathbf{x}) f_\mathbf{y}(\mathbf{y}_i; \boldsymbol{\mu}_\mathbf{y}, \mathbf{\Sigma}_\mathbf{y}) \tag{94}$$

$$= L_\mathbf{x}(\boldsymbol{\mu}_\mathbf{x}, \mathbf{\Sigma}_\mathbf{x}) L_\mathbf{y}(\boldsymbol{\mu}_\mathbf{y}, \mathbf{\Sigma}_\mathbf{y}) \tag{95}$$

where $L_\mathbf{x}(\boldsymbol{\mu}_\mathbf{x}, \mathbf{\Sigma}_\mathbf{x})$ and $L_\mathbf{y}(\boldsymbol{\mu}_\mathbf{y}, \mathbf{\Sigma}_\mathbf{y})$ are the general multivariate normal likelihood expression for $\mathbf{x}$ and $\mathbf{y}$ respectively (both are multivariate normal by the multivariate normal property that any subset of components of a multivariate normal vector is multivariate normal too). Realize that the minimum of both $L_\mathbf{x}(\boldsymbol{\mu}_\mathbf{x}, \mathbf{\Sigma}_\mathbf{x})$ and $L_\mathbf{y}(\boldsymbol{\mu}_\mathbf{y}, \mathbf{\Sigma}_\mathbf{y})$ is at the unrestricted MLE for each expression, and it is straightforward to see that the product of these evaluates

to the minimum of the likelihood under the null hypothesis. Therefore, we can just use our log likelihood expression derived above for getting the log likelihood under the null hypothesis too,

$$l(\hat{\theta_0}) = l_{\mathbf{x}}(\hat{\theta_{\mathbf{x}}}) + l_{\mathbf{y}}(\hat{\theta_{\mathbf{y}}}) \tag{96}$$

$$= -\frac{n}{2}\log|\widehat{\mathbf{\Sigma}_{\mathbf{x}}}| - \frac{np_{\mathbf{x}}}{2}\log(2\pi e) - \frac{n}{2}\log|\widehat{\mathbf{\Sigma}_{\mathbf{y}}}| - \frac{np_{\mathbf{y}}}{2}\log(2\pi e) \tag{97}$$

$$= -\frac{n}{2}\log\left(|\widehat{\mathbf{\Sigma}_{\mathbf{x}}}||\widehat{\mathbf{\Sigma}_{\mathbf{y}}}|\right) - \frac{np}{2}\log(2\pi e) \tag{98}$$

Finally, we can put all our work together for the required formula for G^2,

$$G^2 = 2\left(-l(\hat{\theta_0}) - \left[-l(\hat{\theta})\right]\right) \tag{99}$$

$$= 2\left(\frac{n}{2}\log\left(|\widehat{\mathbf{\Sigma}_{\mathbf{x}}}||\widehat{\mathbf{\Sigma}_{\mathbf{y}}}|\right) + \frac{np}{2}\log(2\pi e) - \left[\frac{n}{2}\log|\hat{\mathbf{\Sigma}}| + \frac{np}{2}\log(2\pi e)\right]\right) \tag{100}$$

$$= n\left(\log\left(|\widehat{\mathbf{\Sigma}_{\mathbf{x}}}||\widehat{\mathbf{\Sigma}_{\mathbf{y}}}|\right) - \log|\hat{\mathbf{\Sigma}}|\right) \tag{101}$$

$$= n\log\left(\frac{|\widehat{\mathbf{\Sigma}_{\mathbf{x}}}||\widehat{\mathbf{\Sigma}_{\mathbf{y}}}|}{|\hat{\mathbf{\Sigma}}|}\right) \tag{102}$$

Furthermore, according to the lecture slides, under the null hypothesis, G^2 has an an approximate chi-square distribution for large sample size, with degrees of freedom equal to number of non-redundant, linear equalities specified by the null hypothesis. In our case, we want the number of unique x_i and y_j pairings (here by x_i and y_j we mean the ith and jth element of the $\mathbf{x}$ and $\mathbf{y}$ vectors respectively). We have $p_{\mathbf{y}}$ elements of the $\mathbf{y}$ vector for each element of the $\mathbf{x}$ vector (of which there are $p_{\mathbf{x}}$). So we have $p_{\mathbf{x}}p_{\mathbf{y}}$ unique pairings, and thus under the null hypothesis, G^2 is an approximate chi-square distribution for large sample size with $p_{\mathbf{x}}p_{\mathbf{y}}$ degrees of freedom.

N.B. In case it was not mentioned before, from the slides, $\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{d}_i - \bar{\mathbf{d}})(\mathbf{d}_i - \bar{\mathbf{d}})^T$. Likewise, $\widehat{\mathbf{\Sigma}_{\mathbf{x}}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$ and $\widehat{\mathbf{\Sigma}_{\mathbf{y}}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T$.

7

7.1 Problem

I said there would be just one more question, but this one does not count. It's just set-up for Question 8.

7.1.1 (a)

Let X and Y be scalar random variables and let a be a constant. Using the definition $Cov(X,Y) = E\{(X-\mu_x)(Y-\mu_y)\}$, show $Cov(aX,Y) = a\,Cov(X,Y)$.

7.1.2 (b)

Let X_1, X_2, Y_1 and Y_2 be scalar random variables. Show

$$Cov(X_1+X_2,Y_1+Y_2) = Cov(X_1,Y_1)+Cov(X_1,Y_2)+Cov(X_2,Y_1)+Cov(X_2,Y_2) \tag{103}$$

Clearly, this extends to larger numbers of terms. Don't hesitate to use it.

7.1.3 (c)

Let $(X_1,Y_1),...,(X_n,Y_n)$ be a random sample from a bivariate distribution with $E(X_i)=\mu_x$, $E(Y_i)=\mu_y$, $Var(X_i)=\sigma_x^2$, $Var(Y_i)=\sigma_y^2$, and $Cov(X_i,Y_i)=\sigma_{xy}$. Find $Cov(\bar{X},\bar{Y})$.

7.2 Solution

7.2.1 (a)

$$Cov(aX,Y) = E\{(aX-a\mu_x)(Y-\mu_y)\} \tag{104}$$
$$= a\,E\{(X-\mu_x)(Y-\mu_y)\} = a\,Cov(X,Y) \tag{105}$$

7.2.2 (b)

$$Cov(X_1 + X_2, Y_1 + Y_2) \tag{106}$$
$$= E\{(X_1 + X_2 - E[X_1 + X_2])(Y_1 + Y_2 - E[Y_1 + Y_2])\} \tag{107}$$
$$= E\{(X_1 - \mu_{X_1} + X_2 - \mu_{X_2})(Y_1 - \mu_{Y_1} + Y_2 - \mu_{Y_2})\} \tag{108}$$
$$= \sum_{i=1}^{2}\sum_{j=1}^{2} E\{(X_i - \mu_{X_i})(Y_j - \mu_{Y_j})\} \tag{109}$$
$$= Cov(X_1, Y_1) + Cov(X_1, Y_2) + Cov(X_2, Y_1) + Cov(X_2, Y_2) \tag{110}$$

7.2.3 (c)

$$Cov(\bar{X}, \bar{Y}) = Cov\left(\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{i=1}^{n} Y_i\right) \tag{111}$$
$$= \frac{1}{n} Cov\left(\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{i=1}^{n} Y_i\right) \tag{112}$$
$$= \frac{1}{n^2} Cov\left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Y_i\right) \tag{113}$$
$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, Y_j) \tag{114}$$
$$= \frac{1}{n^2}\left(\sum_{i=1}^{n} Cov(X_i, Y_i)\right) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} Cov(X_i, Y_j) \tag{115}$$
$$= \frac{1}{n^2}\left(\sum_{i=1}^{n} \sigma_{xy}\right) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} 0 \tag{116}$$
$$= \frac{\sigma_{xy}}{n} \tag{117}$$

N.B. Note that the third step above is a result of a similar procedure to what was done in (a) (i.e. showing $Cov(X, aY) = a\,Cov(X, Y)$ is similar to how $Cov(aX, Y) = a\,Cov(X, Y)$ was shown in (a)). Additionally the step after is an extension of the result shown in (b) to a larger number of terms as mentioned. Furthermore, in the second-last line, we made use of

$Cov(X_i, Y_j) = 0$ for $i \neq j$ as we have been provided that $(X_1, Y_1), ..., (X_n, Y_n)$ is a random sample (i.e. sample elements are independent).

8

8.1 Problem

Independently for $i = 1, ..., n$, let $Y_i = \beta X_i + \epsilon_i$, where $E(X_i) = \mu \neq 0$, $E(\epsilon_i) = 0$, $Var(X_i) = \sigma_x^2$, $Var(\epsilon_i) = \sigma_\epsilon^2$, and ϵ_i is independent of X_i. Some data from this model are at http://www.utstat.toronto.edu/~brunner/data/legal/xy.data.txt.

Let

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2} \text{ and } \hat{\beta}_2 = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i}. \tag{118}$$

You know both of these estimators are consistent; there is no need to prove it again.

8.1.1 (a)

$\hat{\beta}_1$ is just the ordinary least-squares estimator, though's (though it's?) not technically least squares anymore because the explanatory variable is random. Use R's **lm** function, fit a regression line through the origin, obtain a numerical $\hat{\beta}_1$, and calculate a 95% confidence interval for β. It's not easy to justify the confidence interval formally, because the error terms are definitely not normal. But please do it anyway for comparison.

8.1.2 (b)

Now calculate $\hat{\beta}_2$ for these data, and obtain the corresponding large-sample 95% confidence interval for β. Here is a bit of discussion to help you along.

You're definitely going to use the delta method, but please don't overthink it. It's possible to walk down a dark path here by finding asymptotically normal estimators for *all* the parameters, and seeking an estimate of their asymptotic covariance matrix. Instead, retreat to the model of Question 7c, and realize that for reasons of your own, you are looking at a function $g(\mu_x, \mu_y)$. The

asymptotic covariance matrix of $(\bar{x}, \bar{y})$ is actually easy, because you obtain the exact covariance matrix for any n. That's why you did Question 7. I used the **var** function to get an estimate of $cov\begin{pmatrix} x_i \\ y_i \end{pmatrix}$. Finally, the lower limit of my confidence interval was 1.474.

8.2 Solution

8.2.1 (a)

The R code implemented for carrying out the computations for this question is disclosed below.

```
# STA2101 Assignment 4 Ques 8

# first retrieve the data
d_source <- "http://www.utstat.toronto.edu/~brunner/data/legal/"
d_source <- paste(d_source, "xy.data.txt", sep = "")
xy_data <- read.table(d_source, header = T)

# part (a)
# fit regression line through origin
lm1 <- lm(y ~ x-1, data = xy_data)

# get summary
summary(lm1)

# check to see if parameter estimate in summary and given by formula
#       match
beta1_form <- sum(xy_data$x*xy_data$y)/sum(xy_data$x^2)
sprintf("via formula: %.4f", beta1_form)

# they do, so just use the one from summary
# compute confidence interval
# get estimate
beta1 <- summary(lm1)$coefficients[1]
# get standard error
beta1_se <- summary(lm1)$coefficients[2]
# compute confidence interval bounds and output
```

```
# inferential investigation of regression parameters derived from
#       normality assumption
# so will use t-distribution
# length of data
n <- dim(xy_data)[1]
# degrees of freedom n-1 as only one parameter estimated in this
#       regression model
beta1_clb <- beta1 - qt(0.975, df=n-1)*beta1_se
beta1_cub <- beta1 + qt(0.975, df=n-1)*beta1_se
# print output
sprintf("beta1: %.3f", beta1)
sprintf("beta1 0.95 CI: (%.3f, %.3f)", beta1_clb, beta1_cub)
```

Executing the above code in R console, we get $\hat{\beta}_1 = 1.981$ and a 95% confidence interval for β of (1.619, 2.343).

8.2.2 (b)

First, we notice that $\hat{\beta}_2 = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i} = \frac{\bar{Y}}{\bar{X}}$. So, if we're planning to use the delta method as hinted, presumably we will use it on the joint distribution of $(\bar{x}, \bar{y})$. Presumably, we can also argue a bivariate normal distribution for $(\bar{x}, \bar{y})$, using CLT as a basis. From our work in the previous question, we have $Cov(\bar{X}, \bar{Y}) = \frac{\sigma_{xy}}{n}$. We also have $Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^n Var(X_i) = \frac{\sigma_x^2}{n}$ and likewise $Var(\bar{Y}) = \frac{1}{n^2}\sum_{i=1}^n Var(Y_i) = \frac{\sigma_y^2}{n}$ (note that any covariance term disappears from these variance expressions because of independence between the data points). Thus, provided that (note that by consulting the slides, $\mathbf{\Sigma}$ below is the covariance matrix of the i.i.d. $\begin{pmatrix} X_i \\ Y_i \end{pmatrix}$)

$$\sqrt{n}\left(\begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix} - \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}\right) \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma}) = N\left(\mathbf{0}, \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}\right) \tag{119}$$

we have by delta method,

$$\sqrt{n}\left(\hat{\beta}_2 - \beta\right) = \sqrt{n}\left(g\left(\begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix}\right) - g\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}\right)\right) = \sqrt{n}\left(\frac{\bar{Y}}{\bar{X}} - \frac{\mu_y}{\mu_x}\right) \tag{120}$$

$$\xrightarrow{d} \begin{pmatrix} -\frac{\mu_y}{\mu_x^2} & \frac{1}{\mu_x} \end{pmatrix} N(\mathbf{0}, \mathbf{\Sigma}) = N\left(0, \begin{pmatrix} -\frac{\mu_y}{\mu_x^2} & \frac{1}{\mu_x} \end{pmatrix} \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} \begin{pmatrix} -\frac{\mu_y}{\mu_x^2} \\ \frac{1}{\mu_x} \end{pmatrix}\right) \quad (121)$$

$$= N\left(0, \begin{pmatrix} -\frac{\mu_y\sigma_x^2}{\mu_x^2} + \frac{\sigma_{xy}}{\mu_x} & -\frac{\mu_y\sigma_{xy}}{\mu_x^2} + \frac{\sigma_y^2}{\mu_x} \end{pmatrix} \begin{pmatrix} -\frac{\mu_y}{\mu_x^2} \\ \frac{1}{\mu_x} \end{pmatrix}\right) \quad (122)$$

$$= N\left(0, \left(\frac{\mu_y^2\sigma_x^2}{\mu_x^4} - \frac{\mu_y\sigma_{xy}}{\mu_x^3} - \frac{\mu_y\sigma_{xy}}{\mu_x^3} + \frac{\sigma_y^2}{\mu_x^2}\right)\right) \quad (123)$$

$$= N\left(0, \frac{\mu_y^2\sigma_x^2}{\mu_x^4} - 2\frac{\mu_y\sigma_{xy}}{\mu_x^3} + \frac{\sigma_y^2}{\mu_x^2}\right) \quad (124)$$

thus ending up with the asymptotic distribution

$$\hat{\beta}_2 \dot{\sim} N\left(\beta, \frac{1}{n}\left(\frac{\mu_y^2\sigma_x^2}{\mu_x^4} - 2\frac{\mu_y\sigma_{xy}}{\mu_x^3} + \frac{\sigma_y^2}{\mu_x^2}\right)\right) \quad (125)$$

hence to compute the confidence interval, we compute $\left(\hat{\beta}_2 - z_{0.025}S_{\hat{\beta}_2}, \hat{\beta}_2 + z_{0.025}S_{\hat{\beta}_2}\right)$, where $S_{\hat{\beta}_2}$ is the square root of the variance expression we have above with the population moments replaced by the sample moments. The code implemented in R to carry out this question's computations is disclosed below.

```
# part (b)
# first get estimate for beta2
beta2 <- sum(xy_data$y)/sum(xy_data$x)
# next need to compute derived asymptotic se
# length of data
n <- dim(xy_data)[1]

# next compute the standard error derived using delta method
# sample means and variances
mu_x <- mean(xy_data$x)
mu_y <- mean(xy_data$y)
# am not using the (n-1)/n correction factor here because otherwise
#       answer is off from prof's (by 0.001); maybe he forgot to do it?
#       Or we're not doing it here for some reason?
Sigma <- var(xy_data)

#compute s.e.
```

```
beta2_var <- (mu_y^2/mu_x^4)*Sigma[1,1] + (Sigma[2,2]/mu_x^2)
beta2_var <- (1/n)*(beta2_var - 2*(mu_y/mu_x^3)*Sigma[1,2])
beta2_se <- sqrt(beta2_var)

# 95% interval
# compute confidence interval bounds and output
beta2_clb <- beta2 - qnorm(0.975)*beta2_se
beta2_cub <- beta2 + qnorm(0.975)*beta2_se
# print output
sprintf("beta2: %.3f", beta2)
sprintf("beta2 0.95 CI: (%.3f, %.3f)", beta2_clb, beta2_cub)
```

Executing the above code in R console, we get $\hat{\beta}_2 = 1.890$ and a 95% confidence interval for β of (1.474, 2.307).