uwa6xv_M08_HW

Alanna Hazlett

2024-03-27

Problem 1

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. The response variable is bwt, the weight of the baby at birth in grams.

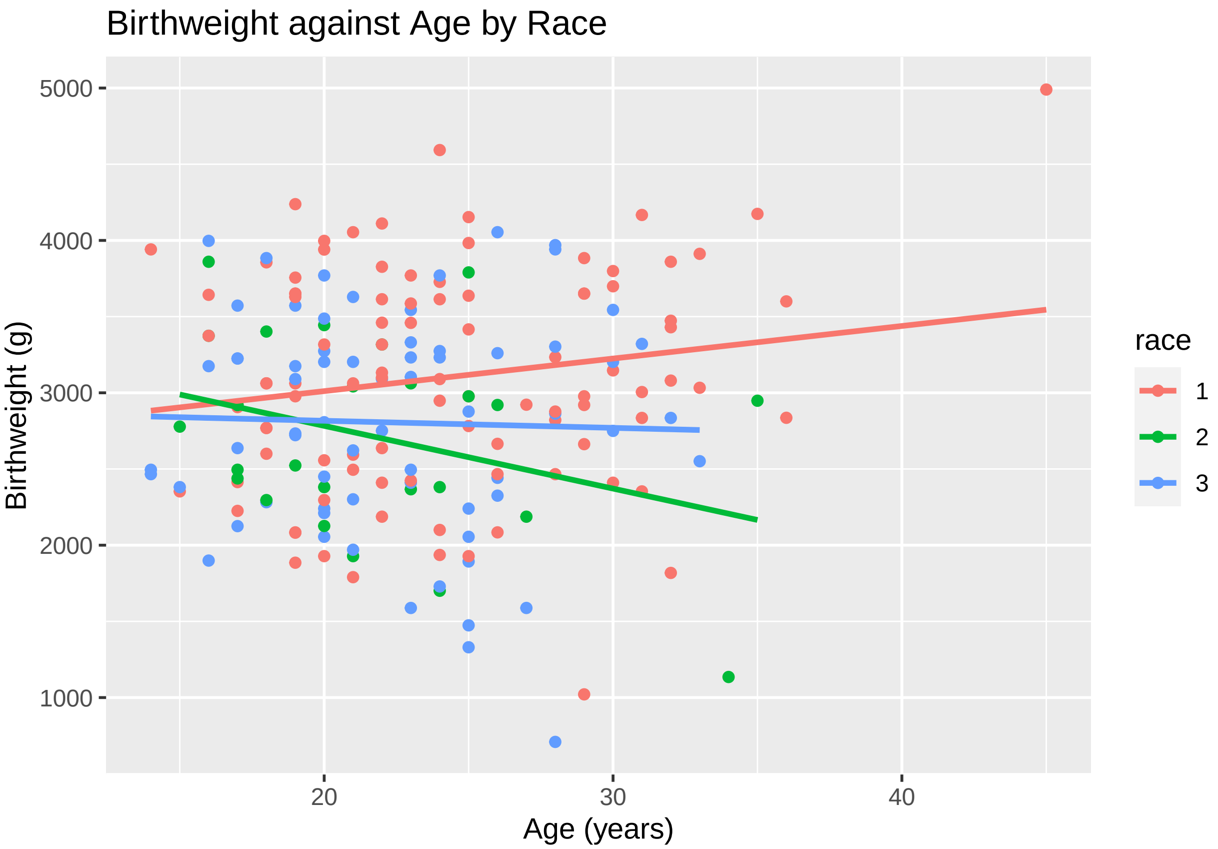
```
Data<-birthwt
```

(a)

Produce a scatterplot of bwt against age. Be sure to have separate colors 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.

```
# Need to convert race to be a factor
Data$race<-factor(Data$race)
ggplot(Data, aes(x=age,y=bwt,color=race))+
  geom_point()+
  geom_smooth(method=lm, se=FALSE)+
  labs(x="Age (years)", y="Birthweight (g)",title="Birthweight against Age by Race")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



We can see that there must be interactive effect(s), because the slopes of all 3 lines are different and the intercepts look like they will all be different as well. This indicates that the effect of a predictor on a response variable depends on the value of the other predictor.

(b)

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?

```
# Checking dummy coding
contrasts(Data$race)
```

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

```
result.inter<-lm(bwt~age*race,data=Data)
summary(result.inter)
```

```
##
## Call:
## lm(formula = bwt ~ age * race, data = Data)
##
## 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 .
## race2         1022.79     694.21   1.473   0.1424
## race3          326.05     545.30   0.598   0.5506
## age:race2      -62.54      30.67  -2.039   0.0429 *
## age:race3      -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
```

The model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 I_1 + \beta_3 I_2 + \beta_4 x_1 I_1 + \beta_5 x_1 I_2 + \in$$

The estimated regression equation is:

$$\hat{y} = 2583.54 + 21.37x_1 + 1022.79I_1 + 326.05I_2 + -62.54x_1 I_1 + -26.03x_1 I_2$$

For Race 1:

$$\hat{y} = 2583.54 + 21.37x_1 + 1022.79(0) + 326.05(0) + -62.54x_1(0) + -26.03x_1(0)$$

$$\hat{y} = 2583.54 + 21.37x_1$$

When we increase age by 1 year (x1) the mean response of birthweight increases by 21.37 grams.

For Race 2:

$$\hat{y} = 2583.54 + 21.37x_1 + 1022.79(1) + 326.05(0) + -62.54x_1(1) + -26.03x_1(0)$$

$$\hat{y} = 3606.33 + -41.17x_1$$

When we increase age by 1 year (x1) the mean response of birthweight decreases by 41.17 grams.

For Race 3:

$$\hat{y} = 2583.54 + 21.37x_1 + 1022.79(0) + 326.05(1) + -62.54x_1(0) + -26.03x_1(1)$$

$$\hat{y} = 2909.59 + -4.66x_1$$

When we increase age by 1 year (x1) the mean response of birthweight decreases by 4.66 grams.

For race one birthweight is increasing as the mother's age increases, then race two displays a slight downward trend of birthweight decreasing as mother's age increases, and race three has a larger downward trend of birthweight decreasing as mother's age increases.

Problem 2

(a)

Briefly comment on the relationship between geographic area and mean teacher pay.
West has the highest mean teacher pay. South has the lowest mean teacher pay. North is in between West and South.

(b)

Briefly comment on the relationship between mean public school expenditure (per student) and mean teacher pay.
We can see that as mean teacher pay increases so does mean spend, so West has the highest mean teacher pay and highest mean spend per student. South has the lowest mean teacher pay and lowest mean spend per

student. However, these two variables do not increase at similar rates. In mean teacher pay there is ~2,000 difference in each region. In mean spend South is significantly lower than the other two regions, where there is an increase of ~700 from South to North and only 18 difference from North to West.

(c)

Briefly 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.

We would be able to assess how each predictor relates to the response variable, when controlling the other; we could determine that there are interactive effects between our predictors that affect the response. We may want to know how teacher pay is affected by spending per student based on each region. It is reasonable to suspect that there are interactive effects, because contextually we know that cost of living varies among these regions, and we would suspect pay and cost (spending) would mirror this for each region.

Problem 3

(a)

Carry out a hypothesis test to see if the interaction terms are significant.

$$H_0 : \beta_4 = \beta_5 = 0$$

$$H_a : at\ least\ one\ coefficient\ \neq 0$$

$$F_0 = \frac{(SS_R(F) - SS_R(R))\ /\ r}{(SS_{res}(F))\ /\ (n-p)} = \frac{(SS_{res}(R) - SS_{res}(F))\ /\ r}{MS_{res}}$$

$$F_0 = \frac{9720281\ /\ 2}{5166633} = 0.9407$$

The critical value and p-value are:

```r
# Partial F test is a one sided test
qf((1-0.05),2,(51-6))
```

```
## [1] 3.204317
```

```r
pf(0.9407,2,(51-6))
```

```
## [1] 0.6021048
```

Our test statistic is less than our critical value and our p-value is greater than our significance level. We fail to reject the null hypothesis. We drop the interactive terms and go with the reduced, additive only model.

(b)

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?

The reference class is AREANorth.

(c)

What is the estimate of beta2? Give an interpretation of this value.

Beta2 is 5.294x10ˆ2 = 529.4. This value indicates the difference in the mean response, mean teacher pay, for AREASouth from AREANorth, when controlling for the other predictor, mean spending per student. On average the mean teacher pay is 529.40 dollars higher for the South region compared to the North, when mean spending per student is controlled.

(d)

Using the Bonferroni procedure, compute the 95% family confidence intervals for the difference in mean response for PAY between teachers in the pairwise comparisons

$$CI = \hat{\beta}_j \pm (t_{1-\frac{\alpha}{2g},(n-p)} * se(\hat{\beta}_j))$$

The t multiplier is: *Venkat helped me realize I forgot to update my p to 6 in this section.*

```
qt(1-(0.05/(2*3)),(51-6))
```

```
## [1] 2.486781
```

i. North region and the South region;

$$CI = 529.4 \pm (2.486781 * 766.9) = (-1374.71, 2436.51)$$

ii. North region and the West region;

$$CI = 1674 \pm (2.486781 * 801.2) = (-318.41, 3666.41)$$

iii. South region and the West region.

$$CI = (\hat{\beta}_2 - \hat{\beta}_3) \pm (t_{1-\frac{\alpha}{2g},(n-p)} * \sqrt{Var(\hat{\beta}_2 - \hat{\beta}_3)})$$

$$\sqrt{Var(\hat{\beta}_2 - \hat{\beta}_3)} = \sqrt{Var(\hat{\beta}_2) + Var(\hat{\beta}_3) - (2 * Cov(\hat{\beta}_2, \hat{\beta}_3))}$$

$$CI = (529.4 - 1674) \pm (2.486781 * \sqrt{588126.71689 + 641873.8 - (2 * 244238.03)})$$

$$CI = -1144.6 \pm (2.486781 * \sqrt{741524.4577}) = (-3286.01, 996.81)$$

(e)
What do your intervals from part 3d indicate about the effect of geographic region on mean annual salary for teachers (while controlling for expenditure)?
All 3 confidence intervals include 0, so there is not a significant difference in the mean teacher pay between all pairs of regions, when controlling for the mean spend per student.