Quiz 2, STATS 401 F18

In lab on 11/16

PRELIMINARY VERSION. QUESTIONS NEED TO BE REWRITTEN AND/OR REARRANGED. 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. No new questions will be added after Wednesday 11/14. Small changes may be made.

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

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

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

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

(3) $\mathrm{Cov}(X,Y) = \mathrm{E}\big[\big(X-\mathrm{E}[X]\big)\big(Y-\mathrm{E}[Y]\big)\big] = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]$

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

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

(6) 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%.

(7) Syntax from `?pnorm`:

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

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

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

Q1-1.

In the sample regression line $\hat{y} = b_1x + b_2$, 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 $\hat{y} = b_1x + b_2$ denotes a line corresponding to the least squares fit for a sample, and substuting $x = 0$ gives $\hat{y} = b_2$.

Q1-2.

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.

Q1-3.

Sometimes a histogram of the residuals deviates considerably from a normal curve, indicating violation of the modeling assumption of normal errors for a linear model. This violation is more problematic for a confidence intervals on a prediction mean than for a prediction interval.

Solution. FALSE. A central limit property applies to the prediction mean - it is the sum of small contributions from many data points. Therefore, a normal approximation is appropriate for the confidence interval even when the residuals indicate non-normality. The prediction interval is dominated by a single measurement error, so is not rescued by a central limit property.

Q1-4.

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.

Q1-5.

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.

Q1-50.

When the fitted values $\hat{y}_1, \dots \hat{y}_n$ and the actual values $y_1, \dots, y_n$ 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 issue.

Q1-6.

We should use 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.

Q1-7.

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.

Q1-80.

`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.

Q1-81.

`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`.

Q1-82.

`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`.

Q1-83.

`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`.

Q1-84.

`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
```

Q1-90.

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.

Q1-91.

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!

Q1-92.

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.

Q1-93.

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.

Q1-94.

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.

Q1-95.

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 the usual probability model for a linear model (which you don't have to write out here) and using a normal approximation, find an expression for the probability that the difference between the coefficient estimate for the data (0.03721) 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:

```
1-pnorm(0.03721,mu=0.03538,sd=0.005939)
```

gives the probability of observing a bigger value of the estimated coefficient under the assumed model, 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=0.03538,sd=0.005939))
```

Q2-2.

Let $X_1, X_2, \dots, X_n$ 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. Find the mean and variance of X_1. Use this to find the mean and variance of $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Now suppose $n = 100$ and suppose that $\bar{X}$ is well approximated by a normal distribution. 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 $X_1, X_2, \dots, X_n$ 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 $X_1, X_2, \ldots, X_n$ be independent random variables each of which take the value 0 with probability 0.25, and 4 with probability 0.75. Find the mean and variance of X_1. Use this to find the mean and variance of $X = \sum_{i=1}^{n} X_i$. Now suppose $n = 200$ and suppose that X is well approximated by a normal distribution. Find a number c such that $\mathrm{P}[X < c]$ 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:

$\mathbb{E}(X_1) = 0 \times 0.25 + 4 \times 0.75 = 3$

$\mathbb{E}(X_1^2) = 0 \times 0.25 + 4^2 \times 0.75 = 12$

$Var(X_1) = \mathbb{E}(X_1^2) - (\mathbb{E}(X_1))^2 = 12 - 9 = 3$

Thus, $\mathbb{E}(X) = \mathbb{E}(\sum_{i=1}^{n} X_i) = n\mathbb{E}X_1 = 600$

$Var(\bar{X}) = Var(\sum_{i=1}^{n} X_i) = nVar(X_1) = 600$

$c = qnorm(0.9, 600, sqrt(600))$

Q2-5.

Let $X_1, X_2, \ldots, X_n$ be independent random variables each of which has possible values 0, 1 and -1. The probability of taking 0 is 0.2 and the probability of 1 is 0.4. Find the mean and variance of $X = \frac{1}{n}\sum_{i=1}^{n} X_i$. Now suppose $n = 100$ and suppose that X is well approximated by a normal distribution. Find a number c such that $\mathrm{P}[X > c]$ is approximately 0.8. 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

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. Calculate the predicted value.

Solution

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

$\hat{y}^* = 195.894 + 100(-0.135) + 1000(-0.045) = 137.394$

(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 $SE_{pred} = s\sqrt{\mathbf{x^*}^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{x^*} + 1}$, where s is the residual standard error.

SE_{pred} is an estimate of $Var[\mathbf{Y} - \mathbf{x}^*\boldsymbol{\beta}]$.

Thus the P.I. is

$$\hat{y}^* \pm 1.96 SE_{pred}$$

.

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

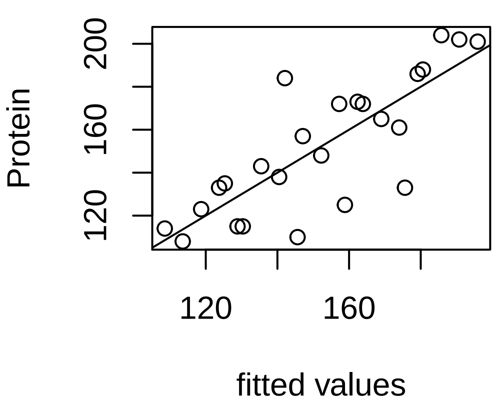
Solution

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

$$\begin{aligned} \hat{\beta}^* &\pm 1.96 SE(\hat{\beta}) \\ -0.045 &\pm 1.96(0.007) \\ &[-0.058, -0.031] \end{aligned}$$

.

(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; 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 observed a new outcome $\mathbf{x^*}$

(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

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

(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

$[\mathbf{x^*}\hat{\boldsymbol{\beta}} - \mathbf{1.96SE_{pred}}]$, where $SE_{pred} = s\sqrt{1 + \mathbf{x^*}(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{x^{*T}}}$. $\mathbf{x^*}$ 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, $\mathbf{x^*}\hat{\boldsymbol{\beta}}$. The prediction interval should contain the confidence interval, i.e. the prediction interval should be wider than the confidence interval. The predicted temperature should be reasonable. Check the data.

(d) Calculate 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

$$\begin{aligned} \hat{\beta}^* &\pm 1.96SE(\hat{\beta}) \\ 1.956 &\pm 1.96(0.024) \\ &[1.909, 2.003] \end{aligned}$$

.

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, $vectx_3$ be the number of X-rays, and $vectx_4$ is 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 $i = 1, \ldots, 113$ where ϵ_i are 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 we would use to answer this question.

Solution

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

(b) What is the distribution of your test statistic from a?

Solution

Under the null hypothesis, our test statistic, b_1 has a normal distribution with a mean of 0 and a standard deviation of 0.00052, i.e. $b_1 \sim N(0, 0.00052)$.

(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 the predicted value for this observation; you do not need to simplify.

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.

We consider a subset of the National Education Longitudinal Study of 1988 which examined schoolchildren's performance on a math test score in 8th grade. `ses` is the socioeconomic status of parents and `paredu` is the parents highest level of education achieved (less than high school, high school, college, BA, MA, PhD). The dataset called `nels88` starts as follows:

```r
head(nels88)
```

```
##      sex  race   ses paredu math
## 1 Female White -0.13     hs   48
## 2   Male White -0.39     hs   48
## 3   Male White -0.80     hs   53
## 4   Male White -0.72     hs   42
## 5 Female White -0.74     hs   43
## 6 Female White -0.58     hs   57
```

We fit a regression model to the data. The rounded co-efficients for the model are provided below:

```r
fit <- lm(math ~ ses + paredu, data = nels88)
round(summary(fit)$coef)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)          59          2      33        0
## ses                   3          1       2        0
## pareducollege        -8          2      -4        0
## pareduhs            -12          3      -5        0
## paredulesshs        -13          3      -4        0
## pareduma             -1          2       0        1
## pareduphd            -2          3      -1        0
```

(a) Describe a suitable probability model, in matrix form, to give a sample version of the linear model that has been fit above.

Solution:

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

where

- $\mathbf{y} = (y_1, \ldots, y_n)$ is a vector random variable modeling schoolchildren's performance on a math test in 8th grade.
- $\mathbb{X} = [x_{ij}]$ is a $n \times 7$ matrix with $x_{i1} = 1$ for $i = 1, \ldots, n$, x_{i2} is the parents' socioeconomic status for student i, x_{i3} equals 1 if 'paredu' = college and 0 otherwise, x_{i4} equals 1 if 'paredu' = high school and 0 otherwise, x_{i5} equals 1 if 'paredu' = below high school and 0 otherwise, x_{i6} equals 1 if 'paredu' = MA and 0 otherwise, and x_{i7} equals 1 if 'paredu' = PhD and 0 otherwise.
- $\mathbf{b} = (b_1, \ldots, b_7)$ are the true but unknown vector of coefficients.
- $\mathbf{e} = (e_1, \ldots, e_n)$ is a vector random variable modeling chance variation.
- All vectors are interpreted as column vectors.

(b) Find the predicted math score for a student whose family has an ses value of -0.5 and whose parents' highest education level is high school (`hs`).

Solution:

$\hat{y} = 59 + 3(-0.5) - 8(0) - 12(1) - 13(0) - 1(0) - 2(0)$

$\hat{y} = 59 - 1.5 - 12$

$\hat{y} = 45.5$

The predicted math score for this student is 45.5.

(c) How is the residual standard error calculated for this model? (Give a formula).

Solution:

$s = \sqrt{\frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{\frac{1}{n-p}\sum_{i=1}^{n}(y_i - [\mathbb{X}\mathbf{b}]_i)^2}$

where

- $n - p$ is the degrees of freedom in the model; p is equal to 7
- y_i is the observed math score in 8th grade for student i
- $\hat{y}_i$ is the predicted math score in 8th grade for student i from the model above.
- $\mathbb{X} = [x_{ij}]$ is a $n \times 7$ matrix with $x_{i1} = 1$ for $i = 1, \ldots, n$, x_{i2} is the parents' socioeconomic status for student i, x_{i3} equals 1 if 'paredu' = college and 0 otherwise, x_{i4} equals 1 if 'paredu' = high school and 0 otherwise, x_{i5} equals 1 if 'paredu' = below high school and 0 otherwise, x_{i6} equals 1 if 'paredu' = MA and 0 otherwise, and x_{i7} equals 1 if 'paredu' = PhD and 0 otherwise.
- $\mathbf{b} = (b_1, \ldots, b_7)$ are the estimated coefficients.

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

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 the subscript format. 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.

```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 $i = 1, \ldots, 580$.

```r
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.

```r
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:

```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 $H_a : \beta_2 < 0$
