Quiz 2, STATS 401 F18

In lab on 11/16

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. Q1 and Q2 review material from throughout the course so far. Q3 and Q4 focus on recently covered topics. The quiz will have several TRUE/FALSE questions drawn at random for Q1, and one question drawn at random for each of Q2, Q3 and Q4. Small changes and corrections from this version may be included in the quiz, but no new questions are anticipated.

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

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

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

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

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

Q1-01.

TRUE or FALSE. In the sample regression line $y = b_1 x + b_2$, the term b_2 is the y-intercept; this is the value of y where the line intersects the y-axis whenever $x = 0$.

Q1-02.

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.

Q1-03.

TRUE or FALSE. If the normality assumption for the measurement model is violated, this is more problematic for the prediction interval for a linear model than for confidence intervals on the parameters.

Q1-04.

TRUE or FALSE. A physicist measures extension y_i for a spring at various measures of load x_i. You agree to help with carrying out inference using a linear model. The right model to fit is

$$Y_i = \beta x_i + \epsilon_i, \quad \epsilon_i \sim \text{iid normal}(0, \sigma^2)$$

rather than the usual simple linear regression probability model

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad \epsilon_i \sim \text{iid normal}(0, \sigma^2).$$

Q1-05.

TRUE or FALSE. If we cannot make replications of the data collection procedure then we cannot properly construct a confidence interval.

Q1-06.

TRUE or FALSE. We obtain a smaller standard error when constructing a prediction interval than the standard error used for a confidence interval for the expected value of a new outcome.

Q1-07.

TRUE or FALSE. Suppose we have a factor with three levels. If our linear model includes an intercept, we should include dummy variables for all three factor levels.

Q1-08.

TRUE or FALSE. Suppose we have been recruited to help study the effect of phone use an hour before bed and the amount of sleep undergraduate students get. We survey 30 undergraduate students, recording the number of minutes they report using their phone in the hour before bed and how long they slept. A scatterplot of the data look football-shaped, so we model the data using a linear model with normal measurement error. A friend asks you to guess how much sleep he gets when he uses his phone for 40 minutes before bed. In this case, it is clearly better to use the t-distribution than the normal distribution to construct our prediction interval for how much sleep your friend receives.

Q1-09.

TRUE or FALSE. Data $\mathbf{y}$ are modeled using the probablity model $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. The model-generated fitted vector of fitted values is $\hat{\mathbf{Y}} = \mathbb{X}\hat{\boldsymbol{\beta}}$. The sample residual vector $\mathbf{e}$ can be written as $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ and so a model-generated residual vector is $\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \hat{\mathbf{Y}}$. By the definition of measurement error, $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$. Is it true or false that $\mathrm{E}[\hat{\boldsymbol{\epsilon}}] = \mathbf{0}$?

Q1-10. TRUE or FALSE. If two random variables are uncorrelated, this means they are independent.

Q1-11.

TRUE or FALSE. A 95% confidence interval is narrower than the corresponding 90% confidence interval.

Q1-12.

TRUE or FALSE. If two random variables are independent, this means they are uncorrelated.

Q1-13.

TRUE or FALSE. If two bivariate normal random variables are uncorrelated, this means they are independent.

Q1-14. TRUE or FALSE. Let $X \sim \text{normal}(0, 1)$. Then $\mathrm{P}(X < -c) = 1 - \mathrm{P}(X < c)$ where $c > 0$.

Q1-15. TRUE or FALSE. Let $X \sim \text{normal}(\mu, \sigma)$. Then $\mathrm{P}(X < -c) = 1 - \mathrm{P}(X < c + \mu)$ where $c > 0$.

Q1-16.

TRUE or FALSE. When the fitted values $\hat{y}_1, \ldots \hat{y}_n$ and the actual values $y_1, \ldots, y_n$ are the same, the standard error on the linear model coefficients is 0.0.

Q1-17.

TRUE or FALSE. `pnorm(19.60,mean=0,sd=10)` is 0.95

Q1-18.

TRUE or FALSE. `qnorm(1.960,mean=0,sd=10)` returns `NaN`

Q1-19.

TRUE or FALSE. `qnorm(0.5)` and `pnorm(0)` both return the same value.

Q1-20.

TRUE or FALSE. `qt(0.5,df=10)` is greater than `qnorm(0.5)`.

Q1-21.

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

Q1-22.

TRUE or FALSE. If all covariates are allocated to units at random, for example randomized assignment of treatments to patients in a medical trial, then we can legitimately interpret statistically significant covariates as causal effects. We do not have to pay attention to the saying "Association is not causation."

Q1-23.

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.

Q1-24.

TRUE or FALSE. If unemployment rate is statistically positively associated with change in life expectancy, we can safely conclude that some phenomenon related to the economic boom/bust cycle causes increased mortality in periods of high economic growth.

Q1-25.

TRUE or FALSE. Suppose that a volcanic activity index is statistically positively associated with change in global atmospheric carbon dioxide. We can safely conclude that volcanic activity causes measurable changes in global greenhouse gas levels.

Q1-26.

TRUE or FALSE. Suppose that a volcanic activity index is statistically positively associated with change in global atmospheric carbon dioxide. We can safely conclude that carbon dioxide emitted during volcanic activity causes measurable changes in global carbon dioxide levels.

Q2. Normal approximations, mean and variance

Q2-1.

Recall the following analysis where 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 fits a linear model to predict freshman GPA using ACT exam scores and percentile ranking of each student within their high school, as follows.

```
head(gpa)
```

```
##   ID  GPA High_School ACT Year
## 1  1 0.98          61  20 1996
## 2  2 1.13          84  20 1996
## 3  3 1.25          74  19 1996
## 4  4 1.32          95  23 1996
## 5  5 1.48          77  28 1996
## 6  6 1.57          47  23 1996
```

```
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
```

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 we have checked that the usual probability model for a linear model is appropriate for these data (you are not asked to write out the probability model here).

Use a normal approximation to find an expression for the probability that the difference between the sample coefficient for a draw from the probability model 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.

Q2-2.

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.

Q2-3.

Let $X_1, X_2, \ldots, X_n$ be independent random variables each of which take value 0 with probability 1/3 and 1 with probability 2/3.

(a) Use the definitions and basic properties of expectation and variance to find the expected value and variance of X_1.

(b) Use these results to find the mean and variance of $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. (You may know about the binomial distribution, and you may know a formula for the mean and variance. If so, you can use that to check your work, but you are asked to find the solution directly.)

(c) Now suppose $n = 50$ and suppose that $\bar{X}$ is well approximated by a normal distribution. Find $\mathrm{P}(0.45 < \bar{X} < 0.55)$. 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.

Q2-4.

Let $\mathbf{U} = (W, X, Y)$ be a multivariate normal vector random variable. Suppose that

$$\mathrm{E}(W) = 0, \quad \mathrm{E}(X) = 2, \quad \mathrm{E}(Y) = 2$$

$$\mathrm{Var}(W) = \mathrm{Var}(X) = \mathrm{Var}(Y) = 2, \quad \mathrm{Cor}(X, Y) = -0.5, \quad \mathrm{Cor}(Y, W) = -0.5, \quad \mathrm{Cor}(X, W) = 0.$$

(a) Find the distribution of $W + X - 2Y$.

(b) Find $\mathrm{P}(2Y < X + W + 1)$. 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.

Q2-5.

Let X and Y be bivariate random variables. Suppose that $X \sim \text{normal}(0, 1)$ and $\mathrm{Cor}(X, Y) = 1$.
If $\mathrm{P}(X > Y) = 0.8413448$ and $\mathrm{P}(X < Y + 1) = 0.5$ then find $\mathrm{P}(-2 < Y < 2)$. Write your answer as a call to `pnorm()`.
Hint: `qnorm(0.8413448)=1`.

Q3. Prediction

Q3-1.

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.

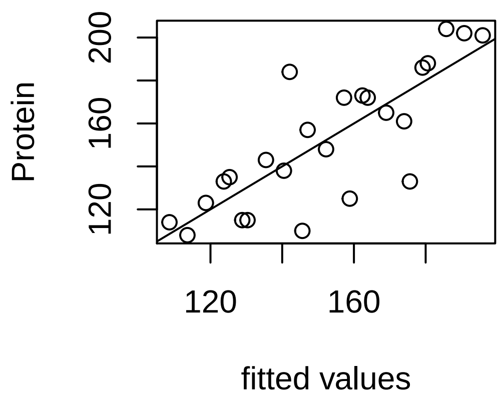
Suppose we're interested in predicting the protein in a new observation at 100ppm copper and 1000ppm zinc.

(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.

Q3-2.

We have been recruited by a California university to explore the relationship between water salinity, water oxygen, and water temperature. We have been given 60 years of oceanographic data collected from the California Current by the California Cooperative Oceanic Fisheries Investigations. Below is a snapshot of the data. (Source: https://www.kaggle.com/sohier/calcofi)

- Depthm: Depth in meters
- T_degC: Water temperture in degrees Celsius
- Salnty: Water Salinity in g of salt per kg of water
- 02ml_L: O_2 mixing ratio in ml/L

We fit a linear model to the data to predict temperature given the other variables; the results are shown below.

```
##             Estimate Std. Error
## (Intercept)  -78.592      3.697
## Depthm        -0.004      0.000
## Salnty         2.482      0.108
## O2ml_L         1.956      0.024
```

Suppose we observe a new outcome with covariate vector $\mathbf{x^*} = (x_1^*, x_2^*, x_3^*, x_4^*)$ corresponding to the intercept, depth, salinity and oxygen level respectively. Call the as-yet-unobserved new temperature y^*.

(a) Suppose we wanted to calculate a 95% confidence interval for the expected value of the new outcome. Write the expression for this calculation and define all terms.

(b) Suppose instead, we wanted to calculate a 95% prediction interval for the new outcome. Write the expression for this calculation and define all terms.

(c) How would you check that your confidence and prediction intervals are plausible?

(d) Find a numeric expression for the 95% confidence interval for the relationship between oxygen levels and water temperature.

Q3-3. The director of the CDC wants to assess how well rates of hospital-acquired infections (`Infection.risk`) can be predicted using properities of a hospital. She expects to use the average length of stay (`Length.of.stay`) in days, the average number of cultures for each patient without signs or symptoms of hospital-acquired infection, times 100 (`Culture`), the number of X-ray procedures divided by number of patients without signs or symptoms of pneumonia, times 100 (`X.ray`), and the number of beds a hospital has (`Beds`).

Let $\mathbf{x_1}$ be the length of stay, $\mathbf{x_2}$ be the culture count, $\mathbf{x_3}$ be the number of X-rays, and $\mathbf{x_4}$ be the number of beds. Consider the probability model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$$

for $i = 1, \ldots, n$ with $n = 113$, and $\epsilon_1, \ldots, \epsilon_n \sim$ iid normal$(0, \sigma)$.

She fits the linear model corresponding to this probability model in R:

```
##                Estimate Std. Error
## (Intercept)     0.41495    0.53089
## Length.of.stay  0.18453    0.05778
## Culture         0.04800    0.01006
## X.ray           0.01304    0.00549
## Beds            0.00134    0.00052
```

(a) The CDC director asks you to determine if the size of the hospital (measured in the number of beds) affects the infection rate of the hospital. Write the null and alternative hypotheses and sample test statistic we would use to answer this question.

(b) What is the distribution of the model-generated test statistic corresponding to your sample test statistic from part (a)?

(c) Suppose we know that a local hopital has an average length of stay of 8 days, the average culture count is 14, the average number of X-rays is 90, and the number of beds is 40. Find a numeric expression for the predicted value for this observation; you are not expected to evaluate it.

(d) Suppose we constructed a confidence interval for the expected infection rate for the hospital in part c. How would you check that your confidence interval is plausible?

Q3-4. Switzerland, in 1888, was entering a period known as the demographic transition; i.e., its fertility was beginning to fall from the high level typical of underdeveloped countries. This Swiss government has commissioned us to determining the factors most contributing to this decline.

We collect the following variables for each of the 47 French-speaking provinces around 1988:

- Fertility: common standardized fertility measure
- Agriculture: % of males involved in agriculture as occupation
- Examination: % draftees receiving highest mark on army examination
- Education: % education beyond primary school for draftees.
- Catholic: % 'catholic' (as opposed to 'protestant').
- Infant.Mortality: live births who live less than 1 year.

Let $\mathbf{x_1}$ be the agriculture rate, $\mathbf{x_2}$ be the examination rate, $\mathbf{x_3}$ be the education rate, $\mathbf{x_4}$ be the catholic rate, and $\mathbf{x_5}$ be the infant mortality rate. Consider the probability model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \epsilon_i$$

for $i = 1, \ldots, n$ with $n = 47$ and $\epsilon_1, \ldots, \epsilon_n \sim$ iid normal$(0, \sigma)$.

We fit a the regression model corresponding to this probability model in R:

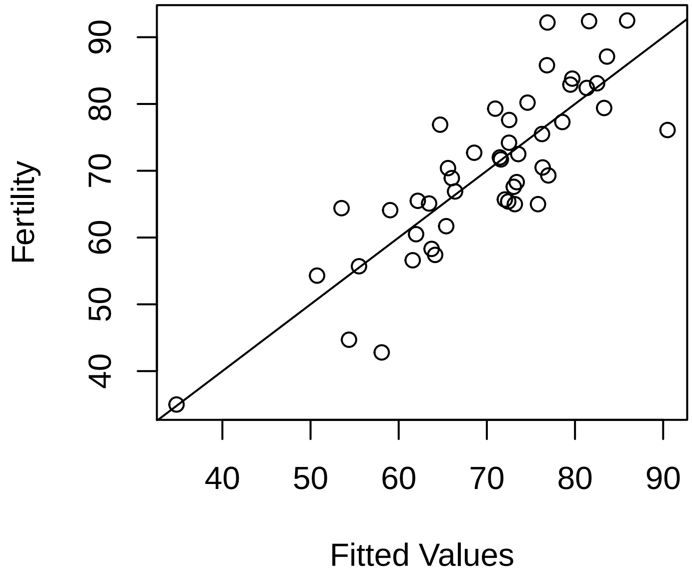
```
##                  Estimate Std. Error
## (Intercept)        66.915     10.706
## Agriculture        -0.172      0.070
## Examination        -0.258      0.254
## Education          -0.871      0.183
## Catholic            0.104      0.035
## Infant.Mortality    1.077      0.382
```

(a) The Swiss governmment is skeptical that the examination percentage affects the fertility rate. Write the null and alternative hypotheses we would use to answer this question.

(b) (i) What is your test statistic for part (a)? (ii) What is the distribution of a model-generated test statistic under the null hypothesis? (iii) What is your conclusion for the hypothesis test in part (a)? No calculations are necessary for this question. Note that if there is no explicit specification of whether the "sample" or "model generated" test statistic is intended, this usually refers to the sample version.

(c) A new province is conquered in 1889 and its statistics are added to our data. This new province had an agriculture rate of 70%, examination rate of 22%, and education rate of 10%, a catholic rate of 50%, and an infant mortality rate of 20%. Find a numeric expression for the predicted fertility rate of this new province. You are not expected to evaluate this expression.

(d)




```r
summary(swiss)
```

```
##    Fertility      Agriculture     Examination      Education
##  Min.   :35.00   Min.   : 1.20   Min.   : 3.00   Min.   : 1.00
##  1st Qu.:64.70   1st Qu.:35.90   1st Qu.:12.00   1st Qu.: 6.00
##  Median :70.40   Median :54.10   Median :16.00   Median : 8.00
##  Mean   :70.14   Mean   :50.66   Mean   :16.49   Mean   :10.98
##  3rd Qu.:78.45   3rd Qu.:67.65   3rd Qu.:22.00   3rd Qu.:12.00
##  Max.   :92.50   Max.   :89.70   Max.   :37.00   Max.   :53.00
##     Catholic       Infant.Mortality
##  Min.   :  2.150   Min.   :10.80
##  1st Qu.:  5.195   1st Qu.:18.15
##  Median : 15.140   Median :20.00
##  Mean   : 41.144   Mean   :19.94
##  3rd Qu.: 93.125   3rd Qu.:21.70
##  Max.   :100.000   Max.   :26.60
```

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

Q4-1. 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.

Q4-2.

In the following data set, we examine the effect of two diets on mice bodyweights. The variable `Diet` is a factor with two levels: "chow" and "hf."

```r
head(mice)
```

```
##   Diet Bodyweight
## 1 chow      21.51
## 2 chow      28.14
## 3 chow      24.04
## 4 chow      23.45
## 5 chow      23.68
## 6 chow      19.79
```

We fit a linear model in R and look at its design matrix $\mathbb{X}$.

```r
lm_mice <- lm(Bodyweight~Diet,data=mice)
model.matrix(lm_mice)
```

```
##    (Intercept) Diethf
## 1            1      0
## 2            1      0
## 3            1      0
## 4            1      0
## 5            1      0
## 6            1      0
## 7            1      0
## 8            1      0
## 9            1      0
## 10           1      0
## 11           1      0
## 12           1      0
## 13           1      1
## 14           1      1
## 15           1      1
## 16           1      1
## 17           1      1
## 18           1      1
## 19           1      1
## 20           1      1
## 21           1      1
## 22           1      1
## 23           1      1
## 24           1      1
## attr(,"assign")
## [1] 0 1
## attr(,"contrasts")
## attr(,"contrasts")$Diet
## [1] "contr.treatment"
```

(a) Write down the sample linear model fitted in `lm_mice` using subscript format–this asks for the usual subscript format for linear models, not the double subscript format introduced to describe models with factors. Make sure to define appropriate notation.

(b) In terms of the coefficients of this sample linear model, explain how to obtain estimates of the means of both treatment groups and the difference between these means.

Q4-3.

We analyze the following data on video game sales in North America. This dataset records sales (in millions of dollars) for 580 games within three genres (shooter, sports and action) from two publishers (Electronic Arts and Activision) with years of release from 2006 to 2010 inclusive, on ten different platforms.

```r
head(vg)
```

```
##                          Name Platform Year   Genre       Publisher Sales
## 1    Call of Duty: Black Ops     X360 2010 Shooter      Activision  9.70
## 2    Call of Duty: Black Ops      PS3 2010 Shooter      Activision  5.99
## 3 Call of Duty: World at War     X360 2008 Shooter      Activision  4.81
## 4 Call of Duty: World at War      PS3 2008 Shooter      Activision  2.73
## 5              FIFA Soccer 11      PS3 2010  Sports Electronic Arts  0.61
## 6               Madden NFL 07      PS2 2006  Sports Electronic Arts  3.63
```

Let $\mathbf{y} = (y_1, \ldots, y_{580})$ be the sales of the games. Let $x_{i,1} = 1$ if game i is published by Activision and 0 otherwise. Similarly, let $x_{i,2} = 1$ if game i is published by Electronic Arts and 0 otherwise.

In R, we fit the sample linear model given by $y_i = m_1 x_{i,1} + m_2 x_{i,2} + e_i$ for $i = 1, \ldots, 580$.

```
lm_vg2 <- lm(Sales ~ Publisher-1, data = vg)
summary(lm_vg2)
```

```
##
## Call:
## lm(formula = Sales ~ Publisher - 1, data = vg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4412 -0.3212 -0.2136  0.0464  9.2588
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## PublisherActivision          0.44124    0.05095   8.661   <2e-16 ***
## PublisherElectronic Arts  0.41361    0.04434   9.327   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8055 on 578 degrees of freedom
## Multiple R-squared:  0.2189, Adjusted R-squared:  0.2162
## F-statistic:    81 on 2 and 578 DF,  p-value: < 2.2e-16
```

(a) What do the coefficients in the summary above measure?

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

(c) Suppose we wish to fit the model $y_i = b_0 + b_1 x_{i,1} + e_i$ for $i = 1, \ldots, 580$. What is the value of b_1?

Q4-4.

We are interested in studying the relationship between the miles per gallon of a car and the number of cylinders its engine has. In the following data set, `mpg` corresponds to the miles per gallon of each car. The variable `cylinders` corresponds to the number of cylinders and takes the values "4 cyl", "6 cyl", or "8 cyl." The variable `horsepower` corresponds to the horse power of each car.

```
head(mpg)
```

```
##   mpg cylinders horsepower
## 1  31     4 cyl         67
## 2  22     4 cyl         98
## 3  27     4 cyl         88
## 4  15     8 cyl        150
## 5  28     4 cyl         86
## 6  21     6 cyl        107
```

Let $\mathbf{x}_1$ be a dummy variable for 6 cylinder cars, $\mathbf{x}_2$ be a dummy variable for 8 cylinder cars, and $\mathbf{x}_3$ be horsepower. Consider the probability model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

for $i = 1, \ldots, 399$ where ϵ_i are iid normal$(0, \sigma)$. We fit the linear model corresponding to this probability model in R:

```
lm_mpg = lm(mpg ~ cylinders + horsepower, data = mpg)
summary(lm_mpg)$coefficients[,1:2]
```

```
##                   Estimate Std. Error
## (Intercept)    37.2708459 0.93803287
## cylinders6 cyl -6.9408552 0.61605263
## cylinders8 cyl -6.1565452 1.04482414
## horsepower     -0.1020284 0.01134433
```

(a) What is the design matrix $\mathbb{X}$? Write out the first 6 rows.

(b) Suppose we have a new car that has 6 cylinders and a horsepower of 110. What is the predicted miles per gallon? You do not need to simplify your calculation.

(c) We want to know if 8 cylinder cars have lower miles per gallon on average than 4 cylinder cars (after controlling for horsepower). What are the null and alternative hypotheses we would use to answer this question?