Module06HW

Diana McSpadden

10/1/2020

Stat 6021: Homework Set 6

Name: H. Diana McSpadden

UID: hdm5s

10.10.2020

Attended Study Group With

- Caprill Wright
- Chelsea Alvarado
- Jing Huang
- Abby Bernhardt
- Brian Nam
- Loren Bushkar
- Katie Barbre

Question 1: You will use the birthwt data set from the MASS package for this question.

The data were collected at Baystate Medical Center, Springfield, Mass during 1986. The data contain information regarding weights of newborn babies as well as potential predictors.

For this question, we will focus on using two predictors: age, the mother's age in years, and race, the mother's race which is coded as 1 for white, 2 for black, and 3 for other.

y = weight of newborn baby in grams (bwt)

x1 = mother's age in years (age)

x2 = categorical variable (race)

1 = white

2 = black

3 = other

(Question 1a) Produce a scatterplot of bwt against age.

Be sure to have separate plots and overlay the regression lines for each of the three racial categories.

Based on this plot, explain why there is an interaction effect between the age of the mother and the race of the mother.

```
# attach the data
library(MASS)

attach(birthwt)
head(birthwt,5)
```

```
##    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
```

```
race<-factor(race) # this asks R to see race as categorical
is.factor(race)
```

```
## [1] TRUE
```

```
contrasts(race)
```

```
##   2 3
## 1 0 0
## 2 1 0
## 3 0 1
```

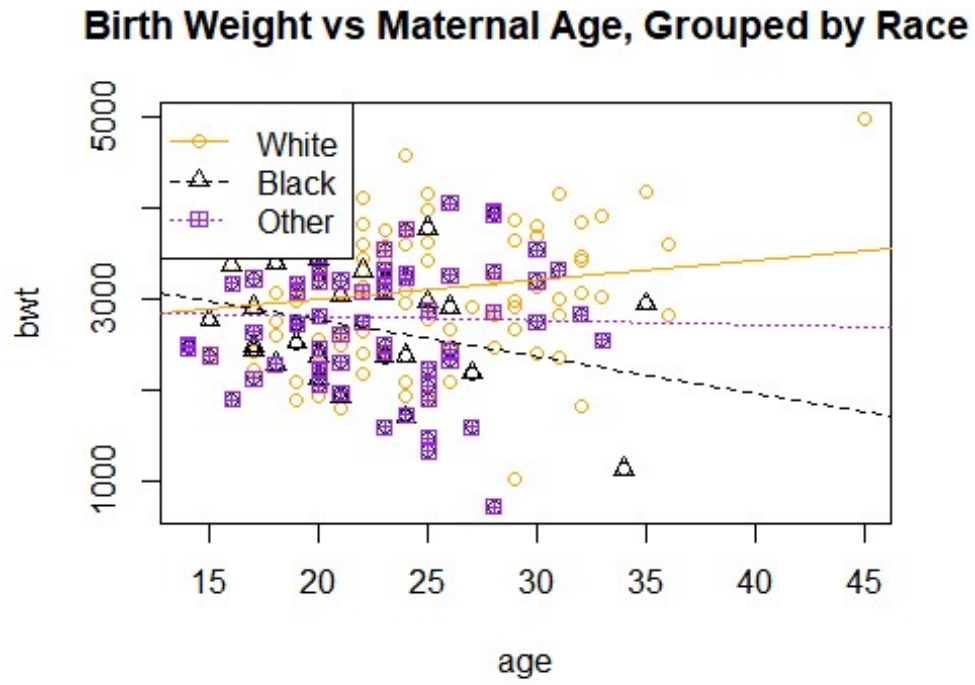
```
# need to label the classes for race
levels(race) <- c("white", "black", "other")  # Dr. Woo HIGHLY recommends
making names for classes in a category
contrasts(race)
```

```
##       black other
## white     0     0
## black     1     0
## other     0     1
```

white is our reference class so our coefficients will be able to interpret the effect on birthweigh of of black and other in comparison to white.

```
# subset the data set by race
# need to use the original data values, not labels
dataW <- subset(birthwt,race=="1")
dataB <- subset(birthwt,race=="2")
```

```r
dataO <- subset(birthwt,race=="3")

##fit 3 separate regressions, one for each region
modW <- lm(bwt~age,data=dataW)
modB <- lm(bwt~age,data=dataB)
modO <- lm(bwt~age,data=dataO)

##create a scatterplot with different colors and symbols for each race
plot(age,bwt,main="Birth Weight vs Maternal Age, Grouped by Race")

points(dataW$age,dataW$bwt, pch=1, col="orange")
points(dataB$age,dataB$bwt, pch=2, col="black")
points(dataO$age,dataO$bwt, pch=12, col="purple")

abline(modW,lty=1, col="orange")
abline(modB,lty=2, col="black")
abline(modO,lty=3, col="purple")

legend("topleft", c("White","Black","Other"), lty=c(1,2,3), pch=c(1,2,12),
col=c("orange","black","purple"))
```



Interaction Effect

There is an obvious interaction effect because as the age of the mother increases the birth weight of babies born to white mothers increases, and the birth weight of babies born to black mothers decreases, and the birth weight of babies born to mothers of other races

stays about the same. These differing relationship are shown by different slopes of bwt~age grouped by race.

(Question 1b) Fit a regression equation with interaction between the two predictors.

How does this regression equation relate the age of the mother and the weight of the baby at birth for each of the three racial categories?

```
modelInteractions <- lm(bwt~age*race)
summary(modelInteractions)
```

```
##
## Call:
## lm(formula = bwt ~ age * race)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2182.35  -474.23    13.48   523.86  1496.51
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2583.54     321.52   8.035 1.11e-13 ***
## age              21.37      12.89   1.658   0.0991 .
## raceblack      1022.79     694.21   1.473   0.1424
## raceother       326.05     545.30   0.598   0.5506
## age:raceblack   -62.54      30.67  -2.039   0.0429 *
## age:raceother   -26.03      23.20  -1.122   0.2633
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 710.7 on 183 degrees of freedom
## Multiple R-squared:  0.07541,    Adjusted R-squared:  0.05015
## F-statistic: 2.985 on 5 and 183 DF,  p-value: 0.01291
```

E(Y|x) = 2583.54 + (21.37 * age) + (1022.79 IF race == black) + (326.05 IF race == other) - (62.54 * age IF race == black) - (26.03 * age IF race == other)

or if:

x1 = age

I1 == 1 IF race is black, otherwise 0

I2 == 1 IF race is other, otherwise 0

E(Y|x) = 2583.54 + (21.37 * x1) + (1022.79 * I1) + (326.05 * I2) - (62.54 * x1 * I1) - (26.03 * x1 * I2)

```
interceptBlack <- 2583.54 + 1022.79
interceptBlack
```

```
## [1] 3606.33

interceptOther <- 2583.54 + 326.05
interceptOther

## [1] 2909.59

slopeBlack <- 21.37 - 62.54
slopeBlack

## [1] -41.17

slopeOther <- 21.37 - 26.03
slopeOther

## [1] -4.66
```

This equation states that:

For white mothers the estimated birth weight is calculated by: 2583.54 + (21.37 * (age)), also stated as: for white mothers, each unit increase of 1 year in maternal age predicts a 21.37g increase in their baby's birth weight.

For black mothers the estimated birth weight is calculated by 3606.33 - ((41.17) * (age)), also stated as: for black mothers, each unit increase of 1 year in maternal age predicts a 41.17 decrease in their baby's birth weight.

For black mothers the estimated birth weight is calculated by 2909.5 - ((4.66) * (age)), also states as for mothers of other race, each unit increase of 1 year in maternal age predicts a 4.66 decrease in their baby's birth weight.

These equations are consistent with the scatter plot: positive slope for white mothers, negative slope for black mothers, and almost stable (but slightly negative) for mothers of other races.

Question 2 (No R required) This question is based on data about teacher salaries …

from the 50 states plus DC (so n = 51) in the mid 1980s.

The variables are: * PAY, y: average annual public school teacher salary, in dollars. * SPEND, x1: Spending on public schools per student, in dollars. * AREA: Region (North, South, West).

Table 1 below provides some summary statistics of the data:

Region n Mean PAY Mean SPEND

North 21 $24424 $3901

South 17 $22894 $3274

West 13 $26159 $3919

Table 1: Summary Statistics of Teacher Pay

(Question 2a) Based only on Table 1, briefly comment on the relationship between geographic area and mean teacher pay.

```
averageAllRegions <- (26159 + 22894 + 22894) / 3
averageAllRegions
```

```
## [1] 23982.33
```

Answer

There may be a relationship between geographic region and teacher pay. In the Northern region mean pay is above average, mean teacher pay in the Western region much more above average, and mean Southern region pay is below average.

(Question 2b) Based only on Table 1, briefy comment on the relationship ...

between mean public school expenditure (per student) and mean teacher pay.

```
averageSpend <- (3901 + 3274 + 3919) / 3
averageSpend
```

```
## [1] 3698
```

Answer

The impact of student expenditure is unclear. Student expenditures in the Northern and Western regions are similar, but mean teacher pay in those regions is different. However, both mean student expenditure and mean teacher pay are less than other regions in the Southern region.

(Question 2c) Briefy explain why using a multiple linear regression model ...

with teacher pay as the response variable with geographic area and public school expenditure (per student) can give further insight into the relationship(s) between these variables.

Answer

With the 51 observation data set using a MLR model with teacher pay estimated by geographic area and per student expenditure we would be able to see how teacher pay changed per region as student expenditure changes within the region. With SLR we can only investigate whether student expenditure affects teacher pay. With MLR we can explore whether there is interaction between region, student expenditure on teacher pay. We can investigate if the degree of change (i.e. slope) is changes by geographic region, in

other words, is there better prediction of teacher pay by student expenditure when considering geographic regions?

Question 3 (No R required) This question is a continuation of question 2...

A regression with interaction was fitted, i.e.,

E(y) = B0 + B1x1 + B2I2 + B3I3 + (B4x1 * I2) + (B5x1 * I3);

where I2 and I3 are the dummy codes for AREA.

I2 = 1 if AREA = South, 0 otherwise,

and

I3 = 1 if AREA = West, 0 otherwise.

The following output from R for the extra sums of squares is shown ...

Work Calculating Missing MSR for SPEND:AREA

```
MSRSpendArea <- 9720281 * 2
print(paste("Mean Squared Regression Spend_Area: ", MSRSpendArea))

## [1] "Mean Squared Regression Spend_Area:  19440562"
```

Analysis of Variance Table

Response: PAY

```
          Df    Sum Sq        Mean Sq         F value     Pr(>F)
```

SPEND 1 608555015 608555015 117.7856 3.764e-14 ***

AREA 2 22606468 11303234 2.1877 0.1240

SPEND:AREA 2 9720281 **19440562** __ __

RESIDUALS 45 232498501 5166633

(Question 3a) Carry out a hypothesis test to see if the interaction terms are significant.

H0: the coefficients B4 and B5 = 0

Ha: at least one of the coefficients B4 or B5 != 0

```
# knowns
dfReducedModel <- 47
```

```
dfFullModel <- 45 #in the anova table, AND also 51 - the six coefficients in
the E(y)
SSEReduced <- 232498501 + 9720281
SSEFull <- 232498501

# calculate that F statistic using SSR's and SSE
# page 272 I think gives the F0 calculation
F0 <- ((SSEReduced - SSEFull) / (dfReducedModel - dfFullModel)) / (SSEFull /
dfFullModel)
print(paste("F0 is: ", F0))
```

```
## [1] "F0 is:  0.940678419685811"
```

```
# create the critical value
# page 272 gives the test statistic
criticalFValue <- qf(0.95,(dfReducedModel - dfFullModel),dfFullModel)
print(paste("The critical value is: ", criticalFValue))
```

```
## [1] "The critical value is:  3.20431729211419"
```

```
pvalue <- 1-pf(F0,(dfReducedModel - dfFullModel),dfFullModel) # remember to
use 1 - pf
print(paste("The p value for the F statistic is: ", pvalue))
```

```
## [1] "The p value for the F statistic is:  0.397903425465255"
```

Answer

Based on these finding I fail to reject the null hypothesis. I determine the interaction terms are not significant because their coefficients equal 0.

(Question 3b) Regardless of your answer from part 3a, …

suppose the interaction terms are dropped. The following is output from the model without interaction.

…

What is the reference class for this model?

Answer 3b

The reference class is **North**.

(Question 3c) What is the estimate of B2? Give an interpretation of this value.

Answer 3c

```
B2hat <- 529.40
```

Answer

The estimate of B2 is **529.40**.

Interpretation

The difference in mean teacher salary between the South and North regions is $529.40.

(Question 3d) Using the Bonferroni procedure, compute the 95% family confidence intervals for

the difference in mean response for PAY between teachers in the:

i. North region and the South region;

```
# multiplier for confidence interval is t((1-.05)/(2*g),n-p)
# g == 3 == number of pairwise comparisons we are making
# n = 51
# p = 4 == parameters Intercept, Spend, AreaSouth, AreaWest
deltaMultiplier <- qt((1 - (.05/6)), (51 - 4))
print(paste("The 95% multiplier is: ", deltaMultiplier))
```

```
## [1] "The 95% multiplier is:  2.48269449656385"
```

```
seB2hat <- sqrt(588126.71689) # sqrt(# from the vcov matrix), BUT also the
same value in the Est. Std Error table for the coef.

# confidence interval = Bi-hat +- deltaMultiplier * se(Bi-hat)
B2CIlow <- B2hat - (deltaMultiplier * seB2hat)
B2CIhigh <- B2hat + (deltaMultiplier * seB2hat)

print(paste("The 95% CI for the mean difference in the mean teacher salary
between the North region and the South region is: ", B2CIlow, " to ",
B2CIhigh))
```

```
## [1] "The 95% CI for the mean difference in the mean teacher salary between
the North region and the South region is:  -1374.56401447498  to
2433.36401447498"
```

ii. North region and the West region;

```
B3hat <- 1674.00

seB3hat <- sqrt(641873.8) # sqrt(# from the vcov matrix), BUT also the same
value in the Est. Std Error table for the coef.

B3CIlow <- B3hat - (deltaMultiplier * seB3hat)
B3CIhigh <- B3hat + (deltaMultiplier * seB3hat)

print(paste("The 95% CI for the mean difference in the mean teacher salary
between the North region and the West region is: ", B3CIlow, " to ",
B3CIhigh))
```

```
## [1] "The 95% CI for the mean difference in the mean teacher salary between
the North region and the West region is:  -315.06101776606  to
3663.06101776606"
```

iii. South region and the West region.

```
B2B3hat <- B2hat - B3hat
#print(paste("B2 B3 hat:", B2B3hat))
#print(paste("t value: ", deltaMultiplier))


# above calculation may be wrong, also trying:
# Var(B2-hat) + Var(B3-hat) - 2Cov(B2,B3)
seB2B3hatv2 <- sqrt(588126.72 + 641873.8 - (2 * 244238.0)) # use the vcov
table

B2B3CIlowv2 <- B2B3hat - (deltaMultiplier * seB2B3hatv2)
B2B3CIhighv2 <- B2B3hat + (deltaMultiplier * seB2B3hatv2)

print(paste("The 95% CI for the mean difference in the mean teacher salary
between the South region and the West region is: ", B2B3CIlowv2, " to ",
B2B3CIhighv2))
```

```
## [1] "The 95% CI for the mean difference in the mean teacher salary between
the South region and the West region is:  -3282.49336648301  to
993.293366483012"
```

(Question 3e) What do your intervals from part 3d indicate about the effect of geographic region on mean annual salary for teachers?

When spending per student is held constant, the difference in mean pay for teachers in the North vs. South or West, the confidence intervals are large and include 0. 0 within the CI range indicates that region does not have an effect on the mean salary of teachers when the model includes student expenditure and region. I conclude that region does not need to be included in the model.