Quiz 2, STATS 401 W18

In lab on 3/29 or 3/30

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.

Instructions. You have a time allowance of 50 minutes, though the quiz may take you much 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 will be 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) $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}$

(3) $\mathrm{Var}(\mathbb{A}\mathbf{Y}) = \mathbb{A}\mathrm{Var}(\mathbf{Y})\mathbb{A}^{\top}$

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

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

(6) The binomial (n, p) distribution has mean np and variance $np(1-p)$.

From `?pnorm`:

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

Q1. Calculating means and variances, and making a normal approximation

Q1-1. 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.

```
gpa <- read.table("gpa.txt",header=T)
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))
```

Q1-2. 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. 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.

Q1-3. 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 might notice that this calculation is related to the binomial distribution. You can use that to check your work, if you like, 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.

Q1-4. 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.

Q1-5. 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.

Q2. Prediction using a linear model

Q2-1. To investigate the consequences of metal poisoning, 25 beakers of minnow larvae were exposed to varying levels of copper and zinc. The data were

```
toxicity <- read.table("toxicity.txt")
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
```

```
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 $\boldsymbol{y} = \mathbb{X}\boldsymbol{b} + \boldsymbol{e}$. 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 $\boldsymbol{x}^*$ so that $\boldsymbol{y}^* = \boldsymbol{x}^*\boldsymbol{b}$ gives a least squares prediction of the new observation.

(b) Explain how to use the data vector $\boldsymbol{y}$, the design matrix $\mathbb{X}$, and your row vector $\boldsymbol{x}^*$ 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.

Q2-2. 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).

```
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
```

```
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 $\boldsymbol{x}^*$ so that $\hat{y}^* = \boldsymbol{x}^*\boldsymbol{b}$ gives the least square predictor.

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

Q2-3. 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 <- read.table("vg_sales.txt") ; 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
```

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:
##     Min      1Q  Median      3Q     Max
## -0.8444 -0.2662 -0.1352  0.0858  8.8556
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.271127   0.061346   4.420 1.18e-05 ***
## PublisherElectronic Arts -0.004955   0.071076  -0.070    0.944
## GenreShooter               0.573315   0.095061   6.031 2.91e-09 ***
## GenreSports                0.118062   0.077585   1.522    0.129
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7818 on 576 degrees of freedom
## Multiple R-squared:  0.06154,    Adjusted R-squared:  0.05665
## F-statistic: 12.59 on 3 and 576 DF,  p-value: 5.546e-08
```

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 $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. 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 $\boldsymbol{x}^*$ such that $y^* = \boldsymbol{x}^*\boldsymbol{b}$ gives the least square predictor of this quantity.

Q2-4. Recall the nels88 dataset from lab. These data are 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 data were as follows:

```
library(faraway)
data(nels88)
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:

```
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:

$\mathbf{Y} = \mathbb{X}\beta + \epsilon$

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.

- $\beta = (\beta_1, \ldots, \beta_7)$ are the true but unknown vector of coefficients.
- $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ is a vector random variable modeling chance variation. It follows the measurement error model, with $E[\epsilon] = 0$ and $Var(\epsilon) = \sigma^2 \mathbb{I}$ where $\mathbb{I}$ is the $n \times n$ identity matrix and σ is the measurement standard deviation.
- 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
- 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.

Q3. Comparing means using a linear model

Q3-1. Consider the following linear model for the mouse diet data that we have studied repeatedly

```
mice <- read.table("femaleMiceWeights.csv",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.

Q3-2. Let's consider the crabs data set we studied in lab. Recall that species (`sp`) 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.

```
library(MASS)
data(crabs)
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 $\mu_1 - \mu_2$.

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. We consider the following analysis

```
vg <- read.table("vg_sales.txt") ; 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.

Q3-4. Recall the crabs dataset used in lab. 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.

```
library(MASS)
data(crabs)
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.

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

$12.583 \pm 1.96 * \sqrt{0.1934376}$

$(11.72096, 13.44504)$

Q4. Making and interpreting an F test

Q4-1. The following is an ANOVA analysis of the football field goal kicking data that we have seen repeatedly. Recall that `Name` is the name of the kicker, `FGt` is the field goal percentage for the kicker in that year, and `FGtM1` is the percentage for that kicker in the previous year.

```
kickers <- read.table("FieldGoals2003to2006.csv",header=T,sep=",")
kickers[1:5,c("Name","Teamt","FGt","FGtM1")]
```

```
##             Name Teamt  FGt FGtM1
## 1 Adam Vinatieri    NE 73.5  90.0
## 2 Adam Vinatieri    NE 93.9  73.5
## 3 Adam Vinatieri    NE 80.0  93.9
## 4 Adam Vinatieri   IND 89.4  80.0
## 5    David Akers   PHI 82.7  88.2
```

```
lm_kickers <- lm(FGt~FGtM1+Name,data=kickers)
anova(lm_kickers)
```

```
## Analysis of Variance Table
##
## Response: FGt
##           Df Sum Sq Mean Sq F value    Pr(>F)
## FGtM1      1   87.2  87.199  2.2597 0.1383978
## Name      18 2252.5 125.137  3.2429 0.0003858 ***
## Residuals 56 2161.0  38.589
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) Write out the null and alternative hypotheses corresponding to the hypothesis test constructed in the `Name` row of the ANOVA table.

(b) Describe how this test is constructed, using formulas when appropriate. You may define a residual sum of squares in words, without giving an explicit formula.

(c) Interpret the outcome of this test.

Q4-2. 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).

```
library(MASS)
data(birthwt)
```

```
head(birthwt)
```

```
##    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
## 88   0  21 108    1     1   0  0  1   2 2594
## 89   0  18 107    1     1   0  0  1   0 2600
## 91   0  21 124    3     0   0  0  0   0 2622
```

We want to study the relationship between birthweight and race using an F test, while mother's weight and age are included as additional explanatory variables. Let the null hypothesis, H_0, be the probability model where birth weight is modeled to depend linearly on mother's weight and age. Let H_a be the probability model where H_0 is extended to include race as a factor, as fitted in R by

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

The results from summary(lm_bw) and anova(lm_bw) are as follows

```
summary(lm_bw)
```

```
##
## Call:
## lm(formula = bwt ~ lwt + age + factor(race), data = birthwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2103.50  -429.68    41.74   486.10  1902.20
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2461.147    314.722   7.820 3.97e-13 ***
## lwt              4.620      1.788   2.584  0.01054 *
## age              1.299     10.108   0.128  0.89789
## factor(race)2 -447.615    161.369  -2.774  0.00611 **
## factor(race)3 -239.357    115.189  -2.078  0.03910 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 704.9 on 184 degrees of freedom
## Multiple R-squared:  0.08536,    Adjusted R-squared:  0.06548
## F-statistic: 4.293 on 4 and 184 DF,  p-value: 0.00241
```

```
anova(lm_bw)
```

```
## Analysis of Variance Table
##
## Response: bwt
##               Df   Sum Sq Mean Sq F value   Pr(>F)
## lwt            1  3448639 3448639  6.9398 0.009148 **
## age            1   334183  334183  0.6725 0.413247
## factor(race)   2  4750632 2375316  4.7799 0.009467 **
## Residuals    184 91436202  496936
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) Write out the null and alternative hypotheses of the F test by completely specifying the probability models.

Solution:

$H_0 : Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, i = 1, \ldots, 189$

$H_a : Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \epsilon_i, i = 1, \ldots, 189$

where Y_i is the birth weight of infant i, X_{i1} is the weight of mother i, X_{i2} is the age of mother i, X_{i3} is the race of mother {i}, and ϵ_i is the chance variation, modeled as an independent random variable with mean zero and variance σ^2.

(b) Interpret the results in `anova(lm_gpa)`. Specifically, read the sample test statistic from R output, give the distribution of the model-generated test statistic under H_0, and explain how the resulting p-value is calculated and interpreted. Your answer should givie relevant formulas, though you may define a residual sum of squares in words without explicitly saying how it is constructed.

Solution:

The sample test statistic is $f = 4.7799$ and the model generated test statistic is $F \sim F_{d,n-p} = F_{2,184}$, the F distribution on 2 and 184 degrees of freedom. The p-value is calculated as $P(F > f) = 0.009467$. Using a level of 0.05, we reject the null hypothesis and conclude that the race of the mother has a significant association with the birth weight of the child.

The sample test statistic is calculated as follows:

$f = \frac{(RSS_0 - RSS_a)/d}{RSS_a/(n-p)}$, where RSS_0 and RSS_a are the residual sum of squares for the null and alternative models respectively, d is the difference in the degrees of freedom between the null and alternative models, and $(n-p)$ is the degrees of freedom in the alternative 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 are interested in considering whether there is a significant pattern in the sales of different game genres, which leads us to carry out the following analysis:

```
vg <- read.table("vg_sales.txt") ; 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_vg3 <- lm(Sales ~ Publisher + Genre, data = vg)
anova(lm_vg3)
```

```
## Analysis of Variance Table
##
## Response: Sales
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Publisher   1   0.11  0.1086  0.1777    0.6735
## Genre       2  22.98 11.4899 18.7971 1.236e-08 ***
## Residuals 576 352.08  0.6113
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) Let π_j be a factor for publisher, where $j = 1, 2$ specifies Electronic Arts and Activision, respectively. Let γ_k be a factor for genre, where $k = 1, 2, 3$ specifies shooter, sports and action respectively. Let y_{ijk} be the sales for the ith game having publisher j and genre k, and let Y_{ijk} be a probability model for y_{ijk}. Using this notation, complete the explicit specification of a the null and alternative hypothesis for an F-test that evaluates whether there is a statistically significant difference between sales of different genres.

(b) Explain how the test statistic and p-value are constructed for this F-test, giving relevant formulas. You may define a residual sum of squares in words, without writing out an explicit formula for this.

(c) Interpret the results of this test, as given in the above ANOVA table.

Q4-4. Recall the nels88 dataset from lab. These data are 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). `race` determines the race of each student which is one of White, Black, Asian or Hispanic.

```
library(faraway)
data(nels88)
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
```

```
fit <- lm(math ~ ses + paredu + race, data = nels88)
summary(fit)
```

```
##
## Call:
## lm(formula = math ~ ses + paredu + race, data = nels88)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.4072  -5.8638  -0.0508   5.7936  23.5985
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    58.7446     1.8514  31.730  < 2e-16 ***
## ses             2.5560     1.3663   1.871 0.062550 .
## pareducollege  -7.4794     2.1637  -3.457 0.000642 ***
## pareduhs      -12.1449     2.6950  -4.507 1.01e-05 ***
## paredulesshs  -13.1456     3.3711  -3.899 0.000124 ***
## pareduma       -0.8131     2.2451  -0.362 0.717532
## pareduphd      -1.9900     2.5241  -0.788 0.431208
## raceAsian       1.6054     3.1137   0.516 0.606583
## raceBlack      -2.4562     1.6192  -1.517 0.130553
## raceHispanic    0.8097     1.9795   0.409 0.682859
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.448 on 250 degrees of freedom
## Multiple R-squared:  0.4445, Adjusted R-squared:  0.4245
## F-statistic: 22.23 on 9 and 250 DF,  p-value: < 2.2e-16
```

```r
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: math
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## ses         1 12391.4 12391.4 173.6329 < 2.2e-16 ***
## paredu      5  1642.4   328.5   4.6029  0.000491 ***
## race        3   241.3    80.4   1.1272  0.338619
## Residuals 250 17841.4    71.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Suppose we are interested in finding whether the race of the student is associated with their math score, after controlling for the other variables.

(a) Write out the null and alternative hypothesis for carrying out the above analysis. (Describe all parameters clearly)

(b) What test would you carry out? Clearly explains how the test is constructed, giving relevant formulas. You may define a residual sum of squares in words, without giving an explicit formula.

(c) Report on the conclusions of your test based on the R output provided.
