Practice midterm exam, STATS 401 F18

Name:

UMID:

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.

You are not allowed to bring any notes into the exam.

The following formulas will be provided. To use these formulas properly, you need to make appropriate definitions of the necessary quantities.

(1) $\mathbf{b} = \big(\mathbb{X}^\top \mathbb{X}\big)^{-1} \mathbb{X}^\top \mathbf{y}$

(2) $\mathrm{Var}(\mathbb{A}\mathbf{Y}) = \mathbb{A}\mathrm{Var}(\mathbf{Y})\mathbb{A}^\top$

(3) $\mathrm{Var}(X) = \mathrm{E}\big[(X - \mathrm{E}[X])^2\big] = \mathrm{E}[X^2] - \big(\mathrm{E}[X]\big)^2$

(4) $\mathrm{Cov}(X,Y) = \mathrm{E}\big[\big(X - \mathrm{E}[X]\big)\big(Y - \mathrm{E}[Y]\big)\big] = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]$

(5) The probability density function of the standard normal distribution is $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

(6) If a random variable is normally distributed, the probability it falls within one standard deviation of the mean is 68

From `?pnorm`:

```
pnorm(q, mean = 0, sd = 1)
qnorm(p, mean = 0, sd = 1)
q:  vector of quantiles.
p:  vector of probabilities.
```

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.

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

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

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

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)
```

(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.
```

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

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

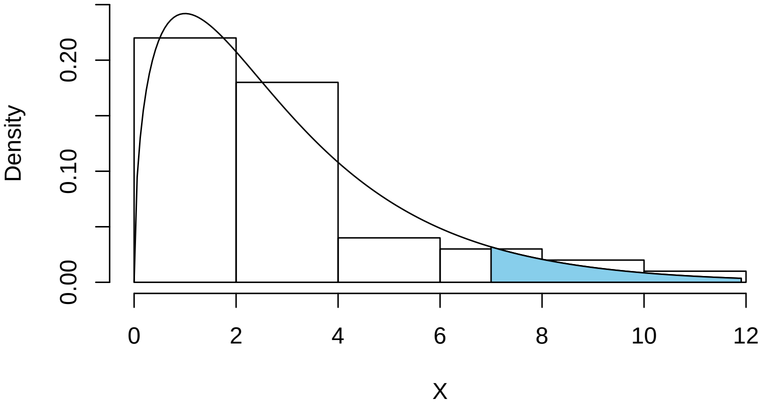
$$X_F = \frac{\sum_{i=1}^m Y_i^2/m}{\sum_{j=1}^n Z_j^2/n}$$

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.

(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!

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

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.

(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.

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

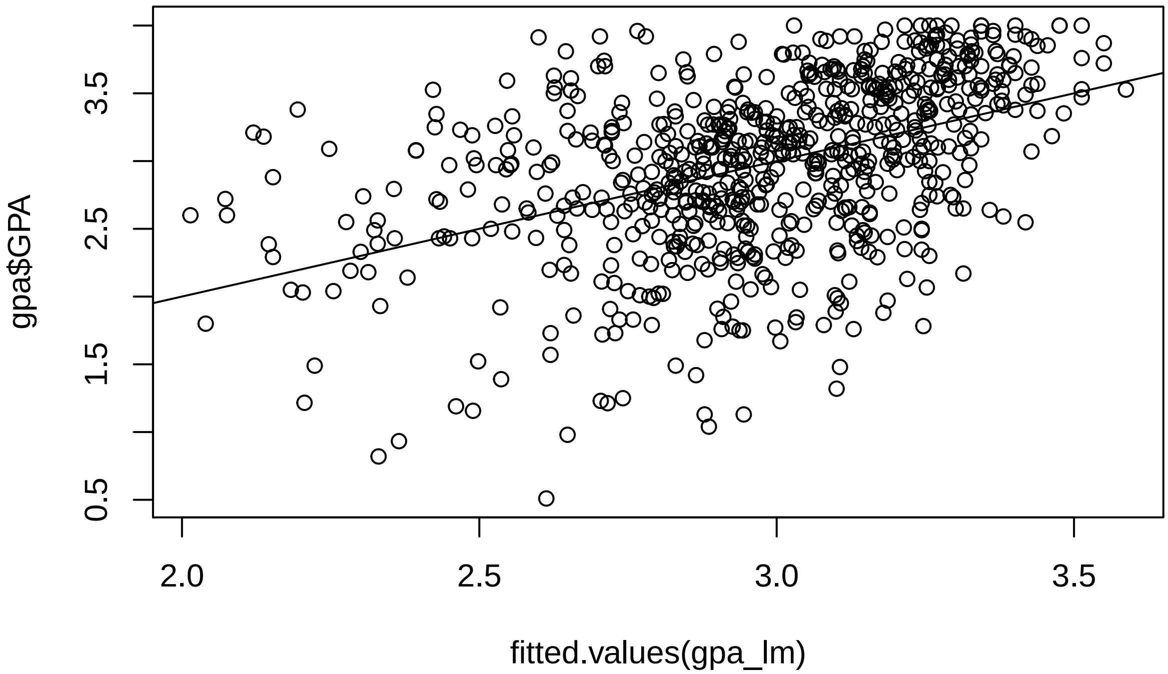
```r
gpa_lm <- lm(GPA~ACT+High_School,data=gpa)
summary(gpa_lm)
```

```
##
## 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.

(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)
```



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

The sample variance/covariance matrix for the GPA dataset is

```r
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.

(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

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

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 $\boldsymbol{\mu}$.

(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.

(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).