Practice midterm exam, STATS 401 F18

Instructions. You have a time allowance of 80 minutes. The exam is closed book. Any electronic devices in your possession must be turned off and remain in a bag on the floor. This includes cell phones, calculators and internet-enabled watches. If you need extra paper, please number the pages and put your name and UMID on each page.

Q1. Summation and matrix exercises.

(a) [1 point] Let $\mathbb{X} = [x_{ij}]$ be a 3×2 matrix with (i, j) entry given by $x_{ij} = 2i$. Write out $\mathbb{X}$, evaluating each of the six entries of the matrix.

Solution.

$$\mathbb{X} = \begin{bmatrix} 2 & 2 \\ 4 & 4 \\ 6 & 6 \end{bmatrix}$$

(b) [1 point] Hence, evaluate the double sum $\sum_{i=1}^{3} \sum_{j=1}^{2} 2i$.

Solution. The double sum adds up all (i, j) entries in the matrix, so

$$\sum_{i=1}^{3} \sum_{j=1}^{2} 2i = 24.$$

(c) [2 points] Evaluate $\mathbb{X}^{\mathrm{T}}\mathbb{X}$ where

$$\mathbb{X} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}$$

Solution.

$$\mathbb{X}^{\mathrm{T}}\mathbb{X} = \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{bmatrix}$$

Q2. Manipulating vectors and matrices in R.

(a) [2 points] Write the output of

```
matrix(c(rep(1,2), rep(0, 2), rep(0,2), rep(1,2)), nrow = 4)
```

Solution.

$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

(b) [1 point] Which of the following is the output to `pnorm(c(-2,2))`

(i)
```
[1] 0.02275013 0.97724987
```

(ii)
```
Error in pnorm(c(-2,2)) : vector argument to scalar function
```

(iii)
```
[1] 0.1586553 0.8413447
```

(iv)
```
0.02275013
Warning message:
In pnorm(c(-2,2)) :
Vector argument to scalar function.
Function applied to only the first vector component.
```

(v)
```
0.1586553
Warning message:
In pnorm(c(-2,2)) :
Vector argument to scalar function.
Function applied to only the first vector component.
```

Solution. (i)

Q3. Investigating a probability distribution.

Homework 4 involved investigating the t and F distributions. Recall that $X_t \sim t(n)$ and $X_F \sim F(m,n)$ if

$$X_t = \frac{Y}{\sqrt{\sum_{j=1}^{n} Z_j^2/n}}$$

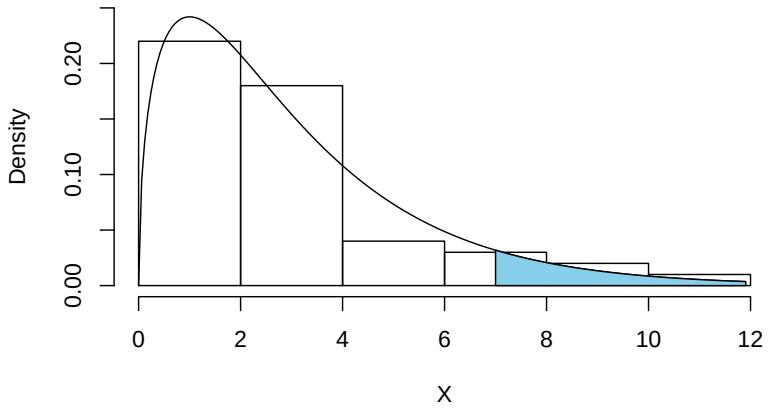
$$X_F = \frac{\sum_{i=1}^{m} Y_i^2/m}{\sum_{j=1}^{n} Z_j^2/n}$$

where Y, $Y_1, \ldots, Y_m$ and $Z_1, \ldots, Z_n$ are independent normal$[0,1]$ random variables. Another related distribution is the chi-square distribution. We say $X \sim \text{chisq}(n)$ if

$$X = \sum_{i=1}^{n} Z_i^2$$

where $Z_1, \ldots, Z_n \sim$ iid normal$[0,1]$. Code simulating a sample from the chi-square distribution with $n = 3$ and comparing it to the chi-square density function is given below. The code for the shaded tail is not given.

```
X <- rchisq(50,df=3)
hist(X,prob=T)
x <- seq(from=0,to=max(X),length=200)
y <- dchisq(x,df=3)
lines(x,y)
```



(a) [2 points] Do you think that the chi-square density function will look like a normal curve if we picked a large value of n? Explain your reasoning.

Solution. Yes. X is written as a sum of n random variables, so as n gets large a central limit theorem should apply. We could also note that these random variables all have the same distribution and they are independent, so we do not expect any single event to dominate the sum. For the examples we have seen in class, if a variable is the sum of many contributions and no single contribution dominates then a normal approximation is appropriate.

(b) [2 points] Guess an R command that will give the area under the probability density curve to the right of 7, corresponding to the shaded right tail on the plot. The shading may or may not be printed well on your copy of the exam, but that is not critical to the question. You are not expected to have previously seen how the chi-squared distribution works in R, but you can assume it works similarly to things you have seen: it does!

Solution. `1-pchisq(7,df=3)`

(c) [2 points] Describe a way to numerically obtain the expected value, E[X], when $X \sim$ chisq(3). Your description should include either R code or mathematical notation. Also, explain why your approach is valid.

Solution. The expected value of a random variable is the average of a large number of draws from that distribution. Therefore, we can numerically obtain the expected value from a large sample obtained using `rchisq()`, for example `mean(rchisq(10000,df=3))`. This will not give an exact answer, since there will be some chance variation in the sample average, but for a large sample it will be close.

Q4. Fitting a linear model by least squares

The director of admissions at a large state university wants to assess how well academic success can be predicted based on information available at admission. She collects a dataset of 705 students. She decides to take freshman GPA as the response variable, and she has access to ACT exam scores and percentile ranking of each student within their high school.

```
gpa <- read.table("gpa.txt",header=T)
```

```
head(gpa)
```

```
##    GPA High_School ACT
## 1 0.98          61  20
## 2 1.13          84  20
## 3 1.25          74  19
## 4 1.32          95  23
## 5 1.48          77  28
## 6 1.57          47  23
```

(a). Write the sample version of a linear model to address this question in subscript form.

Solution. [2 points]

The model is

$$y_i = b_1 x_{i1} + b_2 x_{i2} + b_3 + e_i, \quad i = 1, \ldots, n$$

where y_i is freshman GPA for as the response variable for student i, x_{i1} is the ACT exam score for this student, x_{i2} is the percentile ranking of the student within their high school, and $n = 705$. e_i is the residual error for student i. b_1, b_2 and b_3 are coefficients chosen by least squares.

(b). Write the sample version of this linear model in matrix form. Some of the quantities you have to define may be the same as the quantities you defined previously. Nevertheless, please make this model description self-contained.

Solution. [2 points]

The model is

$$\mathbf{y} = \mathbb{X}\mathbf{b} + \mathbf{e},$$

where

- $\mathbf{y} = (y_1, \ldots, y_n)$ is a vector of freshman GPA scores with $n = 705$
- $\mathbb{X} = [x_{ij}]$ is a $n \times 3$ matrix with x_{i1} being the ACT exam score for student i, x_{i2} being the percentile ranking of the student within their high school, and $x_{i3} = 1$ for $i = 1, \ldots, n$.
- $\mathbf{b} = (b_1, b_2, b_3)$ is a vector of coefficients, chosen by least squares.
- $\mathbf{e} = (e_1, \ldots, e_n)$ is a vector of residuals.
- All vectors are interpreted as column vectors.

(c). The following output fits a linear model in R.

```
gpa_lm <- lm(GPA~ACT+High_School,data=gpa)
summary(gpa_lm)
```

```
##
## Call:
## lm(formula = GPA ~ ACT + High_School, data = gpa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -2.10265 -0.29862  0.07311  0.40355  1.31336
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.292793   0.136725   9.455  < 2e-16 ***
## ACT         0.037210   0.005939   6.266 6.48e-10 ***
## High_School 0.010022   0.001279   7.835 1.74e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5672 on 702 degrees of freedom
## Multiple R-squared:  0.2033, Adjusted R-squared:  0.2011
## F-statistic: 89.59 on 2 and 702 DF,  p-value: < 2.2e-16
```

Explain how the coefficient estimates and the residual standard error presented in this output were calculated.

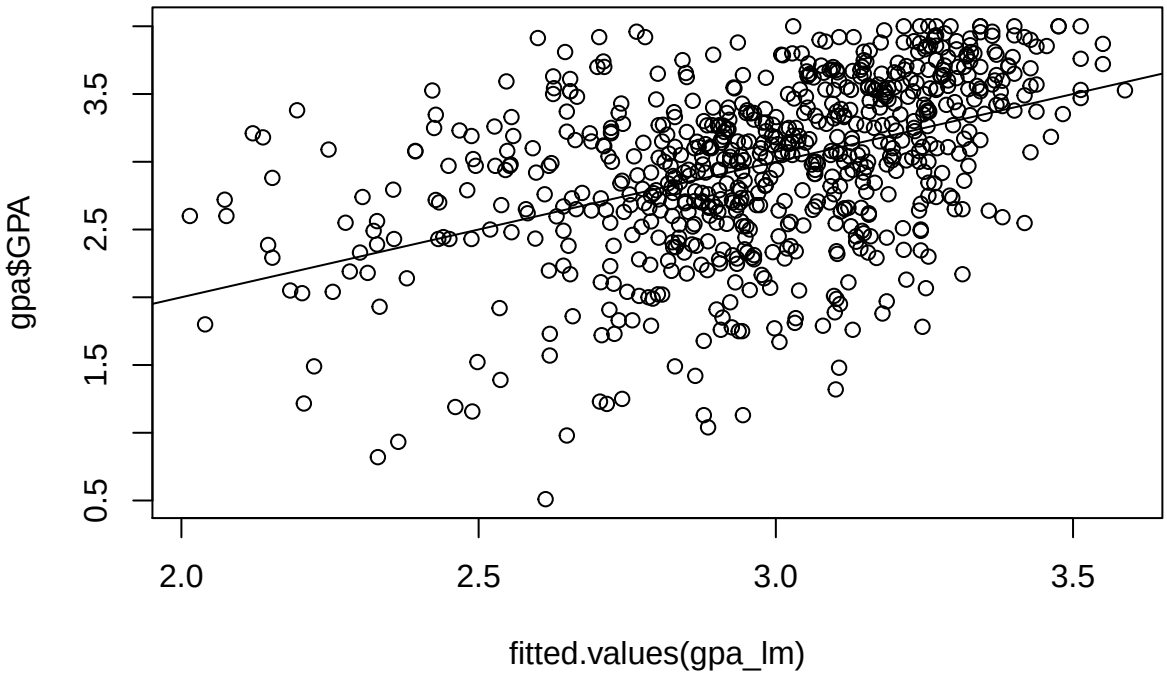
Solution. [2 points]

The coefficient estimates are the vector $\mathbf{b}$ from F2, calculated by least squares using the formula

$$\mathbf{b} = \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y}.$$

(d). Explain what the **fitted values** are for a linear model. Comment briefly on what the admissions director should learn (if anything) from the following plot of the freshman GPA of each patient plotted against the fitted value.

```
plot(x=fitted.values(gpa_lm),y=gpa$GPA)
abline(a=0,b=1)
```



Solution. [2 points]

The fitted values are the values of the response variables with the rersidual errors removed. The vector $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)$ of fitted values is calculated as

$$\hat{\mathbf{y}} = \mathbb{X}\mathbf{b} = \mathbb{X}\big(\mathbb{X}^{\mathrm{T}}\mathbb{X}\big)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y}.$$

Plotting the response against the fitted values, we see that the explanatory variables can explain around 1 GPA point out of the total spread of around 3 GPA points. Other things to look for are (a) there are no noticeable extreme points, known as outliers; (b) the points are roughly football shaped, but with somewhat higher variability at lower values of fitted GPA.

Q5. Working with the mean, variance, and covariance with a normal approximation.

The sample variance/covariance matrix for the GPA dataset is

```
V <- var(gpa)
round(V,2)
```

```
##              GPA High_School   ACT
## GPA         0.40        4.71  0.93
## High_School 4.71      347.22 33.10
## ACT         0.93       33.10 16.11
```

(a) [1 point] Write a numeric calculation using values in this matrix to give the sample correlation between ACT score and high school rank. You are not expected to evaluate it.

Solution. The sample covariance is 33.1. The sample variances are 347.22 and 16.11. The correlation is

$$\frac{33.1}{\sqrt{347.22 \times 16.11}}$$

(b) [2 points] An admissions officer proposes using the sum of high school rank and (ACT score multiplied by 3) as an admission criterion. Write a calculation using matrix multiplication to compute the sample variance of this quantity in terms of the matrix

$$\mathbb{V} = \begin{bmatrix} V_{11} & V_{12} & V_{13} \\ V_{21} & V_{2,2} & V_{23} \\ V_{31} & V_{32} & V_{33} \end{bmatrix} = \begin{bmatrix} 0.40 & 4.71 & 0.93 \\ 4.71 & 347.22 & 33.10 \\ 0.93 & 33.10 & 16.11 \end{bmatrix}$$

Solution. The sample variance is $\mathbb{A}\mathbb{V}\mathbb{A}^{\mathrm{T}}$ where $\mathbb{A} = \begin{bmatrix} 0 & 1 & 3 \end{bmatrix}$ is the row vector corresponding to the linear combination of $0 \times$ GPA $+ 1 \times$ High school rank $+ 3 \times$ ACT.

If you write this out in full, it is

$$\begin{bmatrix} 0 & 1 & 3 \end{bmatrix} \begin{bmatrix} V_{11} & V_{12} & V_{13} \\ V_{21} & V_{22} & V_{23} \\ V_{31} & V_{32} & V_{33} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}$$

The sample means of each variable are given by

```
mu <- apply(gpa,2,mean)
mu
```

```
##         GPA High_School         ACT
##    2.977315   76.953191   24.543262
```

(c) [1 point] Let $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)$ be a mathematical representation of `mu`. The sample mean of the sum of high school rank and (ACT score times 3) is $\mu_{\text{sum}} = \mu_2 + 3\mu_3$. Write a matrix multiplication that gives μ_{sum} in terms of $\boldsymbol{\mu}$.

Solution. Taking $\boldsymbol{\mu}$ to be a column vector, we have

$$\mu_{\text{sum}} = \begin{bmatrix} 0 & 1 & 3 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}$$

An alternative solution could represent $\boldsymbol{\mu}$ as a row vector.

(d) Suppose the standard deviation calculated in part (b) is called σ_{sum}. Let X be a normal random variable with mean μ_{sum} and standard deviation σ_{sum}. Write a probability statement using a normal approximation to find the chance that a random applicant will have a sum of high school rank plus (3 times ACT score) larger than 180. Write an integral that evaluates to this probability.

Solution.

$$\mathrm{P}(X > 180) = \int_{180}^{\infty} \frac{1}{\sqrt{2\pi\sigma_{\text{sum}}^2}} e^{-(x-\mu_{\text{sum}})^2/2\sigma_{\text{sum}}^2}\, dx$$

(e) Suppose σ_{sum} and μ_{sum} have been calculated in R and assigned variable names `sigma_sum` and `mu_sum`. Write an R expression to obtain numerically the probability you wrote in (d).

Solution. 1-pnorm(180,mean=mu_sum,sd=sigma_sum)
