Midterm exam, STATS 401 F18

Instructions. You have a time allowance of 80 minutes. The exam is closed book and closed notes. 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.

Consider the pair of simultaneous linear equations,

$$\begin{array}{rcrcl} 3b_1 & - & 2b_2 & = & 4 \\ b_1 & + & 3b_2 & = & 2 \end{array}$$

(a) [1 point] Write these linear equations in the matrix form $\mathbb{A}\mathbf{b} = \mathbf{c}$.

Solution.

$$\begin{bmatrix} 3 & -2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$$

(b) [2 points] Find the inverse matrix $\mathbb{A}^{-1}$ and make a calculation to check that your answer is correct.

Solution.

$$\begin{bmatrix} 3 & -2 \\ 1 & 3 \end{bmatrix}^{-1} = \frac{1}{11} \begin{bmatrix} 3 & 2 \\ -1 & 3 \end{bmatrix}$$

And we check

$$\frac{1}{11} \begin{bmatrix} 3 & 2 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} 3 & -2 \\ 1 & 3 \end{bmatrix} = \frac{1}{11} \begin{bmatrix} 11 & 0 \\ 0 & 11 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

(c) [2 points] Use matrix methods to solve the simultaneous equations for b_1 and b_2. This requires solving part (b) but partial credit will be available if you explain how to do this without successfully completing (b).

Solution. $\mathbf{b} = \mathbb{A}^{-1}\mathbf{c}$, so in this case we get

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \frac{1}{11} \begin{bmatrix} 3 & 2 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} 4 \\ 2 \end{bmatrix} = \frac{1}{11} \begin{bmatrix} 16 \\ 2 \end{bmatrix} = \begin{bmatrix} 16/11 \\ 2/11 \end{bmatrix}$$

It is acceptable to either bring the factor of 1/11 inside the vector or leave it outside. It is also acceptable to end by writing values of b_1 and b_2 separately rather than leaving b_1 and b_2 in a column vector.

Q2. R exercises.

Define the matrix `A` in R as:

```
A = matrix(c(0,1,1,3,3,2),nrow = 3)
A
```

```
##      [,1] [,2]
## [1,]    0    3
## [2,]    1    3
## [3,]    1    2
```

(a) [1 point]. What is the output of `apply(A,2,mean)`?

(i). A vector of length 3 corresponding to the average of each row of `A`.

(ii). A vector of length 2 corresponding to the average of each column of `A`.

(iii). The mean of all the values in `A`.

(iv). The mean of the second column of `A`.

(v). The mean of the second row of `A`.

Solution:

(ii). A vector of length 2 corresponding to the average of each column of `A`.

(b) [3 points]. For each of the lines of code below, say whether it will correctly make 50 draws from the normal$(100, 20)$ distribution. Among the correct answers, comment briefly on some strengths and weaknesses from the perspective of writing good R code. Which answer do you think is the best code, and why?

(i) `rnorm(50,20,100)`

(ii) `rnorm(100,20,50)`

(iii) `rnorm(100,20,n=50)`

(iv) `rnorm(mean=100,sd=20,n=50)`

(v) `rnorm(n=50,mean=100,sd=20)`

(vi) `replicate(rnorm(100,20),50)`

(vii) `replicate(rnorm(n=1,mean=100,sd=20),n=50)`

(viii) `rnorm(50)*20+100`

(ix) `100+sqrt(20)*rnorm(50)`

Solution:

(iii), (iv), (v), (vii), (viii) are all correct.

(i), (ii), (vi) make incorrect assumptions about how R matches arguments and how R chooses default arguments when they are not provided. The convenience of default arguments, and matching arguments by position or name, comes at a price. If in doubt, use named arguments.

(ix) has the wrong scaling. Recall $\mathrm{SD}(aX) = a\,\mathrm{SD}(X)$ and $\mathrm{Var}(aX) = a^2\,\mathrm{Var}(X)$.

The easiest to read is probably (v). Arguments are labeled but also appear in the order matching the default, for familiarity.

```
head( rnorm(50,20,100) )
```

```
## [1] -128.056759  177.716947  -75.674448  -72.000525 -179.764210   -7.229604
```

```
head( rnorm(100,20,50) )
```

```
## [1]  17.86575  14.36647  42.84136 121.01674 -32.54450  56.73261
```

```
head( rnorm(100,20,n=50) )
```

```
## [1] 105.04680 112.84629  93.06128  89.69474 107.96338 119.07114
```

```
head( rnorm(mean=100,sd=20,n=50) )
```

```
## [1]  40.27088  98.98905 125.99026  72.97198  81.00402 133.72204
```

```
head( rnorm(n=50,mean=100,sd=20) )
```

```
## [1]  90.90273  98.23211 100.79325 110.94140  95.48626 107.87313
```

```
dim( replicate(50,rnorm(100,20)) )
```

```
## [1] 100  50
```

```
head( replicate(rnorm(n=1,mean=100,sd=20),n=50) )
```

```
## [1] 128.09837 113.73778  90.19083 118.01821 132.02530 135.67920
```

```
head( rnorm(50)*20+100 )
```

```
## [1] 106.72071  81.46925  94.54331 124.45090 130.74709 114.87494
```

```
head( 100+sqrt(50)*rnorm(50) )
```

```
## [1] 101.78136 108.95946 114.68784 100.15610 100.22505  98.00305
```

Q3. Investigating a probability model.

The value of a large public company is determined by its share price. The share price varies daily as people buy and sell shares in the company. Day-to-day changes in the share price are commonly modeled using normal random variables.

(a) [2 points]. Suggest some reasons why a normal distribution might be appropriate to model daily changes in the share price of a large company. Also, comment on limitations that you think the normal distribution might have for modeling a company share price. You are not expected to have detailed knowledge of the stock market. You are expected to apply to this situation your understanding about how a normal distribution arises in data.

(b) [4 points]. Let X and Y be bivariate random variables modeling the daily change in price of two different companies. Suppose X and Y are bivariate normal with respective means $\mu_X = \mu_Y = 0$, standard deviations $\sigma_X = 1$ and $\sigma_Y = 2$ and correlation $\mathrm{Cor}(X, Y) = 0.5$. Find the distributions of $X + Y$ and $X - Y$.

Solution. $X + Y$ and $X - Y$ are both normally distributed since they are linear combinations of normal random variables (see homework 5). We have $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y] = 0 + 0 = 0$ and $\mathrm{E}[X - Y] = \mathrm{E}[X] - \mathrm{E}[Y] = 0 - 0 = 0$. Also,

$$\begin{aligned} \mathrm{Var}(X + Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y) \\ &= \sigma_X^2 + \sigma_Y^2 + 2\,\mathrm{Cor}(X, Y)\,\sigma_X\,\sigma_Y = 1 + 4 + 2 \times 0.5 \times 1 \times 2 = 7. \end{aligned}$$

Then,

$$\begin{aligned} \mathrm{Var}\big(X + (-Y)\big) &= \mathrm{Var}(X) + \mathrm{Var}(-Y) + 2\,\mathrm{Cov}(X, -Y) \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\,\mathrm{Cov}(X, Y) \\ &= \sigma_X^2 + \sigma_Y^2 + 2\,\mathrm{Cor}(X, Y)\,\sigma_X\,\sigma_Y = 1 + 4 - 2 \times 0.5 \times 1 \times 2 = 3. \end{aligned}$$

(c) [1 point]. Write an integral representing the probability that $X > Y$. Hint: you can use the answer from part (b).

Solution.

$$\mathrm{P}(X > Y) = \mathrm{P}(X - Y > 0) = \int_0^\infty \frac{1}{\sqrt{2\pi \times 3^2}} e^{-x^2/(2\times 3^2)}\, dx.$$

(d) [1 point]. If you can find an exact value for the integral you wrote in (c), give it. Otherwise, write R code to evaluate it.

Solution. Since the normal curve is symmetric about the line $x = 0$, the area under the region with $x > 0$ and $x < 0$ are both $1/2$. In this case, the answer can be checked by `pnorm(0,mean=0,sd=sqrt(3))`.

Q4. Fitting a linear model by least squares

This question concerns the US unemployment and life expectancy data seen in class and homework. Let u_i be annual percent unemployment in the ith year and let ℓ_i be the life expectancy for the corresponding year, with i ranging over years 1948-2015 for which both variables are available. The data are entered into R as a dataframe called `health`.

```
head(health)
```

```
##   year        u     l
## 1 1948 3.766667 67.25
## 2 1949 5.908333 67.63
## 3 1950 5.325000 68.07
## 4 1951 3.333333 68.17
## 5 1952 3.033333 68.39
## 6 1953 2.925000 68.72
```

Recall that the first step in our analysis was to detrend the data, by fitting a line to the plots of life expectancy and unemployment against time. Let x_i be the detrended unemployment and y_i the detrended life expectancy. These are computed in R as

```
x <- lm(u~year,data=health)$residuals
y <- lm(l~year,data=health)$residuals
```

(a) [2 points]. What is the purpose of detrending the data before looking for a linear relationship between life expectancy and unemployment rate?

Solution. Life expectancy has a clear upward trend over this time period. To tell whether life expectancy is influenced by economic conditions, we can look at fluctuations around the trend, to see whether good (or bad) economic conditions are associated with above trend (or below trend) life expectancy.

(b) [3 points]. Write in subscript form the linear model used to detrend life expectancy. (The model for detrending unemployment is similar, but you are not required to write that out too.)

Solution. The model is

$$\ell_i = b_1 x_i + b_2 + e_i, \quad i = 1, \ldots, n$$

where ℓ_i is life expectancy for the ith row of the dataset, $x_i = 1948 + i - 1$ is the corresponding calendar year, and $n = 2015 - 1948 + 1 = 68$ is the number of years of data. e_i is the residual error for year i. b_1 and b_2 are coefficients chosen by least squares. It is almost equivalent to write the model as

$$\ell_i = b_1 i + b_2 + e_i, \quad i = 1, \ldots, n$$

but the question technically asks for the explanatory variable to be x_i not i. Both models give the same fitted values.

(c) [2 points]. Write the design matrix for representing your linear model from (b) in matrix form.

Solution. Let $x_i = 1948 + i - 1$ for $i = 1, \ldots, n$ with $n = 2015 - 1948 + 1 = 68$. The design matrix is

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

The relationship between fluctuations in life expectancy and fluctuations in unemployment was then investigated by fitting the following model:

```
lm1 <- lm(y~x)
summary(lm1)$coef
```

```
##                  Estimate Std. Error      t value     Pr(>|t|)
## (Intercept) 3.046417e-17 0.04406620 6.913274e-16 1.000000e+00
## x           1.313673e-01 0.02981692 4.405798e+00 3.959878e-05
```

(d) [3 points]. What is the principle used to calculate the coefficients of the linear model? Write a matrix calculation to implement this calculation. Any vectors and matrices you use should be defined, though you can refer back to any definitions you may have made in earlier parts of this question.

Solution. The coefficients are calculated by least squares, meaning they are chosen to minimize $\sum_{i=1}^{n} e_i^2$. This is carried out by a matrix calculation

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

where $\mathbb{X}$ is the design matrix from (c) and $\mathbf{y}$ is a column vector of the data.

(e) [2 points]. How can a probability model help us to interpret the estimated coefficients? You are not yet expected to have studied in detail the probability model for the linear model. Here, you are only asked to discuss in words the goals and purposes of having a probability model for this data analysis.

Solution. A probability model lets us make probability statements about alternative outcomes we might hypothetically have observed. For example, we could ask what is the probability we would have observed a coefficient as large as the observed 0.131 if in fact there is no association between percentage unemployment and life expectancy. Also, a probability model lets us calculate standard errors for the estimated coefficients.

Q5. Working with means, variances and covariances.

This question concerns the dataset on global climate change studied in Homework 5. Carbon dioxide (CO_2) levels in the atmosphere have been increasingly steadily, as recorded by the measurements taken at Mauna Loa observatory in Hawaii. An increasing trend in CO_2 matches increasing trends in both global economic activity and the global population, as well as many other socioeconomic phenomena. However, on shorter timescales, fluctuating geophysical processes such as volcanic activity and the El Nino Southern Oscillation (ENSO) may be important. The first three years and the last year of the dataset contain some missing values so are removed for this analysis using the following R code.

```
X <- read.table("climate.txt",header=TRUE)[-c(1:3,54),]
X$Pop <- X$Pop/1000 # rescale population from millions to billions
head(X)
```

```
##   Year    CO2   GDP   Pop    ENSO Volcanic Emissions
## 4 1961 317.64  7.54 3.069 -0.2322   0.0024       9.5
## 5 1962 318.45  7.97 3.123 -0.7650   0.0024       9.8
## 6 1963 318.99  8.38 3.189 -0.1629   1.8454      10.4
## 7 1964 319.62  8.95 3.255 -0.4784   2.1758      11.0
## 8 1965 320.04  9.45 3.323  0.3252   0.9864      11.5
## 9 1966 321.38 10.03 3.393  0.5084   0.3699      12.1
```

- CO2: Mean annual concentration of atmospheric CO2 (parts per million by volume) at Mauna Loa.

- GDP: world gross domestic product reported by the World Bank.
- Pop: world population, in billions, reported by the World Bank.
- ENSO: an El Nino Southern Oscillation index from NOAA.
- Volcanic: an index of monthly estimated sulfate aerosols derived from NOAA.
- Emissions: estimated emissions of CO2 (million Kt) reported by the World Bank.

Let $\mathbb{V} = [V_{ij}]_{p\times p}$ be the sample variance/covariance matrix for these $p = 7$ variables, computed as

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

```
##              Year    CO2    GDP   Pop ENSO Volcanic Emissions
## Year       212.50 316.01 144.37 16.75 2.31    -2.52     89.16
## CO2        316.01 474.95 217.34 24.94 3.01    -3.98    133.03
## GDP        144.37 217.34  99.89 11.39 1.21    -1.99     61.23
## Pop         16.75  24.94  11.39  1.32 0.18    -0.20      7.01
## ENSO         2.31   3.01   1.21  0.18 0.54     0.19      0.83
## Volcanic    -2.52  -3.98  -1.99 -0.20 0.19     0.60     -1.27
## Emissions   89.16 133.03  61.23  7.01 0.83    -1.27     38.73
```

(a) [1 point] Give a formula that computes the entry V_{12} in terms of the data matrix $\mathbb{X} = [x_{ij}]_{n\times p}$.

Solution. Write $\mathbb{X} = [\mathbf{x}_1 \dots \mathbf{x}_p]$ where $\mathbf{x}_j = (x_{1j}, \dots, x_{nj})$ interpreted as a column vector with $p = 7$ and $n = 50$.

$$V_{12} = \operatorname{cov}(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{n-1}\sum_{i=1}^{n}\big(x_{i1} - \bar{x}_1\big)\big(x_{i2} - \bar{x}_2\big)$$

where $\bar{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{i1}$ and $\bar{x}_2 = \frac{1}{n}\sum_{i=1}^{n} x_{i2}$.

(b) [2 points] Let $\mathbf{b} = (b_1, \dots, b_7)$ be a vector, interpreted as a column vector. We can use $\mathbf{b}$ to construct a linear combination $\mathbb{X}\mathbf{b}$ with ith row equal to $b_1x_{i1} + b_2x_{i2} + \cdots + b_7x_{i7}$ for $i = 1, \dots, n$. Here, b_1, $b_2, \dots, b_7$ represent numbers which are not necessariy positive. Is it possible to find a choice of b_1, $b_2, \dots, b_7$ to get a negative value of $\mathbf{b}^{\mathrm{T}}\mathbb{V}\mathbf{b}$? Explain.

Solution. $\mathbf{b}^{\mathrm{T}}\mathbb{V}\mathbf{b}$ is the sample variance of the linear combination $\mathbb{X}\mathbf{b}$. This is the sample version of the formula $\operatorname{Var}(\mathbb{A}\mathbf{Y}) = \mathbb{A}\operatorname{Var}(\mathbf{Y})\mathbb{A}^{\mathrm{T}}$. Since $\mathbf{b}^{\mathrm{T}}\mathbb{V}\mathbf{b}$ is a sample variance, it cannot be negative.

Now we look at the sample correlation matrix $\mathbb{C} = [C_{ij}]_{p\times p}$.

```
C <- cor(X)
round(C,2)
```

```
##              Year   CO2   GDP   Pop ENSO Volcanic Emissions
## Year         1.00  0.99  0.99  1.00 0.22    -0.22      0.98
## CO2          0.99  1.00  1.00  1.00 0.19    -0.24      0.98
## GDP          0.99  1.00  1.00  0.99 0.16    -0.26      0.98
## Pop          1.00  1.00  0.99  1.00 0.21    -0.22      0.98
## ENSO         0.22  0.19  0.16  0.21 1.00     0.34      0.18
## Volcanic    -0.22 -0.24 -0.26 -0.22 0.34     1.00     -0.26
## Emissions    0.98  0.98  0.98  0.98 0.18    -0.26      1.00
```

(c) [1 point] Give a formula using subscript notation that computes the entry C_{ij} in terms of the entries of $\mathbb{V} = [V_{ij}]_{p\times p}$.

Solution.

$$C_{ij} = \operatorname{cor}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\operatorname{cov}(\mathbf{x}_i, \mathbf{x}_j)}{\sqrt{\operatorname{var}(\mathbf{x}_i)\operatorname{var}(\mathbf{x}_j)}} = \frac{V_{ij}}{\sqrt{V_{ii}V_{jj}}}$$

(d) [2 points] In the table above, the variables `Year`, `CO2`, `GDP`, `Pop` and `Emissions` all have correlations close to 1. Explain your interpretation of this result. Also, explain why this is not very insightful and propose another way to analyze the data to resolve this problem.

Solution. Variables with a correlation close to 1 have a scatterplot with points lying close to a straight line with positive slope. What we learn from this is that these variables all share a similar trend. We probably knew in advance that all these variables have been increasing steadily over the past 50 years. Looking at fluctuations around the trend, for example by taking differences or looking at residuals after fitting a linear model with a trend, might be more insightful.

(e) **[Optional extra credit, 1 point]**. Write a matrix calculation using R code to obtain the sample correlation matrix `C` from the sample variance/covariance matrix `V`. As well as usual matrix and vector operations, you may use other R functions including

(i) `diag(M)` which returns a vector of the diagonal entries when `M` is a square matrix.

(ii) `sqrt(M)` which returns the elementwise square root of a matrix `M`.

Solution: One way to get a matrix of all the products of variances needed to turn covariances into correlations is `diag(V) %*% t(diag(V))`. We can then carry out elementwise division to rescale all the entries in the variance/covariance matrix into correlations, as follows.

```
C <- V / sqrt(diag(V)%*% t(diag(V)))
round(C,2)
```

```
##             Year   CO2   GDP   Pop ENSO Volcanic Emissions
## Year        1.00  0.99  0.99  1.00 0.22    -0.22      0.98
## CO2         0.99  1.00  1.00  1.00 0.19    -0.24      0.98
## GDP         0.99  1.00  1.00  0.99 0.16    -0.26      0.98
## Pop         1.00  1.00  0.99  1.00 0.21    -0.22      0.98
## ENSO        0.22  0.19  0.16  0.21 1.00     0.34      0.18
## Volcanic   -0.22 -0.24 -0.26 -0.22 0.34     1.00     -0.26
## Emissions   0.98  0.98  0.98  0.98 0.18    -0.26      1.00
```
