Homework 6

Tobias Boggess

2022-03-30

Chapter 9 Problems

Problem 2: Calculate and interpret a 95% confidence interval for the mean age of mothers from the Gestation data set from the mosaicData package.

Code:

```
fem_only <- select(.data = Gestation, age)
num_samp <- nrow(fem_only)
xbar <- mean(fem_only$age, na.rm = TRUE)
stand_dev <- sd(fem_only$age, na.rm = TRUE)
conf_int <- c(xbar - 2 * stand_dev/sqrt(num_samp), xbar + 2 * stand_dev/sqrt(num_samp))
round(conf_int, digits = 2)
```

```
## [1] 26.93 27.58
```

The above is the 95% confidence level for the age of the mother.

Problem 3: Use the bootstrap to generate and interpret a 95% confidence interval for the median age of mothers for the Gestation data set from the mosaicData package.

Code:

```
gest_rows <- nrow(Gestation)
B <- 1000

boot_samp_gest_med <- rep(NA, B)

for(i in 1:B) {
  resamp <- slice_sample(.data = Gestation,
                         n = gest_rows,
                         replace = TRUE)
  boot_samp_gest_med[i] <- median(resamp$age, na.rm = TRUE)
}

xbar_med <- median(boot_samp_gest_med)
xbar_med_sd <- sd(boot_samp_gest_med)

ci <- c(xbar_med - 2 * xbar_med_sd, xbar_med + 2 * xbar_med_sd)
round(ci, digits = 2)
```

```
## [1] 25.08 26.92
```

The 95% confidence level for the Gestation data set for the age is shown above.

Appendix E Problems

Problem 1: Why of the following is FALSE? Justify your answers.

Code:

```
mod <- lm(Foster ~ Biological, data = twins)
coef(mod)
```

```
## (Intercept)  Biological
##    9.207599    0.901436
```

```
summary(mod)$r.squared
```

```
## [1] 0.7779022
```

- Alice and Beth were raised by their biological parents. If Beth's IQ is 10 points higher than Alice's, then we would expect that her foster twin Bernice's IQ is 9 points higher than the IQ of Alice's foster twin Ashley.
- Roughly 78% of the foster twins' IQs can be accurately predicted by the model.
- The linear model is $\textit{Foster} = 9.2 + 0.9 \times \textit{Biological}$.
- Foster twins with IQs higher than average are expected to have biological twins with higher than average IQs as well.

The linear model bullet point is true based on the information given in the linear model given in the problem. Another correct statement in the bullet points is the 78% accuracy of the twins IQ's based on the rsquared function output. The bullet point about Alice and Beth IQ's seems accurate based on the linear model determined. So, the only false statement is the one where the foster twins having IQs higher than average have biological twins with higher IQs as well. This would be false because a biological twin could have an average IQ but given the model, the foster twins would have an IQ a little higher than the average IQ.

Problem 3: Do the following.

- a) Fit a linear regression model for birthweight (wt) as a function of the mother's age (age).
- b) Find a 95% confidence interval and p-value for the slope coefficient.
- c) What do you conclude about the association between a mother's age and her baby's birthweight?

```
# Part A linear regression model
my.file <- file.choose()
my.gest <- read.csv(my.file, header = TRUE, sep = "", stringsAsFactors = FALSE)

my.reg <- lm(wt ~ age, data = my.gest)

# Includes P Value for the slope coefficient
summary(my.reg)
```

```
##
## Call:
## lm(formula = wt ~ age, data = my.gest)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.401 -11.222   0.342  11.342  57.298
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 116.68346    2.50438  46.592   <2e-16 ***
## age           0.10622    0.08989   1.182    0.238
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.25 on 1232 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.001132,   Adjusted R-squared:  0.0003215
## F-statistic: 1.396 on 1 and 1232 DF,  p-value: 0.2375
```

```
# Part B 95% Confidence Interval
confint(my.reg, level = 0.95)
```

```
##                    2.5 %     97.5 %
## (Intercept) 111.77014517 121.596776
## age          -0.07012632   0.282573
```

The equation for the linear model: *Weight* = 116.68346 + 0.10622 * *Age*.
The *p-value* for the given linear model is "<2e-16".
The confidence interval consists of the equations: *Weight* = 111.77014517 + -0.07012632 * *Age* and *Weight* = 121.596776 + 0.282573 * *Age*.
Based on the p-value and the multiple R-squared value being so low, I would say the mother's age and the baby's birth weight are associated with one another.

Worksheet Problems

Problem 1: Do the following with the Gestation data set.

a) Fit the multiple regression model to the data. Write out the equation of the fitted multiple regression model.
Code:

```
gest.reg <- lm(wt ~ gestation + age + ht + wt.1, data = my.gest)
summary(gest.reg)
```

```
##
## Call:
## lm(formula = wt ~ gestation + age + ht + wt.1, data = my.gest)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.617 -10.598   0.475  10.181  54.406
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -86.33412   14.79126  -5.837 6.87e-09 ***
## gestation     0.46096    0.02993  15.402  < 2e-16 ***
## age           0.12771    0.08328   1.534  0.12542
## ht            1.00880    0.21061   4.790 1.88e-06 ***
## wt.1          0.07074    0.02582   2.739  0.00625 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.39 on 1179 degrees of freedom
##   (52 observations deleted due to missingness)
## Multiple R-squared:  0.2062, Adjusted R-squared:  0.2035
## F-statistic: 76.57 on 4 and 1179 DF,  p-value: < 2.2e-16
```

The equation of the fitted regression model is *Weight(baby)* = -86.33412 + 0.46096 * *Gestation(mom)* + 0.12771 * *Age(mom)* + 1.00880 * *Height(mom)* + 0.07074 * *Weight(mother)*.

b) What is the predicted weight of an infant born after a gestation period of 280 days to a 27 year old, 64 inch tall, 130 pound mother?
Code:

```
-86.33412 + 0.46096 * 280 + 0.12771 * 27 + 1.00880 * 64 + 0.07074 * 130
```

```
## [1] 119.9422
```

```
newGestPred <- data.frame(gestation = 280,
                          age = 27,
                          ht = 64,
                          wt.1 = 130)

predict(gest.reg, newdata = newGestPred)
```

```
##       1
## 119.942
```

The weight of the baby is predicted to be 119.942 ounces given the gestation period of 280, age of the mother is 27 years old, the mother's height is 64 inches, and the weight of the mother is 130 pounds.

c) By how much does the weight of an infant increase for each 1-day increase in the gestation period (holding mother's age, height, and weight constant)?
For each additional day of gestation, the weight of the infant will increase by 0.46096.

d) The p-value for a coefficient (labeled Pr(>|t|) in the summary() output) is a measure of the strength of evidence that the explanatory variable is related to the response variable – a smaller p-value indicates stronger evidence. Which of the four explanatory variables shows the strongest evidence for a relationship to infant weight? Which shows the weakest evidence?
Based on the p-values in the summary, the variable with the strongest evidence for a relationship is the gestation period. The weakest variable

e) This data set contains missing values – there are 1,236 observations (rows) in the data set, but some rows contain NAs. If a row contains NA for any one of the five variables used to build the model, lm() omits that entire row. How many rows were omitted? Hint: Look at the summary() output.
There were 52 rows deleted due to NA values.

Problem 2: Do the following with the CDI data set.

a) For each geographic region, carry out a multiple regression analysis with response variable number of serious crimes (Y) and three explanatory variables: population density(X1, total population divided by land area), per capita personal income (X2), and percent high school graduates (X3). Write out the equations of the four fitted multiple regression models.
Code:

```
my.file1 <- file.choose()
cdi.data <- read.csv(my.file1, header = TRUE, sep = "", stringsAsFactors = FALSE)
cdi.data <- mutate(.data = cdi.data, PopDen = TotPop/LandArea)

by_region <- nest_by(.data = cdi.data, Region)

models <-
  mutate(.data = by_region, cdi.mod = list(summary(lm(
    nCrimes ~ PopDen + PerCapInc + PctHSGrad, data = data
  ))))

models$cdi.mod
```

```
## [[1]]
##
## Call:
## lm(formula = nCrimes ~ PopDen + PerCapInc + PctHSGrad, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -149353   -3776     339    5091  130462
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.447e+04  4.256e+04  -1.515   0.1330
## PopDen       1.738e+01  8.336e-01  20.854   <2e-16 ***
## PerCapInc   -1.406e+00  7.606e-01  -1.849   0.0674 .
## PctHSGrad    1.183e+03  6.411e+02   1.845   0.0681 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28060 on 99 degrees of freedom
## Multiple R-squared:  0.8352, Adjusted R-squared:  0.8302
## F-statistic: 167.2 on 3 and 99 DF,  p-value: < 2.2e-16
##
##
## [[2]]
##
## Call:
## lm(formula = nCrimes ~ PopDen + PerCapInc + PctHSGrad, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -148517   -4535   -1011    4321  257651
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4163.2673 49958.0969  -0.083    0.934
## PopDen         33.6193     3.8553   8.720 4.79e-14 ***
## PerCapInc       0.1024     1.6708   0.061    0.951
## PctHSGrad      -2.7616   809.2473  -0.003    0.997
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32980 on 104 degrees of freedom
## Multiple R-squared:  0.5285, Adjusted R-squared:  0.5149
## F-statistic: 38.86 on 3 and 104 DF,  p-value: < 2.2e-16
##
##
## [[3]]
##
## Call:
## lm(formula = nCrimes ~ PopDen + PerCapInc + PctHSGrad, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -75116 -15583  -9111   3217 217684
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38862.667  30662.721   1.267   0.2070
## PopDen          5.537      2.377   2.329   0.0212 *
## PerCapInc       1.957      1.096   1.787   0.0761 .
## PctHSGrad    -670.884    511.472  -1.312   0.1917
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36970 on 148 degrees of freedom
## Multiple R-squared:  0.09251,    Adjusted R-squared:  0.07411
## F-statistic: 5.029 on 3 and 148 DF,  p-value: 0.002392
##
##
## [[4]]
##
## Call:
## lm(formula = nCrimes ~ PopDen + PerCapInc + PctHSGrad, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -101554  -25633  -14428    3976  608081
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 129323.415  95555.937   1.353    0.180
## PopDen           5.717      5.774   0.990    0.325
## PerCapInc        4.342      2.731   1.590    0.116
## PctHSGrad    -2159.920   1353.579  -1.596    0.115
##
## Residual standard error: 81820 on 73 degrees of freedom
## Multiple R-squared:  0.08665,    Adjusted R-squared:  0.04912
```

```
## F-statistic: 2.309 on 3 and 73 DF,  p-value: 0.08351
```

```
# summary(models$cdi.mod)
```

Four Equations:

Region 1

- *Number of Serious Crimes* = -6.447e+04 + 1.738e+01 * *Population Density* - 1.406e+00 * *Per Capita Income* + 1.183e+03 * *Percent High School Graduate*

Region 2

- *Number of Serious Crimes* = -4163.2673 + 33.6193 * *Population Density* + 0.1024 * *Per Capita Income* - 2.7616 * *Percent High School Graduate*

Region 3

- *Number of Serious Crimes* = 38862.667 + 5.537 * *Population Density* + 1.957 * *Per Capita Income* - 670.884 * *Percent High School Graduate*

Region 4

- *Number of Serious Crimes* = 129323.415 + 5.717 * *Population Density* + 4.342 * *Per Capita Income* - 2159.920 * *Percent High School Graduate*

b) Are the equations of the fitted models similar for the four regions? Discuss.
The equations are not really similar to each other. For example, region 1 has drastically different results compared to region 4 even though the same variables are used. Specifically, the coefficients aren't even close to one another. On one hand the population density variable coefficient varies by over 10. Another aspect is the fact some of the variables get subtracted in different regions and the same ones will be added in another region.

c) Obtain the √MSE and R2values for each region. Are these measures of model fit similar for the four regions? Discuss.

Region 1:
Residual standard error: 28060 on 99 degrees of freedom
Multiple R-squared: 0.8352, Adjusted R-squared: 0.8302

Region 2:
Residual standard error: 32980 on 104 degrees of freedom
Multiple R-squared: 0.5285, Adjusted R-squared: 0.5149

Region 3:
Residual standard error: 36970 on 148 degrees of freedom
Multiple R-squared: 0.09251, Adjusted R-squared: 0.07411

Region 4:
Residual standard error: 81820 on 73 degrees of freedom
Multiple R-squared: 0.08665, Adjusted R-squared: 0.04912

Three of the residual standard error values are similar to relatively close to each other while the other region (4) is way higher than the others. On the other hand, the multiple R-squared values are low for regions 3 and 4 but regions 1 and 2 have much higher values.

d) Obtain the residuals for each fitted model and plot them in side-by-side boxplots. Interpret your plots and state your findings.

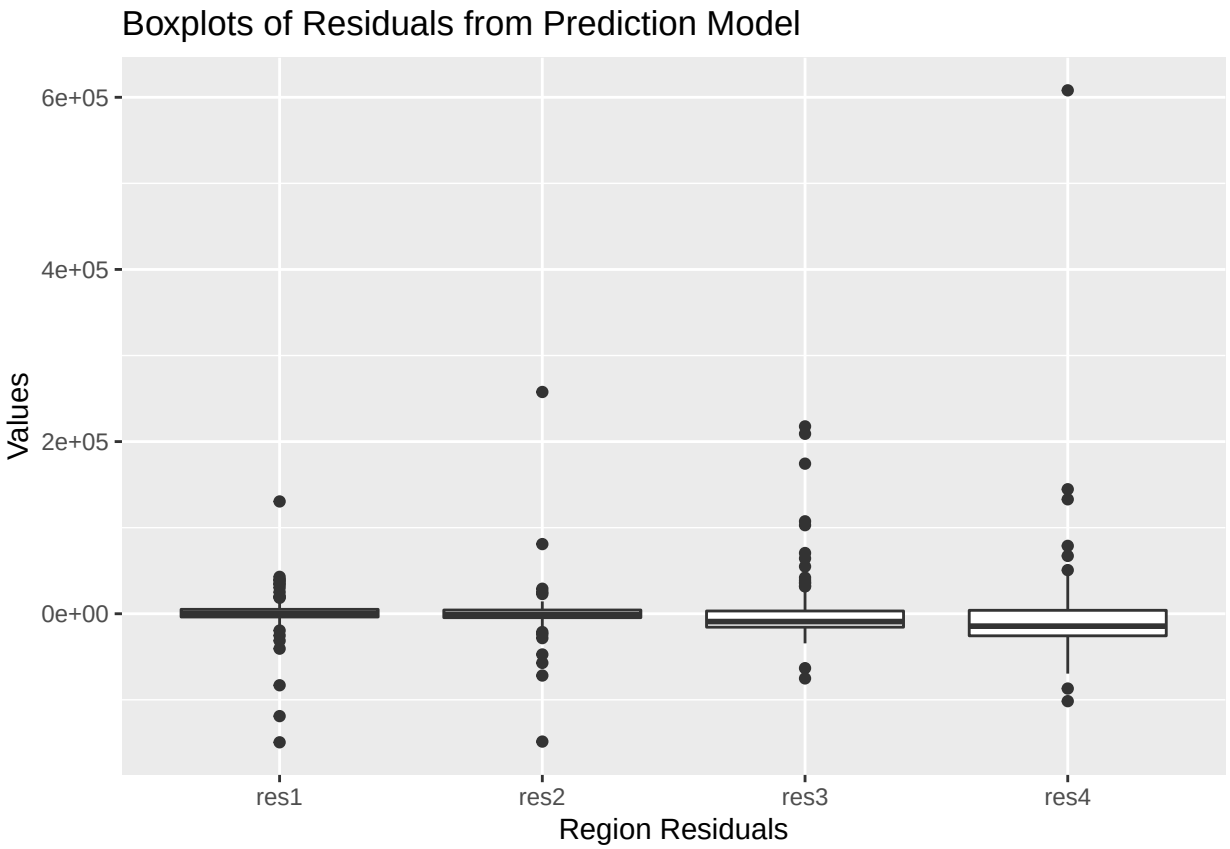
```
cdi.data <- read.csv(my.file1, header = TRUE, sep = "", stringsAsFactors = FALSE)
cdi.data <- mutate(.data = cdi.data, PopDen = TotPop / LandArea)
grp_by_cdi_reg <- nest_by(.data = cdi.data, Region)
grp_by_cdi_reg <- mutate(.data = grp_by_cdi_reg, mod = list(summary(lm(
    nCrimes ~ PopDen + PerCapInc + PctHSGrad, data = data
  ))))

res1 <- grp_by_cdi_reg[[3]][[1]]
res1 <- res1$residuals
res1 <- append(res1, rep(NA, 200 - length(res1)))
res2 <- grp_by_cdi_reg[[3]][[2]]
res2 <- res2$residuals
res2 <- append(res2, rep(NA, 200 - length(res2)))
res3 <- grp_by_cdi_reg[[3]][[3]]
res3 <- res3$residuals
res3 <- append(res3, rep(NA, 200 - length(res3)))
res4 <- grp_by_cdi_reg[[3]][[4]]
res4 <- res4$residuals
res4 <- append(res4, rep(NA, 200 - length(res4)))

res_all <- data.frame(res1, res2, res3, res4)
res_all <- pivot_longer(data = res_all, res1:res4)

ggplot(data = res_all) +
  geom_boxplot(mapping = aes(x = name, y = value)) +
  xlab("Region Residuals") + ylab("Values") +
  ggtitle("Boxplots of Residuals from Prediction Model")
```

```
## Warning: Removed 360 rows containing non-finite values (stat_boxplot).
```



The average of the residual boxplots seem to be around 0 with plenty of outliers that vary greatly. In Region 4, the residuals vary quite a bit such as one of the values approaches 6e+05. From the residuals in the boxplots, the linear models based on the three variables to predict the number of serious crimes seems to do an adequate representation of the data but there could be a better fit overall.