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

(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

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 probability model in P1 and P2, 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.

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.

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.

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 $\mathrm{P}[-c < X < c]$ 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. Prediction using a linear model

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

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

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

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 $Y_{ijk} = a + b_j + c_k + \epsilon_{ijk}$ 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 $b_1 = c_1 = 0$. 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 (a, b_2, b_3, c_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 $\mathbf{Y} = \mathbb{X}\beta + \epsilon$. 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 $\curvearrowleft^*$ such that $y^* = \curvearrowleft^*$ gives the least square predictor.

Q3. Comparing means using a linear model

This question will be based on HW7. It will involve either confidence interval construction or making a hypothesis testing when comparing means of two samples in the context of a linear model.

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.

Let's consider the crabs data set we studied in lab. Recall that species(sp) is a factor with 2 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 model $Y_i = \mu_1 x_{Bi} + \mu_2 x_{Oi} + \epsilon_i, i = 1, ..., 200$. Y_i is the FL of observation i. x_{Bi} is 1 if $sp = B$ for obervation i and 0 otherwise. Similarly, x_{Oi} is 1 if $sp = O$ for obervation i 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). What do μ_1 and μ_2 measure 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 $\hat{\mu} = (\hat{\mu}_1, \hat{\mu}_2)$ 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 $\mu_1 - \mu_2$.

Q4. Making and interpreting an F test

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.

(c) Interpret the outcome of this test.

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.

(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.
