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

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

(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 $x = 0$.

Solution. TRUE. The equation $y = b_1 x + b_2$ denotes a line corresponding to the least squares fit for a sample when b_1 and b_2 are the least squares coefficients for a simple linear regression model. Substituting $x = 0$ gives $y = b_2$.

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.

Solution. TRUE. You can imagine fitted lines with arbitrarily high residual sum of squares (RSS). There is a unique line mimimizing RSS (except in the special case where $x_1 = x_2 = \cdots = x_n$).

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.

Solution. TRUE. A sample coefficient of the linear model is a sum of contributions from all the data points, and so a central limit principle can apply as long as the number of data points is not small. Thus, a normal approximation for the confidence interval can be a good even if a normal model does not hold well for the measurement error. The prediction interval is largely due to measurement uncertainty from a single measurement and so a central limit principle does not apply.

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

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

Solution. TRUE. Since extension is necessarily zero for an unloaded spring, there is no particular reason to include an intercept here.

Q1-05.

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

Solution. FALSE. A confidence interval is defined using a probability model. Replicability helps us justify a model and the corresponding confidence interval. However, we can (and do) write down models for non-replicable phenomena and we can properly construct confidence intervals for the postulated probability models.

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.

Solution. FALSE. In a prediction interval, we are making a prediction for a single new observation $\mathbf{x}^*$. In a confidence interval for the expected value, we are estimating the expected value for all observations with that $\mathbf{x}^*$ value. There is more uncertainty when predicting the outcome for a single new observation, so we should have a larger standard error.

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.

Solution. FALSE. If we include a dummy variable for all three factor levels, then our model will be over-specified. For example, suppose the three factor levels have sample means of 1, 2, and 3. We could have an estimated intercept of 0 and coefficients 1, 2, and 3. We could also have an estimated intercept of 10 and coefficients of -9, -8, and -7.

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.

Solution. TRUE. Since we only have 30 observations, the t prediction interval will be noticeably wider than the normal interval. The assumptions for the t interval are appropriate here, so it is better to use the exact t distribution (taking into account uncertainty in estimating σ) rather than making a normal approximation.

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 $\mathrm{E}[\hat{\boldsymbol{\epsilon}}] = \mathbf{0}$?

Solution. TRUE.

$$\mathrm{E}[\hat{\boldsymbol{\epsilon}}] = \mathrm{E}[Y] - \mathrm{E}[\hat{Y}] = \mathbb{X}\boldsymbol{\beta} - \mathbb{X}\mathrm{E}[\hat{\boldsymbol{\beta}}] = 0,$$

using the property that $\mathrm{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$.

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

Solution. FALSE. Correlation is a measure of linear dependence. Zero correlation means that the veriables are LINEARLY independent. For instance, if X is uniformly distributed on the interval $[-1, 1]$ and $Y = X^2$, then $\mathrm{Cor}(X, Y) = 0$ but X and Y are not independent.

Q1-11.

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

Solution. FALSE. A 95% confidence interval gives us the range of a population parameter with 95% confidence. If we generated many intervals constructed in the same way from draws of the probability model, 95% of them would contain the true population parameter of interest. The 90% CI would only contain the true parameter 90% of the time (on average). Thus, the parameter may lie outside the 90% CI more than the corresponding 95% CI which is why the 95% CI would be broader thab the 90% CI.

Q1-12.

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

Solution. TRUE. If two variables are independent then larger values of one variable cannot be associated with larger values of the other so their linear dependence (correlation) is zero as well.

Q1-13.

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

Solution. TRUE. This is a useful special property of the multivariate normal distribution.

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

Solution. TRUE. We can calculate as follows.

$$\begin{aligned} \mathrm{P}(X < -c) &= \mathrm{P}(X > c) && \text{(by symmetry)} \\ &= 1 - \mathrm{P}(X < c) \end{aligned}$$

Q1-15.

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

Solution. FALSE. This is most easily seen by sketching a normal curve, marking on the center μ and a point c and applying symmetry. We can also write it out in math.

$$\begin{aligned} \text{P}(X < -c) &= \text{P}\big(X < \mu - (\mu + c)\big) \\ &= \text{P}\big(X > \mu + (\mu + c)\big) \quad \text{using symmetry around the center, } \mu \\ &= 1 - \text{P}(X < c + 2\mu) \end{aligned}$$

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.

Solution. TRUE. A slight issue arises if the model is over-parameterized and so the least squares coefficients are not uniquely identified. In this case, the standard error on the coefficients is infinite even when the model fits the data perfectly. TRUE remains a better answer despite this special case.

Q1-17.

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

Solution. FALSE. This is equivalent to `pnorm(1.960,mean=0,sd=1)` which has a right tail of 2.5% not 5%, leading to the well-known fact that the mean ± 1.96 times the standard error is an approximate 95% confidence interval.

Q1-18.

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

Solution. TRUE. `qnorm()` gives the normal quantile corresponding to the specified left tail probability. Since a probability must be between 0 and 1, `qnorm(1.96,...)` cannot give a numeric answer so returns `NaN`.

Q1-19.

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

Solution. FALSE. `qnorm(0.5)` is the x-value on the standard normal curve with 0.5 probability to the left, and so 0.5 to the right. This is the center of the distribution, hence `qnorm(0.5)=0`. `pnorm(0)` is the probability to the left of 0 on the standard normal curve. This distribution is symmetric about its mean of zero, so `pnorm(0)=0.5`.

Q1-20.

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

Solution. FALSE. `qt(0.5,df=10)` is the x-value of the t distribution on 10 degrees of freedom with 0.5 probability to the left, and so 0.5 to the right. This distribution is symmetric about its mean of zero, so `qt(0.5,df=10)=0`. By similar reasoning, `qnorm(0.5)=0`.

Q1-21.

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

Solution. TRUE. These are the x-values on the standard normal curve and t distribution with a left tail of 0.025. The t distribution has a longer tail, so the 0.025 point shifts left (larger in magnitude, but smaller in value along the number line). This corresponds to the property that t confidence intervals are wider than the corresponding normal approximation confidence intervals.

```
qnorm(0.025)
```

```
## [1] -1.959964
```

```
qt(0.025,df=10)
```

```
## [1] -2.228139
```

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."

Solution. TRUE. A statisically robust association between A and B implies A causes B, B causes A or both have a common cause. A randomized assignment of covariates rules out all possibilities other than a causal one. Formally, this randomization has to include all relevant covariates (one might not think to randomize the treating physicians in a study, for example). We also have to bear in mind that the causal story might not be the one we want: if physicians measuring an outcome are not blind to the treatment and therefore make measurements subconsciously biased toward a new treatment, this is a causal story linking a treatment to a favorable measured outcome, but not the causal interpretation that first comes to mind!

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.

Solution. FALSE. Recall that "association is not causation." Many quantities in the economy follow the same boom/bust cycle. There are many candidates for common causes of both unemployment and life expectancy. For example, reduced overall economic activity leads to both less employment and less air pollution, so perhaps a causal chain from economic activity to air pollution to human health could explain the observed association.

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.

Solution. TRUE. A statisically robust association between A and B implies A causes B, B causes A or both have a common cause. Supposing we can rule out life expectancy fluctuations as a major cause of economic boom/bust fluctuations (which seems safe) we are left with only the possibility that something about fluctuations in economic activity causally affects both unemployment and mortality.

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.

Solution. TRUE. A statisically robust association between A and B implies A causes B, B causes A or both have a common cause. Suppose we can rule out atmospheric processes (such as anthropogenic global climate change) as a cause of major volcanic events, and we can also rule out other geophysical phenomena as causes of both major volcanic events and atmospheric composition. These are not accepted as relevant considerations in general discussion of global climate change. So, we are left with the conclusion that volcanic activity has a non-negligible effect on atmospheric carbon dioxide 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.

Solution. FALSE. A statisically robust association between A and B implies A causes B, B causes A or both have a common cause. Suppose we can rule out atmospheric processes (such as anthropogenic global climate change) as a cause of major volcanic events, and we can also rule out other geophysical phenomena as causes of both major volcanic events and atmospheric composition. These are not accepted as relevant considerations in general discussion of global climate change. So, we are left with the conclusion that volcanic activity has a non-negligible effect on atmospheric carbon dioxide levels. However, we cannot safely conclude the causal chain of events. For example, perhaps volcanoes emit negligible CO_2 but produce large amounts of particles which block sunlight and reduce photosynthesis and hence indirectly affect global CO_2 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.

Solution. The probability of observing a bigger value of the estimated coefficient under the assumed model is approximately

```
1-pnorm(0.03721,mu=0.03528,sd=0.005939)
```

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=0.03528,sd=0.005939))
```

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.

Solution.

(a) We calculate the mean and variance of X_1.

$\mathrm{E}(X_1) = 0 \times 0.5 + 1 \times 0.25 - 1 \times 0.25 = 0$

$\mathrm{E}(X_1^2) = 0 \times 0.5 + 1^2 \times 0.25 + (-1)^2 \times 0.25 = 0.5$

$\mathrm{Var}(X_1) = \mathrm{E}(X_1^2) - \big(\mathrm{E}(X_1)\big)^2 = 0.5 - (0)^2 = 0.5$

(b) Use the linearity property of expectation: the expectation of a sum is the sum of the expectations, and multiplicative constants can be pulled outside expectation.

$$\mathrm{E}(\bar{X}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{1}{n}\big(n \times \mathrm{E}(X_1)\big) = 0$$

Then use the scaling property of variance, together with the property that the variance of a sum of independent random variables is the sum of the variances.

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2} n\,\mathrm{Var}(X_1) = \frac{1}{2n}$$

(c) Since $n = 200$, we have $\mathrm{Var}(\bar{X}) = \frac{1}{400}$ and so $\mathrm{SD}(\bar{X}) = \frac{1}{20}$. Thus, our normal approximation is $\bar{X} \sim \text{normal}\big(0, \frac{1}{20}\big)$. We write the central probability $\mathrm{P}(-c < \bar{X} < c)$ in terms of a left tail probability as

$$\mathrm{P}(-c < \bar{X} < c) = 1 - 2\mathrm{P}(\bar{X} < -c)$$

So, for $\mathrm{P}(-c < \bar{X} < c) = 0.9$ we need $\mathrm{P}(\bar{X} < -c) = 0.05$. For this, we need

```
c = - qnorm(0.05,mean=0,sd=1/20)
```

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.

Solution.

(a) We calculate the mean and variance of X_1.

$\mathrm{E}(X_1) = 0 \times \frac{1}{3} + 1 \times \frac{2}{3} = \frac{2}{3}$

$\mathrm{E}(X_1^2) = 0 \times \frac{1}{3} + 1^2 \times \frac{2}{3} = \frac{2}{3}$

$\mathrm{Var}(X_1) = \mathrm{E}(X_1^2) - \big(\mathrm{E}(X_1)\big)^2 = \frac{2}{3} - (\frac{2}{3})^2 = \frac{2}{9}$

(b) Use the linearity property of expectation: the expectation of a sum is the sum of the expectations, and multiplicative constants can be pulled outside expectation.

$$\mathrm{E}(\bar{X}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{1}{n}\big(n \times \mathrm{E}(X_1)\big) = \mathrm{E}(X_1) = \frac{2}{3}$$

Then use the scaling property of variance, together with the property that the variance of a sum of independent random variables is the sum of the variances.

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2} n\,\mathrm{Var}(X_1) = \frac{2}{9n}$$

(c) Since $n = 50$, we have $\mathrm{Var}(\bar{X}) = \frac{1}{9\times 25}$ and so $\mathrm{SD}(\bar{X}) = 1/15$. Thus, our normal approximation is $\bar{X} \sim \text{normal}\,(2/3, 1/15)$. Hence,

$$\begin{aligned} \mathrm{P}(0.45 < \bar{X} < 0.55) &= \mathrm{P}(\bar{X} < 0.55) - \mathrm{P}(\bar{X} < 0.45) \\ &= \texttt{pnorm}(0.55, \texttt{mean} = 2/3, \texttt{sd} = 1/15) - \texttt{pnorm}(0.45, \texttt{mean} = 2/3, \texttt{sd} = 1/15) \end{aligned}$$

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.

Solution.

(a) $W + X - 2Y$ is a normal variable, since linear combinations of multivariate normal variables are normal. We use linearity to find $\mathrm{E}(W + X - 2Y)$.

$$\mathrm{E}(W + X - 2Y) = \mathrm{E}[W] + \mathrm{E}(X) - 2\mathrm{E}(Y) = 0 + 2 - 2 \times 2 = -2.$$

There are different approaches to finding $\mathrm{Var}(W + X - 2Y)$. Here, we find the variance/covariance matrix of $\mathbf{U} = (W, X, Y)$. First, we calculate the covariances.

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathrm{Cor}(X, Y)\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)} = -0.5\sqrt{2 \times 2} = -1 \\
\mathrm{Cov}(X, W) &= \mathrm{Cor}(X, W)\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(W)} = 0\sqrt{2 \times 2} = 0 \\
\mathrm{Cov}(Y, W) &= \mathrm{Cor}(Y, W)\sqrt{\mathrm{Var}(Y)\,\mathrm{Var}(W)} = -0.5\sqrt{2 \times 2} = -1
\end{aligned}$$

Then,

$$\mathrm{Var}\,(\mathbf{U})) = \begin{bmatrix} \mathrm{Var}(W) & \mathrm{Cov}(W, X) & \mathrm{Cov}(W, Y) \\ \mathrm{Cov}(X, W) & \mathrm{Var}(X) & \mathrm{Cov}(X, Y) \\ \mathrm{Cov}(Y, W) & \mathrm{Cov}(Y, X) & \mathrm{Var}(Y) \end{bmatrix} = \begin{bmatrix} 2 & 0 & -1 \\ 0 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}$$

Then, we write $W + X - 2Y$ in the form

$$W + X - 2Y = \begin{bmatrix} 1 & 1 & -2 \end{bmatrix} \begin{bmatrix} W \\ X \\ Y \end{bmatrix} = \mathbb{A}\mathbf{U}.$$

We can now use the formula for the variance of a linear combination,

$$\begin{aligned}
\mathrm{Var}(W + X - 2Y) &= \mathrm{Var}(\mathbb{A}\mathbf{U}) = \mathbb{A}\mathrm{Var}(\mathbf{U})\mathbb{A}^T \\
&= \begin{bmatrix} 1 & 1 & -2 \end{bmatrix} \begin{bmatrix} 2 & 0 & -1 \\ 0 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix} \\
&= \begin{bmatrix} 1 & 1 & -2 \end{bmatrix} \begin{bmatrix} 2(1) + 0(1) + (-1)(-2) \\ 0(1) + 2(1) + (-1)(-2) \\ (-1)1 + (-1)1 + 2(-2) \end{bmatrix} \\
&= \begin{bmatrix} 1 & 1 & -2 \end{bmatrix} \begin{bmatrix} 4 \\ 4 \\ -6 \end{bmatrix} \\
&= 1(4) + 1(4) + (-2)(-6) \\
&= 20
\end{aligned}$$

So, $W + X - 2Y \sim \mathrm{normal}(-2, \sqrt{20})$.

(b) We rewrite $\mathrm{P}(2Y < X + W + 1)$ as $\mathrm{P}(W + X - 2Y > -1)$. From (a) this can be computed as

```
1-pnorm(-1,mean=-2,sd=sqrt(20))
```

Q2-5.

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

Solution.
Since $\text{Cor}(X,Y) = 1$ this means that X and Y are linear, so we can write $Y = aX + b$ for some constants $a > 0$ and b.

The fact that $\text{P}(X > y) = 0.8413448$ gives one equation for a and b.

$$\text{P}(X > Y) = \text{P}(Y - X < 0) = \text{P}(aX + b - X < 0) = \text{P}((a-1)X + b < 0) = \text{P}(X < \frac{-b}{a-1}) = 0.8413448$$

Since `qnorm(0.8413448)`=1 this means that $\text{P}(X < 1) = 0.8413448$. Since we just found $\text{P}(X < \frac{-b}{a-1}) = 0.8413448$ we deduce that $\frac{-b}{a-1} = 1$ and so

$$-b = a - 1.$$

The fact that $\text{P}(X < Y + 1)0.5$ gives us a second equation.

$$\begin{aligned} \text{P}(X < Y + 1) &= \text{P}(Y - X + 1 > 0) = \text{P}(aX + b + 1 - X > 0) \\ &= \text{P}((a-1)X + b + 1 > 0) = \text{P}(X > \frac{-b-1}{a-1}) = 0.5 \end{aligned}$$

Since $\text{P}(X > 0) = 0.5$ this means that $\frac{-b-1}{a-1} = 0$, so $b = -1$. From above, $-b = a - 1$ so $a = 1 - b$ which means that $a = 2$.

So, $Y = 2X - 1$ where $X \sim \mathcal{N}(0,1)$
So Y is also normal, since the linear combination of normal variables is normal. We can work out its mean and variance,

$$\text{E}(Y) = \text{E}(2X - 1) = 2\text{E}(X) - 1 = -1$$

$$\text{Var}(Y) = \text{Var}(2X - 1) = 2^2 \text{Var}(X) = 4.$$

So, $Y \sim \text{normal}(-1, 2)$ and

$$\begin{aligned} \text{P}(-2 < Y < 2) &= \text{P}(Y < 2) - \text{P}(Y < -2) \\ &= \texttt{pnorm}(2, -1, 2) - \texttt{pnorm}(-2, -1, 2). \end{aligned}$$

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.

Solution

$\mathbf{x}^* = (1, 100, 1000)$

$\hat{y}^* = 195.894 + 100(-0.135) + 1000(-0.045)$

(Evaluation gives $\hat{y}^* = 137.394$)

(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.

Solution

Define

$$\text{SE}_{\text{pred}} = s\sqrt{\mathbf{x^*}^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{x^*} + 1},$$

where s is the residual standard error. SE_{pred} is an estimate of $\text{SD}\big(\mathbf{Y^*} - \mathbf{x^*}\boldsymbol{\beta}\big)$. Thus the prediction interval is

$$\hat{y}^* \pm 1.96\,\text{SE}_{\text{pred}}.$$

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

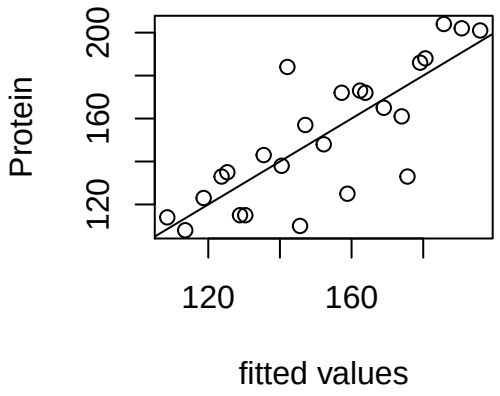
Solution

The 95% confidence interval for the relationship between zinc exposure and protein in minnow larvae is

$$b_3 \pm 1.96\,\text{SE}(b_3) = -0.045 \pm 1.96(0.007)$$

(This evaluates numerically to $[-0.058, -0.031]$.)

(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.

Solution

The model is a good fit for the data. There are no trends or clusters in the plot of the fitted values against the Protein level of the minnow larvae. We have no concerns about using our model to make our prediction, because our x^* contains copper and zinc levels that were observed in our data.

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.

Solution

$[\mathbf{x}^*\mathbf{b} - 1.96\,\mathrm{SE}, \mathbf{x}^*\mathbf{b} + 1.96\,\mathrm{SE}]$, where $SE = s\sqrt{\mathbf{x^*}(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{x^{*T}}}$. $\mathbf{x^*}$ is a row vector of the new observed covariates and $\mathbb{X}$ is the design matrix; s is an approximation of σ, the standard deviation of the errors; $\mathbf{b} = (b_1, b_2, b_3, b_4)$ is the vector of least squares coefficients corresponding to the intercept, depth, salinity and mixing ratio respectively.

(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.

Solution

$[\mathbf{x}^*\mathbf{b} - 1.96\,\mathrm{SE}_{\mathrm{pred}}, \mathbf{x}^*\mathbf{b} + 1.96\,\mathrm{SE}_{\mathrm{pred}}]$, where $\mathrm{SE}_{\mathrm{pred}} = s\sqrt{1 + \mathbf{x^*}(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{x^{*T}}}$. $\mathbf{x^*}$ is the new observed value and $\mathbb{X}$ is the design matrix. s is an approximation of σ, the standard deviation of the errors.

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

Solution

The confidence and the prediction intervals should both contain the predicted value, $\mathbf{x^*b}$. The prediction interval should contain the confidence interval, i.e. the prediction interval should be wider than the confidence interval. The interval should look reasonable, based on inspecting a plot of the data.

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

Solution

The 95% confidence interval for the relationship between oxygen levels and water temperature is

$$b_4 \pm 1.96\,\mathrm{SE}(b_4) = 1.956 \pm 1.96(0.024)$$

This evaluates to $[1.909, 2.003]$.

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.

Solution

Null and alternative hypotheses are

$$H_0 : \beta_4 = 0$$
$$H_a : \beta_4 \neq 0$$

Write the sample model as $\mathbf{y} = \mathbb{X}\mathbf{b} + \mathbf{e}$ with $\mathbf{b} = (b_0, b_1, b_2, b_3, b_4)$. Then, $\mathbf{b} = (\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y}$ and the usual test statistic is b_4.

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

Solution

Under the null hypothesis, the model-generated test statistic, $\hat{\beta}_4$ has a normal distribution with a mean of 0 and a standard deviation of 0.00052, i.e. $\hat{\beta}_4 \sim \text{normal}(0, 0.00052)$.

(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.

Solution

$\mathbf{x}^* = [1, 8, 14, 90, 40]$. The predicted value is $\mathbf{x}^*\mathbf{b} = 0.41495 + 0.18453(8) + 0.04800(14) + 0.01304(90) + 0.00134(40) = 3.79039$

(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?

Solution

We should check that our observed values for the average length of stay, the average culture count, the average number of X-rays, and the number of beds for the new hospital are similar to values observed in the data. We should also check that the predicted infection rate for the hospital makes sense given the observed explanatory variables based on similar hospitals in the data.

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.

Solution

$$H_0 : \beta_2 = 0$$
$$H_a : \beta_2 \neq 0$$

(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.

Solution

(i) Our sample test statistic is b_2, the least squares estimate of β_2.

(ii) The model-generated test statistic $\hat{\beta}_2$ under the null hypothesis has approximately a normal distribution with a mean of 0 and a standard deviation of 0.254, i.e. $\hat{\beta}_2 \sim$ normal$(0, 0.254)$.

(iii) We do not have evidence to reject our null hypothesis from part (a) because our standard error is about the same size as our estimate. For example, if we were to construct a 95% confidence interval for β_2, the confidence interval would include 0.

(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.

Solution The linear combination for the prediction is given by

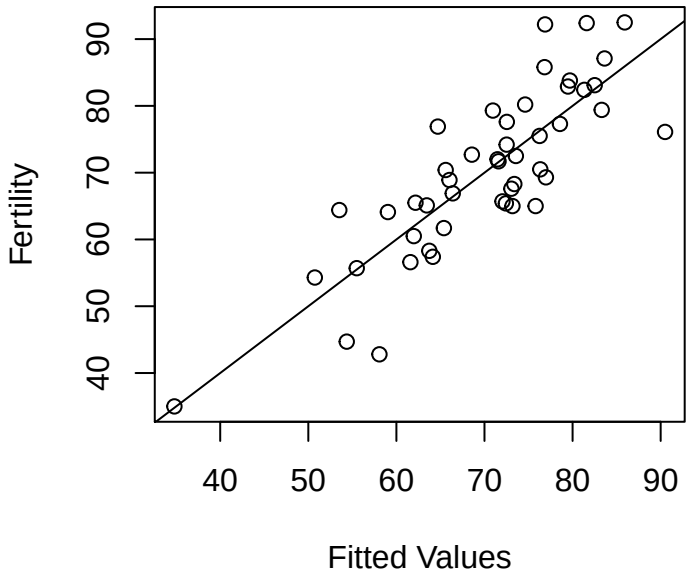
$$\mathbf{x}^* = [1, 70, 22, 10, 50, 20].$$

The predicted value is

$$\mathbf{x}^*\mathbf{b} = 66.915 - 0.172(70) - 0.258(22) - 0.871(10) + 0.104(50) + 1.077(20)$$

(This numeric expression evaluated to 67.229.)

(d)




```
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.

Solution

The model is a relatively good fit for the data. There are no trends or clusters in the plot of the fitted values against the fertility rate. However, there is a province that has a very low fertility rate compared to the other provinces. We may have some concerns about using our model to make our prediction, because of this potential outlier province.

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.

```
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.

```
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

Solution:

μ_1 is the population mean frontal lobe size for blue crabs. μ_2 is the population mean frontal lobe size for orange crabs.

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

Solution:

$(14.056 - 1.96 * 0.315, 14.056 + 1.96 * 0.315) = (13.44, 14.67)$

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

Solution: All of the first six crabs are blue. Therefore the design matrix is given by:

$$\mathbb{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \end{bmatrix}$$

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."

```
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 $\mathbb{X}$.

```
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.

Solution: Let $\mathbf{x} = (x_1, \ldots, x_{24})$ be a dummy variable for high fat diet. That is $x_i = 1$ if `Diet` for observation i is hf and 0 if `Diet` is chow. Let $\mathbf{y} = (y_1, \ldots, y_{24})$ be the weights of the 24 mice, and $\mathbf{e} = (e_1, \ldots, e_{24})$ be the corresponding residuals. Finally, let b_0 be the intercept and b_1 be the sample coefficient corresponding to a high fat diet.

The sample linear model is given by $y_i = b_0 + b_1 x_i + e_i$ for $i = 1, \ldots, 24$.

(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.

Solution: The mean of the "chow" group is given by the intercept, b_0. The mean of the "hf" group is given by $b_0 + b_1$. The difference between these two means is given by b_1.

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.

```
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?

Solution:

0.44124 is the sample mean sales for Activision and 0.41361 is the sample mean sales for Electronic Arts.

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

Solution: The first four games were published by Activision, and the next two by EA. We therefore have:

$$\mathbb{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ \vdots & \vdots \end{bmatrix}$$

(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?

Solution: In this model, b_0 corresponds to the sample mean sales of Electronic Arts, which is equal to $m_2 = 0.41361$. On the other hand, $b_0 + b_1$ corresponds to the sample mean for Activision, which is equal to $m_1 = 0.44124$. We therefore have $b_1 = m_1 - m_2 = 0.44124 - 0.41361$

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.

Solution: The fitted model contains 4 variables: an intercept, a dummy variable for 6 cylinders, a dummy variable for 8 cylinders, and the horsepower. As an example, since observation 1 is 4 cylinders, x_{11} and x_{12} are both equal to 0. The design matrix is:

$$\mathbb{X} = \begin{bmatrix} 1 & 0 & 0 & 67 \\ 1 & 0 & 0 & 98 \\ 1 & 0 & 0 & 88 \\ 1 & 0 & 1 & 150 \\ 1 & 0 & 0 & 86 \\ 1 & 1 & 0 & 107 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

(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.

Solution:

Because this new observation has 6 cylinders, the value of $x^*_{i2} = 1$ and $x^*_{i3} = 0$. Thus $\mathbf{x}^* = \begin{bmatrix} 1 & 1 & 0 & 110 \end{bmatrix}$. The predicted value is $\mathbf{x}^*\mathbf{b} = 37.27 - 6.94 + 110 \times -0.102$.

(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?

Solution:

The interpretation of the parameter β_2 is the difference in means between 8 cylinder cars and 4 cylinder cars for a fixed horsepower level. We therefore wish to test $H_0 : \beta_2 = 0$ against $H_a : \beta_2 < 0$. This is written as a 1-sided test; it is a judgement call whether to consider also the possibility that $\beta_2 > 0$.
