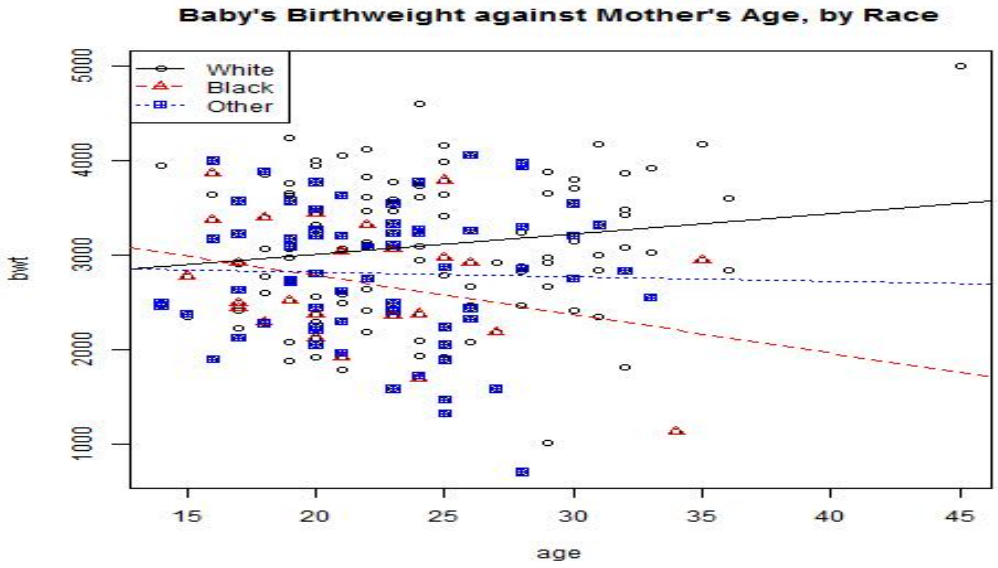
Stat 6021: Homework Set 8 Solutions

1. (a) The scatterplot is shown below. There is an interaction between the mother's age and race; the effect of the mother's age on the baby's weight differs between the racial groups. For example, for white mothers, their age is positively correlated with the baby's weight, whereas for black mothers, their age is negatively correlated with the baby's weight. For mothers who are neither black nor white, their age seems to have little correlation with the baby's weight. If there is no interaction, the slopes will be the same for all the racial groups.



(b) **Note: The dataset originally coded the racial groups as 1 for white, 2 for black, and 3 for others. When converting this variable to categorical, R chose white as the reference class, since its numeric code was the smallest, by default.** The regression output is shown below.

```
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
```

The regression equation is

$$\begin{aligned}\hat{y} &= 2583.54 + 21.37age + 1022.79I_1 + 326.05I_2 \\ &- 62.54ageI_1 - 26.03ageI_2\end{aligned}$$

where $I_1 = 1$ if mother is black, 0 otherwise; and $I_2 = 1$ if mother is neither black nor white, 0 otherwise. Breaking down the regression equation for each racial group, we have the following:

- For white mothers, the regression equation becomes

$$\begin{aligned}\hat{y} &= 2583.54 + 21.37age + 1022.79 \times 0 + 326.05 \times 0 \\ &- 62.54age \times 0 - 26.03age \times 0 \\ &= 2583.54 + 21.37age\end{aligned}$$

- For black mothers, the regression equation becomes

$$\begin{aligned}\hat{y} &= 2583.54 + 21.37age + 1022.79 \times 1 + 326.05 \times 0 \\ &- 62.54age \times 1 - 26.03age \times 0 \\ &= 3606.33 - 41.17age\end{aligned}$$

- For mothers of other races, the regression equation becomes

$$\begin{aligned}\hat{y} &= 2583.54 + 21.37age + 1022.79 \times 0 + 326.05 \times 1 \\ &- 62.54age \times 0 - 26.03age \times 1 \\ &= 2909.59 - 4.66age\end{aligned}$$

For white mothers, their age is positively correlated with their baby's weight. For black mothers, their age is negatively correlated with their baby's weight. For other races, the relationship is also negative, but less strong than for black mothers.

Note: If you had chosen another racial group as the reference class, the regression equations for each racial group will still be the same.

2. (a) Teachers in the west have the highest mean pay, followed by teachers in the north, and teachers in the south have the lowest mean pay.

   (b) Based on the mean teacher pay and mean expenditure, the higher the mean expenditure, the higher the mean teacher pay.

(c) A multiple linear regression model will allow us to separate the effects of geographic region and expenditure.

A multiple linear regression model will allow us to assess if geographic region influences teacher pay, while controlling for expenditure. We will also be able to assess if increased expenditure is related to increased teacher pay, on average, while controlling for geographic regions.

A multiple linear regression model will allow us to assess how geographic region and expenditure are simultaneously associated with teacher pay.

3. (a) We have a partial F test here.

The full model is: $\mathrm{E}(y) = \beta_0 + \beta_1 x_1 + \beta_2 I_2 + \beta_3 I_3 + \beta_4 x_1 \cdot I_2 + \beta_5 x_1 \cdot I_3$.
The reduced model is: $\mathrm{E}(y) = \beta_0 + \beta_1 x_1 + \beta_2 I_2 + \beta_3 I_3$.
$H_0 : \beta_4 = \beta_5 = 0$.
H_a : At least one of β_4, β_5 is not zero.

$$\begin{aligned} F &= \frac{[SSR(F) - SSR(R)]/r}{SS_{res}(F)/n - p} \\ &= \frac{9720281/2}{5166633} \\ &= 0.9407. \end{aligned}$$

Critical value for $F_{0.95;2,45}$ is 3.204317. Fail to reject null, so we can drop the interaction terms from the model.

(b) Reference class is NE/NC region.

(c) $\hat{\beta}_2 = 529.4$. This indicates that the estimated annual pay for teachers in the south is \$529.40 higher than the pay for teachers in the northeast/north central, while controlling for spending per student.

(d) Using Bonferroni procedure, the multiplier is $B = t_{1-0.05/6,47} = 2.482694$.

i.

$$\begin{aligned} \hat{\beta}_2 \pm Bs\{\hat{\beta}_2\} &= 529.4 \pm 2.482694(766.9) \\ &= (-1374.578, 2433.378) \end{aligned}$$

ii.

$$\begin{aligned} \hat{\beta}_3 \pm Bs\{\hat{\beta}_3\} &= 1674 \pm 2.482694(801.2) \\ &= (-315.1348, 3663.1348) \end{aligned}$$

iii.

$$\begin{aligned} s^2\{\hat{\beta}_2 - \hat{\beta}_3\} &= s^2\{\hat{\beta}_2\} + s^2\{\hat{\beta}_3\} - 2s\{\hat{\beta}_2, \hat{\beta}_3\} \\ &= 588126.71689 + 641873.8 - 2(244238.02959) \\ &= 741524.5 \end{aligned}$$

$$
\begin{aligned}
(\hat{\beta}_2 - \hat{\beta}_3) \pm Bs\{\hat{\beta}_2 - \hat{\beta}_3\} &= (529.4 - 1674) \pm 2.482694\sqrt{741524.5} \\
&= (-3282.4933, 993.2933)
\end{aligned}
$$

(e) Since all the intervals contain 0, there is no significant effect of geographic region on mean annual salary for teachers, while controlling for expenditure per student.