Quiz 1, STATS 401 W18

In lab on 10/5

This document produces different random quizzes each time the source code generating it is run. The actual quiz will be a realization generated by this random process, or something similar.

This version lists all the questions currently in the quiz generator. The quiz will have one question drawn at random from each of the five categories. No new questions will be added after Wednesday 10/3. Small changes may be made.

Instructions. You have a time allowance of 40 minutes, though the quiz may take you less time and you can leave lab once you are done. The quiz is closed book, and you are not allowed access to any notes. Any electronic devices in your possession must be turned off and remain in a bag on the floor.

Formulas

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

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

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

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

(4) Syntax from `?pnorm`:

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

(5) If a random variable is normally distributed, the probability it falls within one standard deviation of the mean is 68%, within two standard deviations of the mean is 95%, and within three standard deviations of the mean is 99.7%.

Q1. Matrix exercises

Q1-1.

(a). Evaluate $\mathbb{AB}$ when

$$\mathbb{A} = \begin{bmatrix} 2 & -2 \\ 3 & 2 \\ -1 & 1 \end{bmatrix}, \quad \mathbb{B} = \begin{bmatrix} 2 & -2 \\ 3 & -1 \end{bmatrix}$$

Solution:

$$\mathbb{AB} = \begin{bmatrix} -2 & -2 \\ 12 & -8 \\ 1 & 1 \end{bmatrix}$$

(b). For $\mathbb{A}$ as above, write down $\mathbb{A}^{\mathrm{T}}$.

Solution:

$$\mathbb{A}^{\mathrm{T}} = \begin{bmatrix} 2 & 3 & -1 \\ -2 & 2 & 1 \end{bmatrix}$$

(c). For $\mathbb{B}$ as above, find $\mathbb{B}^{-1}$ if it exists. If $\mathbb{B}^{-1}$ doesn't exist, explain how you know this.

Solution:

$$\mathbb{B}^{-1} = \frac{1}{4}\begin{bmatrix} -1 & 2 \\ -3 & 2 \end{bmatrix}$$

Q1-2.

(a). Evaluate $\mathbb{AB}$ when

$$\mathbb{A} = \begin{bmatrix} 0 & 1 & -1 \\ -1 & -1 & 0 \end{bmatrix}, \quad \mathbb{B} = \begin{bmatrix} 3 & 3 & -1 \\ 1 & 2 & -2 \\ 3 & -1 & 0 \end{bmatrix}$$

Solution:

$$\mathbb{AB} = \begin{bmatrix} -2 & 3 & -2 \\ -4 & -5 & 3 \end{bmatrix}$$

(b). For $\mathbb{A}$ as above, write down $\mathbb{A}^{\mathrm{T}}$.

Solution:

$$\mathbb{A}^{\mathrm{T}} = \begin{bmatrix} 0 & -1 \\ 1 & -1 \\ -1 & 0 \end{bmatrix}$$

(c). For $\mathbb{A}$ as above, find $\mathbb{A}^{-1}$ if it exists. If $\mathbb{A}^{-1}$ doesn't exist, explain how you know this.

Solution:

Only square matrices can be invertible. $\mathbb{A}$ is 2×3 and so cannot have an inverse.

Q2. Summation exercises

Q2-1.

Calculate $\sum_{i=k}^{k+3}(i+3)$, where k is a whole number. Your answer should depend on k.

Solution:

$\sum_{i=k}^{k+3}(i+3) = (k+3) + [(k+1)+3] + [(k+2)+3] + [(k+3)+3] = 4k+18.$

Q2-2.

Evaluate $\sum_{i=1}^{30} 10 - \sum_{i=10}^{20} 20$.

Solution:

$\sum_{i=1}^{30} 10 - \sum_{i=10}^{20} 20 = 30 \times 10 - 11 \times 20 = 300 - 220 = 80.$

Q2-3.

Calculate $\sum_{k=m}^{n} a$, where m and n are whole numbers and a is a real number.

Solution:

$\sum_{k=m}^{n} a = (n-m+1)a$ since the sum has $(n-m+1)$ terms each of which is a.

Q2-4.

Evaluate $3\sum_{k=1}^{5} 2 - 0.5\sum_{i=2}^{11} 6$.

Solution:

$3\sum_{k=1}^{5} 2 - 0.5\sum_{i=2}^{11} 6 = 3 \times 10 - 0.5 \times 60 = 0$

Q2-5.

Evaluate $\sum_{i=1}^{3} i(i-1)$.

Solution:

$\sum_{i=1}^{3} i(i-1) = \sum_{i=1}^{3}(i^2 - i) = \sum_{i=1}^{3} i^2 - \sum_{i=1}^{3} i = (1+4+9) - (1+2+3) = 14 - 6 = 8.$

Q2-6.

Suppose $F_0 = 0, F_1 = 1, F_2 = 1, F_3 = 2, F_4 = 3, F_5 = 5, F_6 = 8, F_7 = 13$ Evaluate $\sum_{i=4}^{7} F_i - \sum_{i=0}^{3} F_i$.

Solution

$$\begin{aligned}\sum_{i=4}^{7} F_i - \sum_{i=0}^{3} F_i &= (3+5+8+13) - (0+1+1+2) \\ &= 29 - 4 = 25\end{aligned}$$

Q3. R exercises

Q3-1.

(a) Which of the following is the output of `matrix(c(rep(0,times=4),rep(1,times=4)),ncol=2)`

(i). $\begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$ (ii). $\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$ (iii). $\begin{bmatrix} 0 & 0 \\ 1 & 1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}$ (iv). $\begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$

Solution:

(i), since R fills matrices by columns.

(b) Suppose we define an R vector by `y <- c(3,NA,-1,4,NA,-2)`. What will `y[y>0]` give you?

(i). A vector of the positive elements and NA values of `y`.

(ii). A vector of the negative elements of `y`.

(iii). A vector of all `NA`s.

(iv). A vector of `TRUE`s and `FALSE`s.

(v). A vector of `TRUE`s and `FALSE`s and `NA`s.

Solution:

(i). Indexing by `y>0` should pick out positive terms, but `NA` terms remain `NA` since R cannot tell if missing data are positive.

Q3-2.

(a) Which one of the following lines of code successfully constructs the matrix $\mathbb{A} = \begin{bmatrix} 1 & 1 \\ 2 & 2 \\ 3 & 3 \end{bmatrix}$

(i). `A <- matrix(c(1,1,2,2,3,3) ,nrow=3)`

(ii). `A <- cbind(c(1,1),c(2,2),c(3,3))`

(iii). `A <- t(matrix(c(1,1,2,2,3,3) ,nrow=2))`

(iv). `A <- c(c(1:3),c(1:3))`

Solution:

(iii)

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

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

```
cbind(c(1,1),c(2,2),c(3,3))
```

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

```
t(matrix(c(1,1,2,2,3,3) ,nrow=2))
```

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

```
c(c(1:3),c(1:3))
```

```
## [1] 1 2 3 1 2 3
```

(b) Suppose `X` is a matrix in R. Which of the following is NOT equivalent to `X`?

(i). `t(t(X))`

(ii). `X %*% matrix(1,ncol(X)`

(iii). `X*1`

(iv). `X%*%diag(ncol(X))`

Solution:

(ii). for (i), the transpose of the transpose gets back the original matrix. (iii) is elementwise multiplication by 1. (iv) is matrix multiplication with the identity matrix.

Q3-3.

(a) Which of the following is the matrix $\mathbb{A}$ generated by

```
A <- t(matrix(c(rep(1,times=2),rep(3,times=2), 6, 4),ncol=3))
```

(i) $\mathbb{A} = \begin{bmatrix} 1 & 1 \\ 3 & 3 \\ 6 & 4 \end{bmatrix}$

(ii) $\mathbb{A} = \begin{bmatrix} 1 & 3 & 6 \\ 1 & 3 & 4 \end{bmatrix}$

(iii) $\mathbb{A} = \begin{bmatrix} 1 & 3 \\ 1 & 6 \\ 1 & 3 \end{bmatrix}$

(iv) $\mathbb{A} = \begin{bmatrix} 1 & 1 & 3 \\ 3 & 6 & 4 \end{bmatrix}$

Solution:

(i). The matrix generated by `matrix(c(rep(1,times=2),rep(3,times=2), 6, 4),ncol=3)` is answer (ii) and the code then transposes this.

(b) Which of the following successfully select the first five odd elements of the vector `x <- c(1,2,3,4,5,6,7,8,9,10,11)`? (List all that apply. Do not list commands that will give an error)

(i) `x[rep(c(TRUE,FALSE),each=5)]`

(ii) `x[rep(c(TRUE,FALSE),times=5)]`

(iii) `x[rep(c(TRUE,FALSE),length=9)]`

(iv) `x[rep(c(TRUE,FALSE))][1:5]`

(v) `x[rep(c("TRUE","FALSE"),5)]`

(vi) None of the above

(vii) All of the above

Solution:

Only (iv). Here's what they give:

```
x <- c(1,2,3,4,5,6,7,8,9,10,11)
x[rep(c(TRUE,FALSE),each=5)]
```

```
## [1]  1  2  3  4  5 11
```

```
x[rep(c(TRUE,FALSE),times=5)]
```

```
## [1]  1  3  5  7  9 11
```

```
x[rep(c(TRUE,FALSE),length=9)]
```

```
## [1]  1  3  5  7  9 10
```

```
x[rep(c(TRUE,FALSE))][1:5]
```

```
## [1] 1 3 5 7 9
```

```
x[rep(c("TRUE","FALSE"),5)]
```

```
##  [1] NA NA NA NA NA NA NA NA NA NA
```

Q3-4.

(a) Define the matrix `A` as:

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

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) 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 $\text{SD}(aX) = a\text{SD}(X)$ and $\text{Var}(aX) = a^2\text{Var}(X)$.

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]  -87.176004    6.101386  -39.731309 -198.396676   44.081726   -5.935541
```

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

```
## [1] -18.002897   2.930686 -85.116456   4.914886 -43.619172   6.016695
```

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

```
## [1] 89.78165 86.81238 99.19620 97.62612 99.60686 90.28643
```

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

```
## [1] 116.47430  91.44424  97.14712 128.37566 109.74268 112.06883
```

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

```
## [1] 101.49610 143.94859 115.90046  89.22116  67.97434  85.37253
```

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

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

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

```
## [1] 124.82627 121.49103  94.74586  96.63817 110.97790 126.16972
```

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

```
## [1] 115.31483  69.47221  90.25799  66.67923 106.03583  99.82000
```

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

```
## [1] 102.91579 103.37720  96.58727 102.89408 110.50044 101.53718
```

Q3-5.

(a) Which of the following successfully select the diagonal elements of the matrix $\mathbb{A} = \begin{bmatrix} 1 & 0 \\ 2 & 2 \end{bmatrix}$ represented in R by `A<-matrix(c(1,2,0,2),2,2)`?

(i). `A[c(1,1),c(2,2)]`

(ii). `A[rbind(c(1,1),c(2,2))]`

(iii). `A[cbind(c(1,1),c(2,2))]`

(iv). `A[matrix(c(TRUE,FALSE,FALSE,TRUE),2)]`

(v). all of (i,ii,iii,iv)

(vi). none of (i,ii,iii,iv)

(vii). (ii) and (iv) only

(viii). (i) and (ii) only

Solution:

We see below that the answer is (vii) [(ii) and (iv) only]. Indexing a matrix by a $n \times 2$ numeric vector makes a vector with elements picked out by each row of the indexing vector. Indexing a matrix by a logical matrix picks out the `TRUE` terms. (i) is somewhat unusual: this form picks out all the rows and columns identified, with possible replication. Try, for example, `A[c(1,1,1),c(1,2,1,2)]`.

```
A <- matrix(c(1,2,0,2),2,2)
A[c(1,1),c(2,2)]
```

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

```
A[rbind(c(1,1),c(2,2))]
```

```
## [1] 1 2
```

```
A[cbind(c(1,1),c(2,2))]
```

```
## [1] 0 0
```

```
A[matrix(c(TRUE,FALSE,FALSE,TRUE),2)]
```

```
## [1] 1 2
```

(b) Suppose we define a vector `x <- c(3,0,-1,4,0,-2)`. What will `which(x==0)` give you?

(i). A vector of the 0 elements of x.

(ii). A vectors of 0's.

(iii). A vector of `TRUE`'s and `FALSE`'s.

(iv). The vector of the indices of the 0 values.

Solution:

(iv).

```
x <- c(3,0,-1,4,0,-2)
which(x==0)
```

```
## [1] 2 5
```

Q3-6.

(a) Define the matrix `A` as:

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    6    4    3    4
## [2,]    2    1    1    3    1
## [3,]    3    2    1    6    2
```

What is the output of `apply(A[,-1],1,sd)`?

(i). A vector of length 4 corresponding to the standard deviation of each column of `A`, excluding the first column.

(ii). A vector of length 3 corresponding to the standard devation of each row of `A`, excluding the first column.

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

(iv). The standard devation of the first row of `A`.

(v). An error since `A` doesn't have a -1 column.

Solution:

(ii). A vector of length 3 corresponding to the standard devation of each row of `A`, excluding the first column.

(b) Which of the following lines of code successfully contructs the matrix for part (a)? Comment on the strengths and weaknesses of the correct answers.

(i). cbind(c(1,2,3), c(6,1,2), c(4,1,1), c(3,3,6), c(4,1,2))

(ii). matrix(cbind(c(1,2,3), c(6,1,2), c(4,1,1), c(3,3,6), c(4,1,2)))

(iii). t(matrix(c(1,6,4,3,4,2,1,1,3,1,3,2,1,6,2),nrow = 3))

(iv). matrix(c(1,2,3,6,1,2,4,1,1,3,3,6,4),nrow = 3)

Solution:

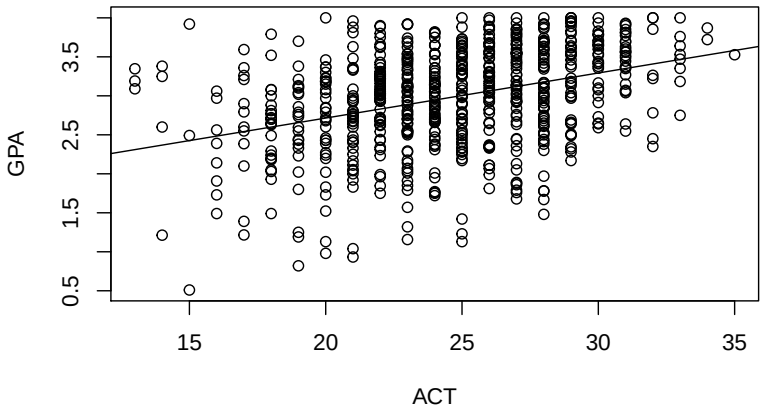
(i) and (iv) are both correct. (i) is the easiest to follow since the columns are layed out directly in the code. However, (iv) is easy to type and still relatively easy to follow if one is familiar with the matrix function.

Q4. Fitting a linear model by least squares

Q4-1.

The admissions officer at a large state university wants to assess how well academic success can be predicted based on information available at admission. She collects data on freshman GPA and highschool ACT exam scores for 705 students in an R dataframe called `gpa`. The plot below shows a line fitted to a scatterplot of the points in the dataset.

```
gpa_lm <- lm(GPA~ACT,data=gpa)
plot(GPA~ACT,data=gpa)
abline(coef(gpa_lm))
```



(a) Explain in words the criterion that is used to obtain the fitted line in the plot above.

Solution:

The line is fitted by least squares. This minimizes the sum of squared residuals, where the residual for each student is the difference between the value of GPA for that student and the value predicted by their ACT score.

(b) Defining appropriate notation, write an equation for the fitted model in subscript form. At this point, you don't have to explain how the coefficients are calculated.

Solution:

Let y_i be the freshman GPA for student i, $i = 1, \ldots, n$ with $n = 705$. Let x_i be the corresponding ACT score. The model in subscript form is

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

where e_i is the residual for student i.

(c) Defining appropriate notation, write an equation for the fitted model in matrix form. You still don't have to explain how the coefficients are calculated.

Solution:

Define the column vector of coefficients as $\mathbf{b} = (b_1, b_2)$. Let $\mathbf{y}$ be the column vector $(y_1, \ldots, y_n)$ and let

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

Finally, let $\mathbf{e} = (e_1, \ldots, e_n)$ be a column vector of residuals. The matrix form of the linear model is

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

(d) Now, explain using matrix notation how the model coefficients are calculated.

Solution:

The least squares choice of $\mathbf{b}$ is calculated using the equation

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

(e) Write an equation using subscript notation for the *fitted value* for the ith student. Write a sentence to explain the interpretation of this fitted value.

Solution:

The fitted value for height x_i of player i is

$$\hat{y}_i = b_1 x_i + b_2$$

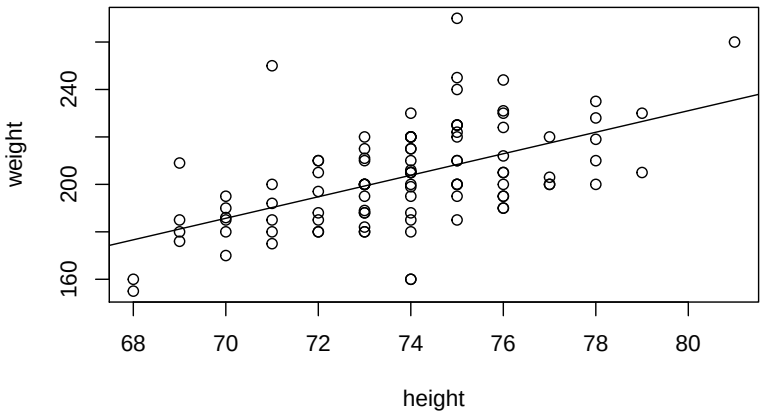
This is the predicted value of weight from the best fit line for weight of an individual with height x_i.

Q4-2.

A statistician employed by a major league baseball team is asked to assess the range of typical weights for major league baseball players of a given height. She obtains data from http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_MLB_HeightsWeights and reads them into R as a dataframe including variables 'Height' (in inches) and 'Weight' (in pounds) for each of 1035 Major League Baseball players. She starts by analyzing just the first 100 players.

She fits a linear model and plots the data and the resulting fitted line using the following R code:

```
weight_lm <- lm(weight ~ height)
plot(height,weight)
abline(coef(weight_lm))
```



(a) Write out the fitted linear model using subscript notation, including the following coefficients from `weight_lm`. This means you are asked to use actual numbers, rather than letters, for the model coefficients. Make sure to define any notation you introduce.

```
round(coef(weight_lm),3)
```

```
## (Intercept)      height
##    -131.652       4.534
```

Solution:

Let y_i be the weight of observation i, $i = 1, \ldots, 100$, and x_i be the corresponding height. The model in subscript form is

$$y_i = 4.53 \times x_i - 131.7 + e_i, \quad i = 1, \ldots, 100$$

where e_i is the residual for observation i.

(b) Use matrix notation to explain how these coefficients were calculated.

Solution:

Define the column vector of coefficients as $\mathbf{b} = (4.53, -131.7)$. Let $\mathbf{y}$ be the column vector $(y_1, \ldots, y_{100})$ and let

$$\mathbb{X} = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_{100} & 1 \end{bmatrix}$$

We obtain $\mathbf{b}$ using the equation

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

(c) The tenth observation corresponds to Adam Stern, an outfielder for the Baltimore Orioles. His recorded height is 71 inches. Write out the formula for the fitted value for this observation. You do not need to simplify your calculation.

Solution:

We have the fitted value

$$\hat{y}_{10} = 4.53 \times 71 - 131.7$$

(d) Use matrix notation to write out an expression for the fitted values of the model. Make sure to define appropriate notation.

Solution:

Define $\mathbb{X}$ and $\mathbf{b}$ as above. Let $\hat{\mathbf{y}}$ be the column vector $(\hat{y}_1, \ldots \hat{y}_{100})$ where $\hat{y}_i$ is the fitted value corresponding to observation i. The fitted values are given by

$$\hat{\mathbf{y}} = \mathbb{X}\mathbf{b}$$

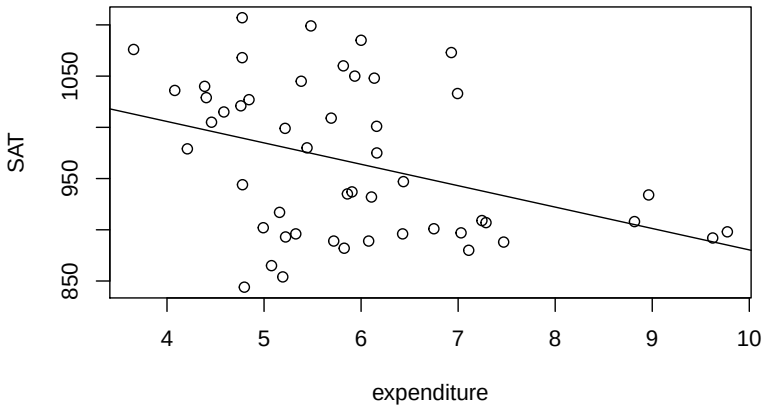
Since $\mathbf{b} = (\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y}$ we can also write

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

Q4-3.

The government wants to understand the relationship between expenditures on public education and test results. The dataset `SAT` contains the per-pupil annual expenditure (in thousands of dollars) and the average SAT score for each of the 50 states in 1994-95. The plot below shows a line fitted to a scatterplot of the points in the dataset.

```
sat_lm <- lm(SAT~expenditure,data=sat)
plot(SAT~expenditure,data=sat)
abline(coef(sat_lm))
```



(a) Write out the regression model in subscript form (including an intercept term). Use letters, rather than actual numbers, for the model coefficients: you don't have the actual numbers at this point. Make sure to define any notation you introduce.

Solution:

Let y_i be the average SAT score in state i, $i = 1, \ldots, n$ with $n = 50$. Let x_i be the corresponding per-pupil expenditure. The model in subscript form is

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

where e_i is the residual for state i.

The table below shows the head of the dataset (the first 6 rows).

```
head(sat)
```

```
##            expenditure  SAT
## Alabama          4.405 1029
## Alaska           8.963  934
## Arizona          4.778  944
## Arkansas         4.459 1005
## California       4.992  902
## Colorado         5.443  980
```

(b) Write out the corresponding design matrix. You only need to put in actual numbers for the first 5 rows, use ... after that and specify the dimension of the matrix.

Solution:

The design matrix $\mathbb{X}$ is a 50×2 matrix,

$$\mathbb{X} = \begin{bmatrix} 4.405 & 1 \\ 8.963 & 1 \\ 4.778 & 1 \\ 4.459 & 1 \\ 4.992 & 1 \\ 5.443 & 1 \\ \vdots & \vdots \end{bmatrix}$$

We could also put the column of ones first. This should be consistent with how you write your model. If b_1 is the intercept and b_2 the slope, then the first column should contain the ones.

(c) Explain how the model coefficients are evaluated. Name the method and give the appropriate formula.

Solution:

Let $\mathbf{y}$ be the column vector $(y_1, \ldots, y_n)$ of average SAT scores in each of the $n = 50$ states. Define the column vector of coefficients as $\mathbf{b} = (b_1, b_2)$. Let $\mathbb{X}$ be the design matrix defined above. We obtain $\mathbf{b}$ by using the equation for the least squares fit,

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

(d) Describe in one line what trend you observe from the plot (what would you interpret from this data). Is this what you would've expected? What could be a possible justification for the trend being observed?

Solution:

One might expect increasing expenditure on education to correspond to higher test scores. We see an opposite pattern, with higher expenditure associated with lower SAT scores. Superficially, this is surprising. Possible explanations are (i) higher expenditure might lead to many more people taking SAT, which could drive down the average test scores. (ii) The association could be driven by selection on who takes ACT with expenditure being just a proxy for geography: For example, some southern states which tend to spend less on education also historically tend to take ACT rather than SAT, and maybe only higher-achieving students considering out-of-state universities tend to take SAT from those states.

You are welcome to propose other possible explanations.

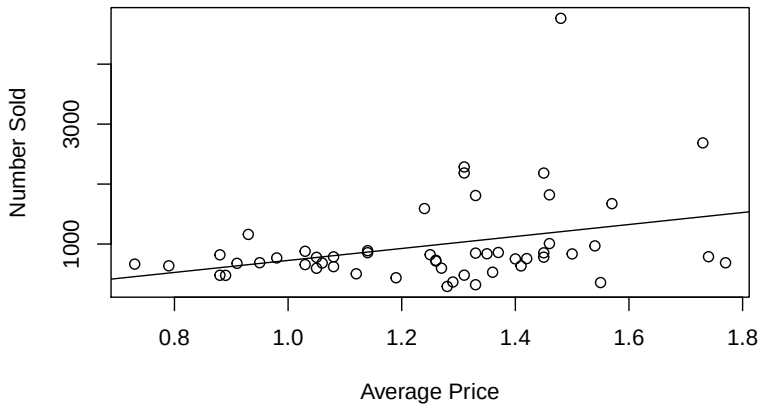
Q4-4.

A statistician employed by an avocado producer is asked to assess the relationship between avocado prices and sales volumn for small Hass avocados. She obtains data from https://www.kaggle.com/neuromusic/avocado-prices and reads them into R as a dataframe. She keeps only the 2016 data for organic avocados sold in the Detroit area and plots the average price in dollars ('AveragePrice') against the number of small Hass avocados sold ('X4046'). This results in a dataset with 52 observations.

```
avocado <- read.csv("avocado.csv")
avocado_2016 <- subset(avocado, year == 2016 & type == 'organic' & region == 'Detroit')
```

She fits a linear model and plots the data and the resulting fitted line using the following R code:

```
price_lm <- lm(X4046 ~ AveragePrice, data = avocado_2016)
plot(avocado_2016$AveragePrice, avocado_2016$X4046,
     xlab = 'Average Price', ylab = 'Number Sold')
abline(coef(price_lm))
```



(a) Write out the fitted linear model using subscript notation, including the following coefficients from `price_lm`. This means you are asked to use actual numbers, rather than letters, for the model coefficients. Make sure to define any notation you introduce.

```
round(coef(price_lm),3)
```

```
##  (Intercept) AveragePrice
##     -272.829      999.062
```

Solution:

Let y_i be the number of small Hass avacodos sold for week i, $i = 1, \ldots, 52$, and x_i be the corresponding average price. The model in subscript form is

$$y_i = 999.06 \times x_i - 272.8 + e_i, \quad i = 1, \ldots, 52$$

where e_i is the residual for observation i.

(b) Use matrix notation to explain how these coefficients were calculated.

Solution:

Define the column vector of coefficients as $\mathbf{b} = (999.06, -272.8)$. Let $\mathbf{y}$ be the column vector $(y_1, \ldots, y_{52})$ and let

$$\mathbb{X} = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_{52} & 1 \end{bmatrix}$$

We obtain $\mathbf{b}$ using the equation

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

(c) Use matrix notation to write out an expression for the residual values of the model. Make sure to define appropriate notation.

Solution:

The residuals are calculated as $\mathbf{e} = \mathbf{y}$ - $\mathbb{X}\mathbf{b}$ where $\mathbf{y}$ be the column vector of average prices $(y_1, \ldots, y_{52})$, $\mathbf{b}$ be the column vector (b_1, b_2) where b_1 and b_2 are the estimated coefficients from the linear model, and

$$\mathbb{X} = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_{52} & 1 \end{bmatrix}$$

is the design matrix.

(d) From the scatter plot above, the statistician notices a potential outlier. This potential outlier corresponds to the the second week in January 2016. This week organic small Hass avocados sold for an average of \$1.48 with a total of 4,763 sold. Write a numeric expression for the residual of this observation—you are not expected to evaluate it.

Solution:

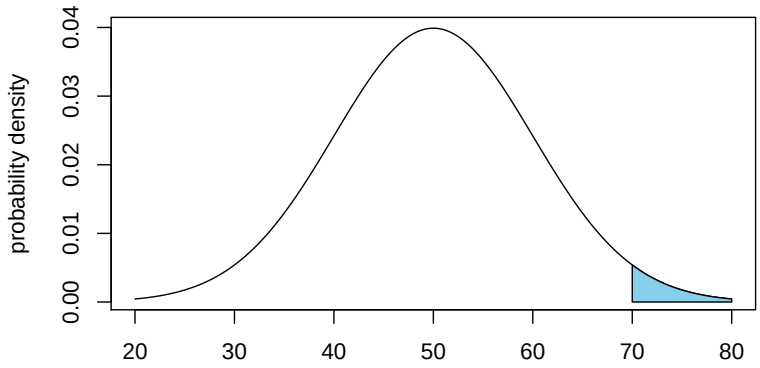
e_i = 4763 - (-272.83 +1.48× 999.06)

e_i = 3557.22

Q5. Probability exercises

Q5-1.

The figure below shows the probability density function of a normal random variable X.



(a) By looking at the probability density function, estimate the mean and standard deviation of X. Use these estimates for the subsequent parts of this question.

Solution:

The center is at about 50. The points of inflection on the density are at about 40 and 60, which should be the mean plus/minus one standard deviation. It looks like about 95% of the area is between 30 and 70, which should be the mean plus/minus two standard deviations. These facts are consistent with a mean of 50 and an SD of 10.

(b) Write a probability statement about the random variable X that corresponding to the shaded area.

Solution:

The shaded area is $\mathrm{P}(X > 70)$.

(c) Write an integral corresponding to this shaded area.

Solution:

$$\int_{70}^{\infty} \frac{1}{\sqrt{2\pi 10^2}} \exp\left\{-\frac{(x-50)^2}{2\times 10^2}\right\} dx$$

(d) Write R code to evaluate this integral numerically.

Solution:

```
1-pnorm(70,mean=50,sd=10)
```

```
## [1] 0.02275013
```

It is acceptable not to label arguments, but then you have to get them in the right order!

Q5-2.

Let Y be a discrete random variable that takes values 0, 1, or 2 with probabilities 0.25, 0.5, and 0.25, respectively.

(a) What is the expected value of Y?

Solution:

The expected value is

$$0.25 \times 0 + 0.5 \times 1 + 0.25 \times 2 = 1$$

(b) What is the variance of Y?

Solution:

The variance is

$$0.25 \times (0-1)^2 + 0.5 \times (1-1)^2 + 0.25 \times (2-1)^2 = 0.5$$

Alternatively, we can calculate the expected value of X^2 first:

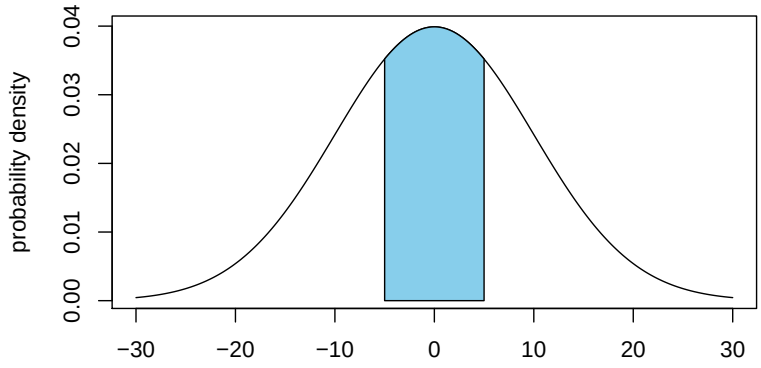
$$\mathrm{E}[X^2] = 0.25 \times 0^2 + 0.5 \times 1^2 + 0.25 \times 2^2 = 0.5 \times 1 + 0.25 \times 4 = 1.5$$

and then

$$\mathrm{Var}(X) = 1.5 - 1^2 = 0.5$$

Q5-3.

The figure below shows the probability density function of a normal random variable X.



(a) By looking at the probability density function, estimate the mean and standard deviation of X. Use these estimates for the subsequent parts of this question.

Solution:

The center is at about 0. The points of inflection on the density are at about -10 and 10, which should be the mean plus/minus one standard deviation. It looks like about 95% of the area is between -20 and 20, which should be the mean plus/minus two standard deviations. These facts are consistent with a mean of 0 and an SD of 10.

(b) Write a probability statement about the random variable X that corresponding to the shaded area.

Solution:

The shaded area is $\mathrm{P}(-5 < X < 5)$.

(c) Write an integral corresponding to this shaded area.

Solution:

$$\int_{-5}^{5} \frac{1}{\sqrt{2\pi 10^2}} \exp\left\{-\frac{x^2}{2 \times 10^2}\right\} dx$$

(d) Write R code to evaluate this integral numerically.

Solution:

The left tail is

```
pnorm(-5,mean=0,sd=10)
```

```
## [1] 0.3085375
```

The right tail is

```
1-pnorm(5,mean=0,sd=10)
```

```
## [1] 0.3085375
```

We could have used symmetry to avoid calculating the same thing twice. Subtracting these tails from 1 gives the total shaded area.

```
1-pnorm(-5,mean=0,sd=10)- (1-pnorm(5,mean=0,sd=10))
```

```
## [1] 0.3829249
```

It is acceptable not to label arguments, but then you have to get them in the right order!

Q5-4.

The average midterm score of 20 students is 40 out of a maximum of 75. The professor realizes that she missed one student and upon including his score, the average went up by one point and became 41.

(a) What is the midterm score of the 21st student?

Solution:

The average score of 20 students is 40. Hence, the total score of the 20 students is $20 \times 40 = 800$. The average score of 21 students is 41. Hence, the total score of the 21 students is $21 \times 41 = 861$. Hence, the midterm score of the 21st student is $861 - 800 = 61$.

(b) Suppose the sample mean and variance are 41 and 36, and we model the midterm scores as being a draw from a normal distributed with these parameters. What is the chance that a student drawn at random gets over 59? Write your answer as a call to `pnorm()` and also give an approximate answer based on your knowledge of the normal distribution.

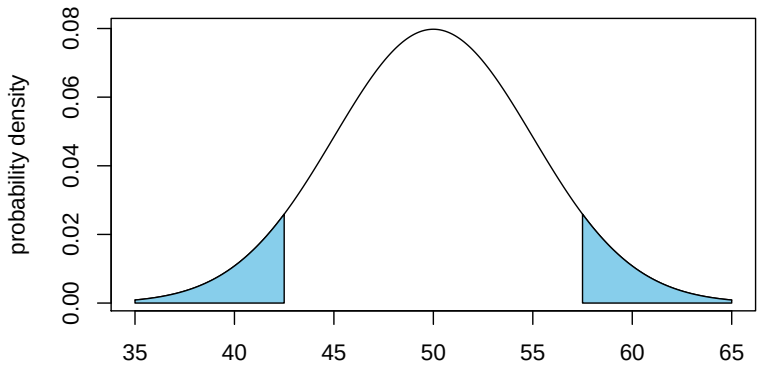
Solution:

Since the variance $\sigma^2 = 36 = 6^2$, the standard deviation is $\sigma = 6$. 59 is three standard deviations above the mean. The probability that a normal random falls more than 3 SDs away from the mean is about 3/1000 (This is in the list of formulas given for the quiz, but is also worth knowing as a fact.) So, the chance of being more than 3 SDs above the mean is about 0.0015.

To evaluate this in R, we can do `pnorm(-3)`. This gives the left tail, which by symmetry matches the probability to the right of 3 for a standard normal random variable. `pnorm(41-3*6,mean=41,sd=6)` gives the same answer.

Q5-5.

The figure below shows the probability density function of a normal random variable X.



(a) By looking at the probability density function, estimate the mean and standard deviation of X. Use these estimates for the subsequent parts of this question.

Solution:

The center is at about 50. The points of inflection on the density are at about 42.5 and 57.5, which should be the mean plus/minus one and a half standard deviations.

(b) Write a probability statement about the random variable X corresponding to the shaded area.

Solution:

The shaded area is $\mathrm{P}(X < 42.5 \text{ or } X > 57.5)$.

(c) Write an integral corresponding to this shaded area.

Solution:

$$\int_{-\infty}^{42.5} \frac{1}{\sqrt{2\pi 5^2}} \exp\left\{-\frac{(x-50)^2}{2\times 5^2}\right\} dx + \int_{57.5}^{\infty} \frac{1}{\sqrt{2\pi 5^2}} \exp\left\{-\frac{(x-50)^2}{2\times 5^2}\right\} dx$$

(d) Write R code to evaluate this integral numerically.

Solution:

```
pnorm(42.5, mean=50, sd=5) + pnorm(57.5, mean=50, sd=5, lower.tail = FALSE)
```

```
## [1] 0.1336144
```

It is acceptable not to label arguments, but then you have to get them in the right order!
