Quiz 2, STATS 401 F18

In lab on 11/16

Name: **UMID:**

Instructions. You have a time allowance of 50 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.

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

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

(2) $\hat{\boldsymbol{\beta}} = \big(\mathbb{X}^{\mathsf{T}}\mathbb{X}\big)^{-1}\,\mathbb{X}^{\mathsf{T}}\mathbf{Y}, \qquad \mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2\big(\mathbb{X}^{\mathsf{T}}\mathbb{X}\big)^{-1}$

(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) $\mathrm{Var}(\mathbb{A}\mathbf{Y}) = \mathbb{A}\mathrm{Var}(\mathbf{Y})\mathbb{A}^{\mathsf{T}}, \qquad \mathrm{var}(\mathbb{X}\mathbb{A}^{\mathsf{T}}) = \mathbb{A}\,\mathrm{var}(\mathbb{X})\,\mathbb{A}^{\mathsf{T}}$

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

(7) 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%.

(8) Syntax from `?pnorm`:

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

(9) $(\mathbb{A}\mathbb{B})^{\mathsf{T}} = \mathbb{B}^{\mathsf{T}}\mathbb{A}^{\mathsf{T}}, \quad (\mathbb{A}\mathbb{B})^{-1} = \mathbb{B}^{-1}\mathbb{A}^{-1}, \quad (\mathbb{A}^{\mathsf{T}})^{-1} = (\mathbb{A}^{-1})^{\mathsf{T}}, \quad (\mathbb{A}^{\mathsf{T}})^{\mathsf{T}} = \mathbb{A}.$

Q1. Circle TRUE or FALSE for the following statements. No explanation is necessary.

TRUE or FALSE. For a given data set of pairs of values $(x_1, y_1), \ldots, (x_n, y_n)$, an infinite number of possible regression equations can be fitted to the corresponding scatter diagram, and each equation will have a unique combination of values for the slope b_1 and y-intercept b_2. However, only one equation will be the "best fit" as defined by the least-squares criterion.

TRUE or FALSE. `qnorm(0.025)` is greater than `qt(0.025,df=10)`.

TRUE or FALSE. If unemployment rate is statistically positively associated with change in life expectancy, we can safely conclude that the short-term consequence of a public policy decreasing unemployment is likely to be a short-term decrease in life expectancy.

Q2. Normal approximations, 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.

(a) Find the mean and variance of X_1.

(b) Use (a) to find the mean and variance of $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

(c) Now suppose $n = 200$ and suppose that $\bar{X}$ is well approximated by a normal random variable. 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.

Q3. Prediction

To investigate the consequences of metal poisoning, 25 beakers of minnow larvae were exposed to varying levels of copper and zinc and the protein content was measured. The data are as follows.

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  195.894      8.548  22.917    0.000
## Copper        -0.135      0.072  -1.879    0.074
## Zinc          -0.045      0.007  -6.207    0.000
```

The sample linear model is $\mathbf{y} = \mathbb{X}\mathbf{b} + \mathbf{e}$. Here, y_i is a measurement of total larva protein at the end of the experiment (in microgram, μg). $\mathbb{X} = [x_{ij}]$ is a 25×3 matrix where $x_{i1} = 1$, x_{i2} is copper concentration (in parts per million, ppm) in beaker i , and x_{i3} is zinc concentration (in parts per million, ppm) in beaker i.

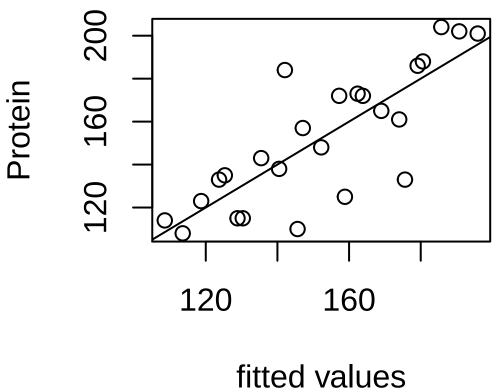
Suppose we're interested in predicting the protein in a new observation at 100ppm copper and 1000ppm zinc.

(a) Specify the values in a row matrix $\mathbf{x}^*$ such that $\mathbf{y}^* = \mathbf{x}^*\mathbf{b}$ gives a least squares prediction of the new observation. Find a numerical expression for this: you are not expected to evaluate the expression.

(b) Explain how to use the data vector $\mathbf{y}$, the design matrix $\mathbb{X}$, and your row vector $\mathbf{x}^*$ to construct a prediction interval that will cover the new measurement in approximately 95% of replications. Your answer should include formulas to construct this interval.

(c) Find a numerical expression for a 95% confidence interval for the relationship between zinc exposure and protein content in minnow larvae.

(d)




```
##      Copper           Zinc          Protein
##  Min.   :  0.0   Min.   :   0   Min.   :108.0
##  1st Qu.: 38.0   1st Qu.: 375   1st Qu.:125.0
##  Median : 75.0   Median : 750   Median :148.0
##  Mean   : 75.2   Mean   : 750   Mean   :152.2
##  3rd Qu.:113.0   3rd Qu.:1125   3rd Qu.:173.0
##  Max.   :150.0   Max.   :1500   Max.   :204.0
```

Based on the graph above and the corresponding summary statistics, is this model a good fit for the data? Do you have any concerns about using this model for this prediction.

Q4. Linear models with factors

We consider a dataset of measurements on crabs. The start of the dataset `crabs` is shown below. The species `sp` corresponds to the color of the crabs, which is a factor with two levels, Blue (`B`) and Orange (`O`). We want to study the difference between the frontal lobe size (`FL`) of the two species.

```r
head(crabs)
```

```
##   sp sex index   FL  RW   CL   CW  BD
## 1  B   M     1  8.1 6.7 16.1 19.0 7.0
## 2  B   M     2  8.8 7.7 18.1 20.8 7.4
## 3  B   M     3  9.2 7.8 19.0 22.4 7.7
## 4  B   M     4  9.6 7.9 20.1 23.1 8.2
## 5  B   M     5  9.8 8.0 20.3 23.0 8.2
## 6  B   M     6 10.8 9.0 23.0 26.5 9.8
```

Consider the probability model $Y_i = \mu_1 x_{Bi} + \mu_2 x_{Oi} + \epsilon_i$ for $i = 1, ..., 200$. Y_i is the frontal lobe size of crab i. x_{Bi} is 1 if crab i is of species Blue and 0 otherwise. Similarly, x_{Oi} is 1 if crab i is of species Orange and 0 otherwise. ϵ_i are i.i.d with mean 0 and variance σ^2. This model can be fit to the `crabs` dataset in R using the `lm()` function. The resulting summary is provided below.

```r
lm_crab <- lm(FL~sp-1, data=crabs)
summary(lm_crab)$coefficients[,1:2]
```

```
##     Estimate Std. Error
## spB   14.056  0.3150194
## spO   17.110  0.3150194
```

(a) Interpret the meaning of μ_1 and μ_2 in the above probability model

(b) Build a 95% confidence interval for μ_1 using the normal approximation. You do not need to simplify your upper and lower bounds.

(c) What is the design matrix used to fit the model above? Write out the first 6 rows.