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.

Formulas

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

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

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

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

(4) Syntax from `?pnorm`:

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

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

Q1. Say whether the following statements are TRUE or FALSE. No explanation is necessary.

Q1-1.

In the sample regression line $\hat{y} = b_1 x + b_2$, 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 $\hat{y} = b_1 x + b_2$ 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.

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 $X_1, X_2, \dots, X_n$ 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. The data are as follows.

```r
head(toxicity)
```

```
##   Copper Zinc Protein
## 1      0    0     201
## 2      0  375     186
## 3      0  750     173
## 4      0 1125     110
## 5      0 1500     115
## 6     38    0     202
```

```r
lm_toxicity <- lm(Protein~Copper+Zinc,data=toxicity)
round(coef(summary(lm_toxicity),3))
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)      196          9      23        0
## Copper             0          0      -2        0
## Zinc               0          0      -6        0
```

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 $\mathbf{x}^*$ so that $\mathbf{y}^* = \mathbf{x}^*\mathbf{b}$ gives a least squares prediction of the new observation.

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

(c) Explain briefly some things you would look for to check whether your prediction interval is reasonable.

Solution:

(a).

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

(b).

$$\begin{aligned} Var(\hat{y}^* + \epsilon_{new}) &= Var(\mathbf{x}^{*T}\hat{\beta} + \epsilon_{new}) \\ &= Var(\mathbf{x}^{*T}\hat{\beta}) + Var(\epsilon_{new}) \\ &= \sigma^2(\mathbf{x}^{*T}(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{x}^* + 1) \end{aligned} \tag{1}$$

Thus the P.I. is

$$\hat{y}^* \pm 1.96s\sqrt{\mathbf{x}^{*T}(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{x}^* + 1},$$

where s is the residual standard error.

(c).

We can check if $\mathbf{x}^*$ is within the support of the given data set to avoid extrapolation; The prediction interval should be large than the confidence intervel; $\hat{y}^*$ should fall within the predection intervel.

Q3-2.

Consider the birth weight data set we have seen in lab. For this question, we will look at columns `bwt` (birth weight), `lwt` (mother's weight), `age` (mother's age) and `race` (mother's race, 1 for white, 2 for black and 3 for other).

```r
library(MASS)
data(birthwt)
head(birthwt,3)
```

```
##    low age lwt race smoke ptl ht ui ftv  bwt
## 85   0  19 182    2     0   0  0  1   0 2523
## 86   0  33 155    3     0   0  0  0   3 2551
## 87   0  20 105    1     1   0  0  0   1 2557
```

```r
lm_bw <- lm(bwt ~ lwt + age +factor(race), data = birthwt)
summary(lm_bw)$coef
```

```
##                   Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)    2461.147482 314.722327  7.8200600 3.968682e-13
## lwt               4.619545   1.787729  2.5840294 1.054066e-02
## age               1.298831  10.107701  0.1284991 8.978943e-01
## factor(race)2  -447.614691 161.369310 -2.7738527 6.110757e-03
## factor(race)3  -239.356515 115.188920 -2.0779474 3.910220e-02
```

Now suppose we are interested in predicting the birthweight of a baby who has a 30-year-old white mother with weight 130.

(a) Specify a row matrix $\mathbf{x}^*$ so that $\hat{y}^* = \mathbf{x}^*\mathbf{b}$ gives the least square predictor.

(b) Write a matrix expression for the variance of $\hat{Y}^* = \mathbf{x}^*\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}}$ is the least squares fit on model-generated data, i.e., $\hat{\boldsymbol{\beta}} = (\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}\mathbb{X}^{\mathsf{T}}\mathbf{Y}$.

Q3-3.

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.

```
vg <- head(vg)
```

Consider the probability model $Y_{ijk} = \alpha + \beta_j + \gamma_k + \epsilon_{ijk}$ where $j = 1, 2, 3$ specifies the genre (shooter, sports and action, respectively), $k = 1, 2$ gives the publisher (Electronic Arts and Activision, respectively), and i ranges over all the games in each (j, k) category. In order to code these factors, we set $\beta_1 = \gamma_1 = 0$. As usual, ϵ_{ijk} gives an independent $N[0, \sigma]$ error for game (i, j, k). Parameters in this probability model are estimated by least squares as follows:

```
lm_vg1 <- lm(Sales ~ Publisher + Genre, data = vg)
summary(lm_vg1)
```

```
##
## Call:
## lm(formula = Sales ~ Publisher + Genre, data = vg)
##
## Residuals:
##       1       2       3       4       5       6
##  3.8925  0.1825 -0.9975 -3.0775 -1.5100  1.5100
##
## Coefficients: (1 not defined because of singularities)
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  5.808      1.374   4.226   0.0134 *
## PublisherElectronic Arts   -3.688      2.380  -1.549   0.1962
## GenreSports                    NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.748 on 4 degrees of freedom
## Multiple R-squared:  0.375,  Adjusted R-squared:  0.2188
## F-statistic:   2.4 on 1 and 4 DF,  p-value: 0.1962
```

Note that the output of `summary(lm_vg1)` tells you that R is using $\boldsymbol{\beta} = (\alpha, \beta_2, \beta_3, \gamma_2)$ as the parameter vector.

(a) Write the first six lines of the design matrix $\mathbb{X}$ in the matrix version of the linear model $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Hint: the output from `head(vg)` tells you what the values of j and k are for each of the first six observations.

(b) Suppose we're interested in the predicting the North American Sales of a shooting game released by Activision. Specify a row matrix $\mathbf{x}^*$ such that $y^* = \mathbf{x}^*\mathbf{b}$ gives the least square predictor of this quantity.

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. Here, `BD` refers to the body depth of the crabs. 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 of frontal lobe size (`FL`) of 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 fitted 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)
```

```
##
## Call:
## lm(formula = FL ~ sp - 1, data = crabs)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.010 -2.410  0.390  2.169  7.244
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## spB   14.056      0.315   44.62   <2e-16 ***
## spO   17.110      0.315   54.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.15 on 198 degrees of freedom
## Multiple R-squared:  0.9615, Adjusted R-squared:  0.9611
## F-statistic:  2470 on 2 and 198 DF,  p-value: < 2.2e-16
```

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

(b) Build a 95% confidence interval for μ_1 using normal approximation

(c) Recall in homework we know that the full estimated covariance matrix of $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \hat{\mu}_2)$ can be found by

```
V <- summary(lm_crab)$cov.unscaled * summary(lm_crab)$s^2
V
```

```
##            spB        spO
## spB 0.09923719 0.00000000
## spO 0.00000000 0.09923719
```

Use V and information provided in `summary(lm_crab)` to write down an expression that constructs a 95% confidence interval for $\mu_1 - \mu_2$.

Solution:

(a).

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).

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

(c).

Let $a = (1, -1)^T$.

$Var(a^T\hat{\mu}) = a^T Var(\hat{\mu}) a = a^T V a = 0.198$

$\hat{\mu}_1 - \hat{\mu}_2 = 14.056 - 17.110 = -3.054$

Thus we have the 95% C.I. $(-3.054 - 1.96 * \sqrt{0.198}, -3.054 + 1.96 * \sqrt{0.198}) = (-3.926, -2.182)$

Q4-2.

Consider the following linear model for the mouse diet data that we have studied repeatedly

```
mice <- read.table(paste(my_dir,"/femaleMiceWeights.csv",sep=""),sep=",",header=TRUE)
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
```

```
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.

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

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. We consider the following analysis

```
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
```

```
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
```

The fitted probability model is $Y_{ij} = \pi_j + \epsilon_{ij}$ where $j = 1, 2$ specifies the publisher (Electronic Arts and Activision, respectively), and i ranges over all the games for each publisher. As usual, ϵ_{ij} gives an independent $N[0, \sigma]$ error for game (i, j). Parameters in this probability model are estimated by least squares as follows:

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

(b) Explain how to build a 95% confidence interval for Activision sales using a normal approximation. You can use the property that $\mathrm{P}[Z < 1.96] = 0.975$ when Z has a $N[0,1]$ distribution.

Q4-4.

We consider a dataset of measurements on crabs. The start of the dataset `crabs` is shown below. Here, `BD` refers to the body depth of the crabs, and `sp` denotes the colour of the crabs, which is one of blue or orange.

```
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
```

```
crabs$mu1 <- (crabs$sp == "B")*1
crabs$mu2 <- (crabs$sp == "O")*1
crabs$mu3 <- 1
crabs$mu4 <- 1-crabs$mu1
crabs$mu_diff <- crabs$mu2
fit1 <- lm(BD ~ mu1+mu2-1, data = crabs)
fit2 <- lm(BD ~ mu3 + mu_diff - 1, data = crabs)
fit3 <- lm(BD ~ mu2, data = crabs)
fit4 <- lm(BD ~ 1-mu1, data = crabs)
fit5 <- lm(BD ~ mu4, data = crabs)
fit6 <- lm(BD~  mu1+mu2, data = crabs)
```

(a) Would any of the models (`fit1` to `fit6`) give the same coefficients? If yes, list them.

Solution:

Yes, `fit2`, `fit3`, and `fit5` would provide the same coefficients. `fit2` would give an intercept term (`mu3`) and an estimate of the difference in body depth between the orange and the blue crabs (`mu2 - mu1`) which is coded as `mu2` or `mu_diff`. From this, we can determine that `fit3` would also provide the same model as `fit2` since by default R includes an intercept value. Similarly, we can determine that `fit5` would provide the same model as `fit2` and `fit3` since `mu4` is the same as `mu2`.

The following are the coefficients obtained from each of the models (for comparison purposes only):

```
summary(fit1)$coef
```

```
##     Estimate Std. Error  t value      Pr(>|t|)
## mu1   12.583  0.3109965 40.46026  1.113535e-97
## mu2   15.478  0.3109965 49.76905 6.839989e-114
```

```
summary(fit2)$coef
```

```
##         Estimate Std. Error   t value     Pr(>|t|)
## mu3       12.583  0.3109965 40.460262 1.113535e-97
## mu_diff    2.895  0.4398155  6.582306 4.059563e-10
```

```
summary(fit3)$coef
```

```
##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   12.583  0.3109965 40.460262 1.113535e-97
## mu2            2.895  0.4398155  6.582306 4.059563e-10
```

```
summary(fit4)$coef
```

```
##             Estimate Std. Error  t value      Pr(>|t|)
## (Intercept)  14.0305   0.242168 57.93706 1.524559e-126
```

```
summary(fit5)$coef
```

```
##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   12.583  0.3109965 40.460262 1.113535e-97
## mu4            2.895  0.4398155  6.582306 4.059563e-10
```

```
summary(fit6)$coef
```

```
##             Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)   15.478  0.3109965 49.769049 6.839989e-114
## mu1           -2.895  0.4398155 -6.582306  4.059563e-10
```

Now consider the probability model $Y_i = \mu_1 x_{Bi} + \mu_2 x_{Oi} + \epsilon_i$, where $i = 1, \ldots, 200$. Y_i models the body weight of observation i. x_{Bi} is 1 if `sp=B` for observation i and 0 otherwise. Similarly, x_{Oi} is 1 if `sp=O` for observation i and 0 otherwise. $\epsilon_1, \ldots, \epsilon_{200}$ are i.i.d with mean 0 and variance σ^2. This model can be fitted to the `crabs` dataset in R using the `lm()` function. The resulting summary is provided below.

```
summary(fit1)
```

```
##
## Call:
## lm(formula = BD ~ mu1 + mu2 - 1, data = crabs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0780 -2.1830  0.0695  2.3170  7.4170
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## mu1   12.583      0.311   40.46   <2e-16 ***
## mu2   15.478      0.311   49.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.11 on 198 degrees of freedom
## Multiple R-squared:  0.9541, Adjusted R-squared:  0.9536
## F-statistic:  2057 on 2 and 198 DF,  p-value: < 2.2e-16
```

(b) Interpret μ_1 and μ_2 in the above model?

Solution:

μ_1 is the mean body depth of blue crabs, and μ_2 is the mean body depth of orange crabs.

(c) Recall from homework that the estimated covariance matrix of $\hat{\boldsymbol{\beta}} = (\hat{\mu}_1, \hat{\mu}_2)$ can be found by

```
V <- summary(fit1)$cov.unscaled * summary(fit1)$s^2; V
```

```
##            mu1        mu2
## mu1 0.09671882 0.00000000
## mu2 0.00000000 0.09671882
```

Construct a 95% confidence interval for $\mu_1 - \mu_2$ using normal approximation. Based on this, do we have sufficient evidence to conclude that $\mu_1 = \mu_2$ at the 95% level?

Solution:

1. Finding the variance of $\mu_1 - \mu_2$

$Var(\mathbb{A}\mathbf{Y}) = \mathbb{A}Var(\mathbf{Y})\mathbb{A}^T$

$\mathbb{A} = \begin{bmatrix}1-1\end{bmatrix}$

$$\begin{bmatrix}1-1\end{bmatrix}\begin{bmatrix}0.09671882 & 0.00000000\\ 0.00000000 & 0.09671882\end{bmatrix}\begin{bmatrix}1\\ -1\end{bmatrix}$$

$$\begin{bmatrix}0.09671882 & -0.09671882\end{bmatrix}\begin{bmatrix}1\\ -1\end{bmatrix}$$

$0.09671882 + 0.09671882 = 0.1934376$

2. Constructing the confidence interval.

$(\mu_1 - \mu_2) \pm z_{\frac{\alpha}{2}} * sd(\mu_1 - \mu_2)$

$(12.583 - 15.478) \pm 1.96 * \sqrt{0.1934376}$

$(-3.757038, -2.032962)$

No; we do not have enough evidence to conclude that $\mu_1 = \mu_2$ ath the 95% level.