Midterm exam, STATS 401 W18

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. If you need extra paper, please number the pages and put your name and UMID on each page.

Formulas

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

(2) $\mathrm{Var}(\hat\beta) = \sigma^2\left(\mathbb{X}^\top\mathbb{X}\right)^{-1}$

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

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

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

(6) The binomial (n,p) distribution has mean np and variance $np(1-p)$.

From `?pnorm`:

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

Summation exercises

S1. A basic exercise.

Let $\mathbb{X} = [x_{ij}]$ be a 3×2 matrix with (i,j) entry given by $x_{ij} = 2i$.

(a) Write out $\mathbb{X}$, evaluating each of the six entries of the matrix.

Solution. [1 point]

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

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

Solution. [1 point]

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

S2. An example involving the summation representation of matrix multiplication.

Evaluate $\mathbb{X}^{\mathsf{T}}\mathbb{X}$ where

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

Solution. [2 points]

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

R exercises

R1. Using `rep()` and `matrix()`.

Write the output of

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

Solution. [2 points]

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

R2. Manipulating vectors and matrices in R.

Which of the following is the output to `pnorm(c(-2,2))`

a) `[1] 0.02275013 0.97724987`

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

c) `[1] 0.1586553 0.8413447`

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

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

Solution. [1 point] (a)

Properties of variance and covariance

V1. A numerical calculation to find the variance of a linear combination using matrix techniques.

Let $\mathbf{X} = (X_1, X_2)$ be a vector random variable with mean $(3, 4)$ and variance matrix

$$\mathbb{V} = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}.$$

Let $Y = X_1 - X_2$. Find a suitable matrix $\mathbb{A}$ for which $Y = \mathbb{A}\mathbf{X}$, noting that the random variable Y can be considered as a 1×1 matrix. Set up and solve a matrix calculation to find the variance of Y.

Solution. [2 points]

Set $\mathbb{A} = [1 \ \ -1]$. Then,

$$\text{Var}(X_1 - X_2) = \text{Var}(\mathbb{A}\mathbf{X}) = \mathbb{A}\text{Var}(X)\mathbb{A}^{\text{T}} = 3.$$

Note that the mean of $\mathbf{X}$ is irrelevant for this question.

V2. An algebraic calculation using basic definitions of variance & covariance, together with the linearity of expectation.

Use formulas (4) and (5) above, together with the linearity of expectation, to show that

$$\text{Var}(3X + Y + 4) = 9\text{Var}(X) + \text{Var}(Y) + 6\text{Cov}(X, Y)$$

.

Solution. [2 points]

First note that, from (4), adding a constant doesn't change the variance so $\text{Var}(3X + Y + 4) = \text{Var}(3X + Y)$. Therefore,

$$\begin{array}{rcl} \text{Var}(3X + Y + 4) & = & \text{E}[(3X + Y)^2] - (\text{E}[3X + Y])^2 \\ & = & \text{E}[9X^2 + Y^2 + 6XY] - (3\text{E}[X] + \text{E}[Y])^2 \\ & = & 9\text{E}[X^2] + \text{E}[Y^2] + 6\text{E}[XY] - \big\{9\text{E}[X]^2 + \text{E}[Y]^2 + 6\text{E}[X]\,\text{E}[Y]\big\} \\ & = & 9\big\{\text{E}[X^2] - \text{E}[X]^2\big\} + \big\{\text{E}[Y^2] - \text{E}[Y]^2\big\} + 6\big\{\text{E}[XY] - \text{E}[X]\,\text{E}[Y]\big\} \\ & = & 9\text{Var}(X) + \text{Var}(Y) + 6\text{Cov}(X, Y) \end{array}$$

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
```

F1. 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.

F2. 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.

F3. 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.

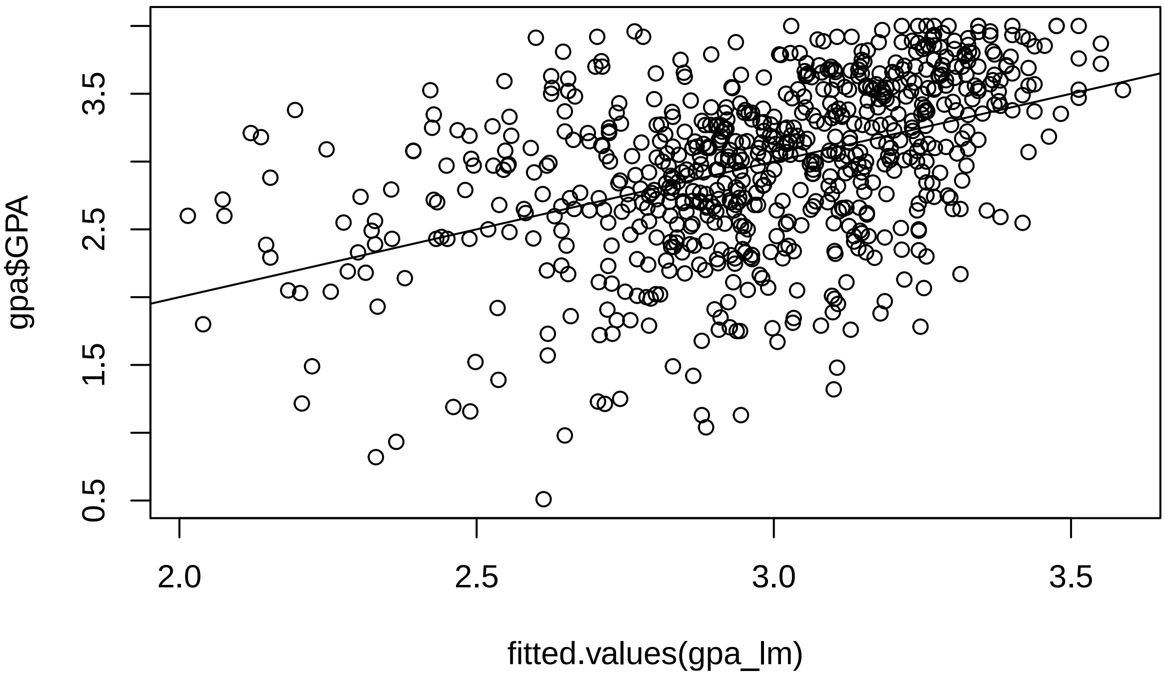
Solution. [2 points]

The coefficient estimates are the vector **b** from F2, calculated by least squares using the formula

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

F4. 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

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

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. (We will see later that 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.)

The population version (or probability version) of the linear model

P1. Write out a suitable probability model, in subscript form, to give a population version of the linear model for freshman GPA in question F3. 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 probability model is

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 + \epsilon_i, \quad i = 1, \ldots, n$$

where Y_i models freshman GPA 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 $n = 705$. The measurement error model is that $\epsilon_1, \ldots, \epsilon_n$ are independent random variables with mean zero and variance σ^2. β_1, β_2 and β_3 are unknown constants which model the relationship between the response and explanatory variables.

P2. Describe a suitable probability model, in matrix form, to give a population version of the linear model in question F3. 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 probability model (also known as the population model) is

$$\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

- $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is a vector random variable modeling 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 $i = 1, \ldots, n$. This exactly matches the definition in F2.
- $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ is the true but unknown vector of coefficients.
- $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ is a vector random variable modeling chance variation. It follows the measurement error model, with $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbb{I}$ where $\mathbb{I}$ is the $n \times n$ identity matrix and σ is the measurement standard deviation.
- All vectors are interpreted as column vectors.

P3. Explain how R produces standard errors for coefficients in a linear model. Also, describe in words how you interpret the standard error of 0.037210 for the coefficient of ACT.

Solution. [2 points]

The estimated residual standard error is computed as

$$s = \sqrt{\frac{1}{n-3}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

The variance of $\hat{\boldsymbol{\beta}}$ is

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}$$

which is estimated by

$$\mathbb{V} = s^2\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}.$$

Then, the standard error for b_1 is $\sqrt{\mathbb{V}_{11}}$. We interpret this as an estimate of the standard deviation of the estimates of the coefficient for ACT score in many draws from the probability model.

Normal probability calculations

N1. A normal approximation to estimate a probability using the mean and variance.

Suppose that an analysis of a large dataset from another comparable university gave a coefficient of 0.03528 for the ACT variable when fitting a linear model using ACT score and high school rank. The admissions director is interested whether the difference could reasonably be chance variation due to having only a sample of 705 students, or whether the universities have differences beyond what can be explained by sample variation. Suppose that population value for this school is also 0.03528. Supposing the probability model in P1 and P2, and using a normal approximation, find an expression for the probability that the difference between the coefficient estimate for the data (0.03721) and the hypothetical true value (0.03528) is larger in magnitude than the observed value (0.03721-0.03528). Write your answer as a call to `pnorm()`. Your call to pnorm may involve specifying any necessary numerical calculations that you can't work out without access to a computer or calculator.

Solution. [2 points] `1-pnorm(0.03721,mu=0.03538,sd=0.005939)` gives the probability of observing a bigger value of the estimated coefficient under the assumed model, making a normal approximation using the calculated standard error. By symmetry, the chance of the difference being larger in magnitude (i.e., too large or too small) is twice the chance of being bigger. So, the answer is `2(1-pnorm(0.03721,mu=0.03538,sd=0.005939))`

N2. A normal approximation to find a region with a given probability using the mean and variance.

Let $X_1, X_2, \ldots, X_n$ be independent random variables each of which take the value 0 with probability 0.5, 1 with probability 0.25 and -1 with probability 0.25. Find the mean and variance of X_1. Use this to find the mean and variance of $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Now suppose $n = 100$ and suppose that $\bar{X}$ is well approximated by a normal distribution. Find a number c such that $\mathrm{P}(-c < \bar{X} < c)$ is approximately 0.9. Write your answer as a call to `qnorm()`. Your call to qnorm may involve specifying any necessary numerical calculations that you can't work out without access to a computer or calculator.

Solution. [3 points]

$$\begin{aligned}
\mathrm{E}[X_1] &= 0 \times 0.5 + (-1) \times 0.25 + 1 \times 0.25 = 0 \\
\mathrm{E}[X_1^2] &= 0 \times 0.5 + (-1)^2 \times 0.25 + 1^2 \times 0.25 = 1 \\
\mathrm{Var}(X_1) &= \mathrm{E}[X_1^2] - \big(\mathrm{E}[X_1]\big)^2 = 1 \\
\mathrm{E}[\bar{X}] &= \mathrm{E}[X_1] = 0 \\
\mathrm{Var}(\bar{X}) &= \tfrac{1}{100^2} \times 100\mathrm{Var}(X_1) = 0.01
\end{aligned}$$

Then, making a normal approximation, we can call `c <- qnorm(0.95,mu=0,sd=0.1)` to obtain a value c with a probability of 0.05 to the right. By symmetry, this also has a probability of 0.05 to the left of -c, giving a probability of 0.9 for $[-c, c]$.
