Stat 6021: Homework Set 8

Tom Lever

10/28/22

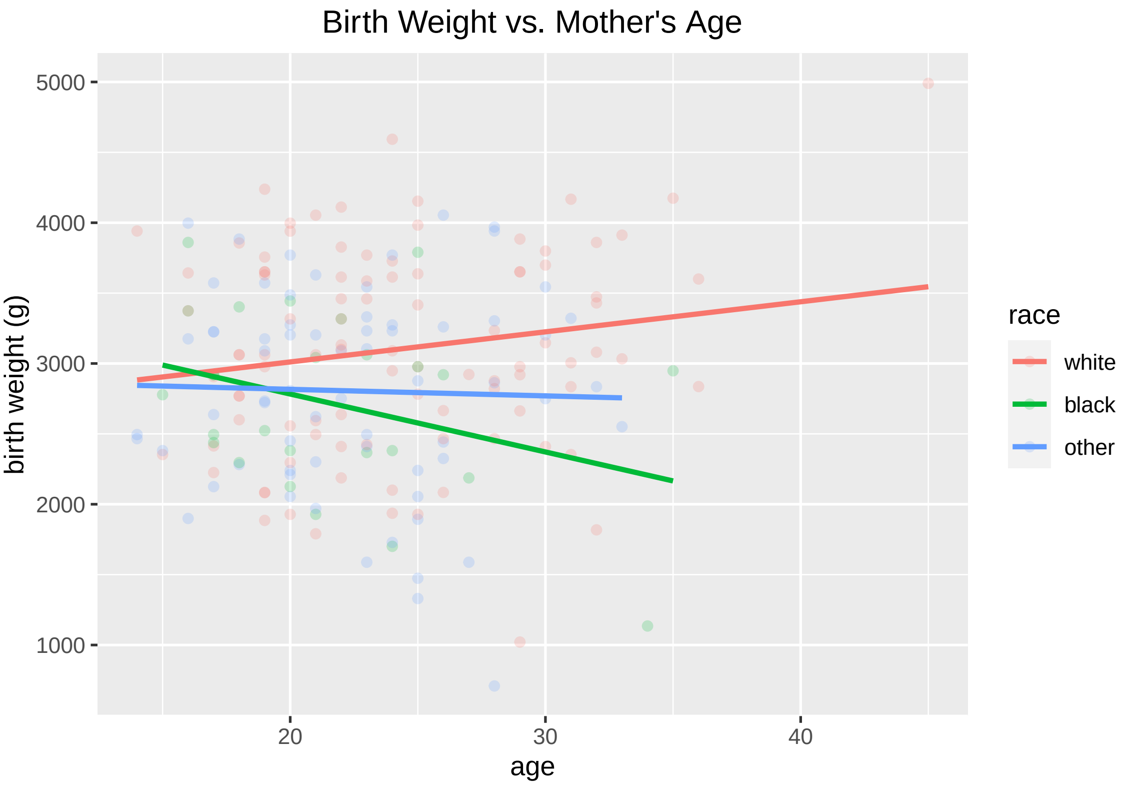
1. You will use the `birthwt` data set from the `MASS` package for this question. The data were collected at Baystate Medical Center, Springfield, MA in 1986. The data contain information regarding weights of newborn babies as well as potential predictors. For this question, we will focus on using two 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. The response variable is `bwt`, the weight of the baby at birth in grams.

   (a) Produce of scatterplot of *bwt* versus *age*.

```
library(MASS)
library(ggplot2)
birthwt$race <- factor(birthwt$race)
levels(birthwt$race) <- c("white", "black", "other")
head(birthwt, n = 3)
```

```
##    low age lwt  race smoke ptl ht ui ftv  bwt
## 85   0  19 182 black     0   0  0  1   0 2523
## 86   0  33 155 other     0   0  0  0   3 2551
## 87   0  20 105 white     1   0  0  0   1 2557
```

```
ggplot(
    birthwt,
    aes(x = age, y = bwt, color = race)
) +
    geom_point(alpha = 0.2) +
    geom_smooth(method = "lm", se = FALSE) +
    labs(
        x = "age",
        y = "birth weight (g)",
        title = "Birth Weight vs. Mother's Age"
    ) +
    theme(
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 0)
    )
```



The slope of simple linear regression equations for birth weight versus mother's age differs for different racial categories. The effect on birth weight of mother's age depends on racial category. The influence on birth weight of mother's age differs for different racial categories. Because the effect on birth weight of mother's age depends on racial category, there is an interaction effect between mother's age and racial category.

(b) Fit a regression equation with interaction between 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?

```
library(TomLeversRPackage)
generate_data_frame_for_indicator_variables(birthwt$race)
```

```
##       I_black I_other
## white       0       0
## black       1       0
## other       0       1
```

```
linear_model <- lm(bwt ~ age * race, birthwt)
summarize_linear_model(linear_model)
```

```
##
## Call:
## lm(formula = bwt ~ age * race, data = birthwt)
##
## 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) =
##     B_0 +
##     B_age * age +
##     B_raceblack * raceblack +
##     B_raceother * raceother +
##     B_age:raceblack * age:raceblack +
##     B_age:raceother * age:raceother
## E(y | x) =
##     2583.53885033192 +
##     21.3727574477423 * age +
##     1022.78795938276 * raceblack +
##     326.048531510237 * raceother +
##     -62.5379307559243 * age:raceblack +
##     -26.0316605033589 * age:raceother
## Number of observations: 189
## Estimated variance of errors: 505088.237828597
## Prediction R2: 0.00212644317471256
## Multiple R:  0.274605107027129   Adjusted R:  0.223932774155654
## Critical value t(alpha/2 = 0.05/2, DFRes = 183): 1.97301191513626
## Critical value F(alpha = 0.05, DFR = 5, DFRes = 183): 2.26347718874258
```

The below multiple linear regression model and equation relates birthweight to mother's age and racial category. When a mother is white, I_{black} and I_{other} are 0. When a mother is black, I_{black} is 1 and I_{other} is 0. When a mother has a racial category of other, I_{black} is 0 and I_{other} is 1.

$$\hat{\beta}_0 = 2583.539\ g$$

$$\hat{\beta}_{age} = 21.373\ \frac{g}{y}$$

$$\hat{\beta}_{black} = 1022.788\ g$$

$$\hat{\beta}_{other} = 326.049\ g$$

$$\hat{\beta}_{age,black} = -62.538\ \frac{g}{y}$$

$$\hat{\beta}_{age,other} = -26.032\ \frac{g}{y}$$

$$y = \beta_0 + \beta_{age}\ age + \beta_{black} I_{black} + \beta_{other} I_{other} + \beta_{age,black}\ age\ I_{black} + \beta_{age,other}\ age\ I_{other} + \epsilon$$

$$E(y|(age, white)) = \beta_0 + \beta_{age}\ age + \beta_{black} I_{black} + \beta_{other} I_{other} + \beta_{age,black}\ age\ I_{black} + \beta_{age,other}\ age\ I_{other}$$

When a mother is white, birth weight varies with mother's age at a positive reference rate of β_{age} with a positive reference bias of β_0.

$$E(y|(age, white)) = \beta_0 + \beta_{age}\ age + \beta_{black}(0) + \beta_{other}(0) + \beta_{age,black}\ age\ (0) + \beta_{age,other}\ age\ (0)$$

$$E(y|(age, white)) = \beta_0 + \beta_{age}\ age$$

When a mother is black, birth weight varies with mother's age at a decreased, negative rate of $\beta_{age} + \beta_{age,black}$ with a increased, positive bias of $\beta_0 + \beta_{black}$.

$$E(y|(age, black)) = \beta_0 + \beta_{age}\ age + \beta_{black}(1) + \beta_{other}(0) + \beta_{age,black}\ age\ (1) + \beta_{age,other}\ age\ (0)$$

$$E(y|(age, black)) = \beta_0 + \beta_{age}\ age + \beta_{black} + \beta_{age,black} age$$

$$E(y|(age, black)) = (\beta_0 + \beta_{black}) + (\beta_{age} + \beta_{age,black})\ age$$

When a mother has a racial category of other, birth weight varies with mother's age at a negative rate of $\beta_{age} + \beta_{age,other}$ between the reference rate and the rate for black mothers with a positive bias of $\beta_0 + \beta_{other}$ between the reference bias and the bias for black mothers.

$$E(y|(age, other)) = \beta_0 + \beta_{age}\ age + \beta_{black}(0) + \beta_{other}(1) + \beta_{age,black}\ age\ (0) + \beta_{age,other}\ age\ (1)$$

$$E(y|(age, other)) = \beta_0 + \beta_{age}\ age + \beta_{other} + \beta_{age,other} age$$

$$E(y|(age, other)) = (\beta_0 + \beta_{other}) + (\beta_{age} + \beta_{age,other})\ age$$

2. You may only use R as a simple calculator or to find p-values or critical values. This question is based on data about teacher salaries from the 50 states plus Washington, DC in the mid 1980's. The variables are

- $AREA$: region and an element of the set {North, South, West}
- n: number of political entities in an area
- PAY, y: average annual public school teacher salary in dollars
- $SPEND$, x_1: spending on public schools per student in dollars

Table 1 below provides some summary statistics of the data:

$AREA$	n	$\bar{y}$	$\bar{x_1}$
North	21	24,424	3,901
South	17	22,894	3,274
West	13	26,159	3,919

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

Mean teacher pay increases going from area South to North to West.

(b) Based only on Table 1, briefly comment on the relationship between mean public school expenditure per student and mean teacher pay.

As mean spending on public schools per students increases, average annual public school teacher salary increases, at an increasing rate.

(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 relationships between these variables.

A multiple linear regression model with teacher pay as the response variable with geographic area and public school expenditure per student would allow us to determine whether the effect on teacher pay of public school expenditure per student depends on geographic area, and to determine whether there is an interaction effect between public school expenditure per student and geographic area.

3. You may only use R as a simple calculator or to find p-values or critical values. This question is a continuation of question 2. A full multiple linear regression model with an interaction term between public school expenditure per student and geographic area M_{full} was fitted: The resulting multiple linear regression equation is

$$E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 I_2 + \beta_3 I_3 + \beta_4 x_1 I_2 + \beta_5 x_1 I_3$$

where I_2 and I_3 are the dummy codes for $AREA$ / the indicator variables for $South$ and $West$, respectively. If $AREA = South$, $I_2 = 1$; else $I_2 = 0$. If $AREA = West$, $I_3 = 1$; else $I_3 = 0$. The prompt for this homework includes output for an R analysis of variance for M_{full}.

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

Let $\boldsymbol{\beta}_{predictor}$ be a column vector of the coefficients of the non-reference indicator variables associated with predictor $predictor$. Let $\boldsymbol{predictor}$ be a column vector of the non-reference indicator variables associated with predictor $predictor$. The R analysis of variance considers a version of the equation for MLR model M_{full}

$$E(y|x) = \beta_0 + \beta_1 x_1 + \boldsymbol{\beta}_{area} \cdot \boldsymbol{area} + \boldsymbol{\beta}_{x_1, area}\ x_1 \cdot \boldsymbol{area}$$

The mean square for the interaction term $x_1\ area$

$$MS_{x_1, area} = \frac{SS_{x_1, area}}{DF_{x_1, area}} = \frac{9,720,281}{2} = 4,860,140.5$$

We conduct a Partial F Test of a null hypothesis that that regression coefficient for the interaction term is 0, is insignificant, and may be dropped from the multiple linear regression model.

The F statistic for the Partial F Test follows a F distribution with $DF_R = 2$ and $DF_{Res} = 45$ degrees of freedom.

$$F_0 = \frac{MS_{x_1, area}}{MS_{Res}} = \frac{4,860,140.5}{5,166,633} = 0.941$$

The critical value of the F distribution for which the probability of a random F statistic being greater is equal to a significance level $\alpha = 0.05$

$$F_c = 3.204$$

```r
F_statistic <- 0.941
significance_level <- 0.05
regression_degrees_of_freedom <- 2
residual_degrees_of_freedom <- 45
qf(
    significance_level,
    regression_degrees_of_freedom,
    residual_degrees_of_freedom,
    lower.tail = FALSE
)
```

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

Since our F statistic is less than the critical value, we have insufficient evidence to reject the null hypothesis that the regression coefficient for the interaction term is 0, is insignificant, and may be dropped from the multiple linear model.

The probability that a random F statistic is greater than our F statistic

$$p = 0.398$$

```r
pf(
    F_statistic,
    regression_degrees_of_freedom,
    residual_degrees_of_freedom,
    lower.tail = FALSE
)
```

```
## [1] 0.3977806
```

Since this probability is greater than a significance level $\alpha = 0.05$, we have insufficient evidence to reject the null hypothesis that the regression coefficient for the interaction term is 0, is insignificant, and may be dropped from the multiple linear model.

We conclude that the regression coefficient for the interaction term is 0, is insignificant, and may be dropped from the multiple linear model.

(b) Regardless of your answer from part 3a, suppose the interaction terms are dropped. The prompt for this homework includes a summary and variance-covariance matrix for the multiple linear model without interaction. What is the reference class for this model?

Consider the multiple linear model of teacher pay versus public school expenditure per student and geographic area without an interaction effect between public school expenditure per student and geographic area $M_{reduced}$ with equation

$$\hat{\beta}_0 = 11,600 \ dollars$$

$$\hat{\beta}_1 = 3.289$$

$$\hat{\beta}_2 = 529.4 \ dollars$$

$$\hat{\beta}_3 = 1,674 \ dollars$$

$$E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 I_2 + \beta_3 I_3$$

Because if $AREA = South$, $I_2 = 1$ and $I_2 = 0$ otherwise, and if $AREA = West$, $I_3 = 1$ and $I_3 = 0$ otherwise, when $AREA = North$, $I_2 = 0$ and $I_3 = 0$ and North is the reference class for $M_{reduced}$. When $Area = North$,

$$E_N = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(0)$$

$$E_N = \beta_0 + \beta_1 x_1$$

(c) What is the estimate of β_2? Give an interpretation of this value.

When $Area = South$, $I_2 = 1$, and

$$E_S = \beta_0 + \beta_1 x_1 + \beta_2(1) + \beta_3(0)$$

$$E_S = \beta_0 + \beta_1 x_1 + \beta_2$$

$$E_S = (\beta_0 + \beta_2) + \beta_1 x_1$$

$$E_S - E_N = \beta_2$$

$\hat{\beta}_2 = 529.4 \ dollars$ is the estimated difference between mean teacher pay when $AREA = South$ and the reference mean teacher pay when $AREA = North$, per $M_{reduced}$.

(d) Using the Bonferroni procedure, compute the 95-percent family confidence intervals for the difference in mean response for PAY between teachers in the

i. North region and the South region
ii. North region and the West region

iii. South region and the West region

When $Area = West$, $I_3 = 1$, and

$$E_W = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(1)$$

$$E_W = \beta_0 + \beta_1 x_1 + \beta_3$$

$$E_W = (\beta_0 + \beta_3) + \beta_1 x_1$$

$\hat{\beta}_3 = 1,674\ dollars$ is the estimated difference between mean teacher pay when $AREA = West$ and the reference mean teacher pay when $AREA = North$, per $M_{reduced}$.

$\hat{\beta}_3 - \hat{\beta}_2 = (1,674\ dollars) - (529.4\ dollars) = 1144.6\ dollars$ is the estimated difference between mean teacher pay when $AREA = West$ and mean teacher pay when $AREA = South$, per $M_{reduced}$.

Considering the variance-covariance matrix for $M_{reduced}$,

$$\widehat{Var\left(\beta_2\right)} = 588,126.717\ dollars^2$$

$$\widehat{SE\left(\beta_2\right)} = \sqrt{\widehat{Var\left(\beta_2\right)}} = \sqrt{588,126.717} = 766.894\ dollars$$

$$\widehat{Var\left(\beta_3\right)} = 641,873.8\ dollars^2$$

$$\widehat{SE\left(\beta_3\right)} = \sqrt{\widehat{Var\left(\beta_3\right)}} = \sqrt{641,873.8} = 801.170\ dollars$$

$$\widehat{Var\left(\beta_3 - \beta_2\right)} = \widehat{Var\left(\beta_3\right)} + \widehat{Var\left(\beta_2\right)} - 2\widehat{Cov\left(\beta_3, \beta_2\right)}$$

$$\widehat{Var\left(\beta_3 - \beta_2\right)} = \left(641,873.8\ dollars^2\right) + \left(588,126.717\ dollars^2\right) - 2\left(244,238.02959\ dollars^2\right)$$

$$\widehat{Var\left(\beta_3 - \beta_2\right)} = \left(641,873.8\ dollars^2\right) + \left(588,126.717\ dollars^2\right) - 2\left(244,238.02959\ dollars^2\right)$$

$$\widehat{Var\left(\beta_3 - \beta_2\right)} = 741,524.458\ dollars^2$$

$$\widehat{SE\left(\beta_3 - \beta_2\right)} = \widehat{SE\left(d_{W,S}\right)} = \sqrt{\widehat{Var\left(\beta_3 - \beta_2\right)}} = \sqrt{741,524.458} = 861.118\ dollars$$

We are making three pairwise comparisons / calculating three differences in mean response for teacher PAY; we consider a family of $g = 3$ confidence intervals.

The residual degrees of freedom $DF_{Res,full} = 45$. When the interaction term is removed from M_{full}, regression degrees of freedom decreases by 2 and residual degrees of freedom increases by 2; the residual degrees of freedom $DF_{Res,reduced} = 47$.

The critical value for each confidence interval

$$t_c = t_{\alpha/(2g)=0.05/(2*3),DF_{Res,reduced}=47} = 2.483$$

```
number_of_confidence_intervals <- 3
residual_degrees_of_freedom <- 47
qt(
    significance_level/(2*number_of_confidence_intervals),
    residual_degrees_of_freedom,
    lower.tail = FALSE
)
```

```
## [1] 2.482694
```

```
calculate_critical_value_tc(
    significance_level,
    number_of_confidence_intervals,
    residual_degrees_of_freedom
)
```

```
## [1] 2.482694
```

Let $d_{i,j}$ be the difference in mean response for PAY between teachers in region i and region j. For example, $d_{W,S} = \hat{\beta}_3 - \hat{\beta}_2 = 1,144.6\ dollars$.

The 95-percent family confidence interval for the difference in mean response for PAY between teachers in the North region and the South region

$$\left[-\hat{d}_{S,N} - t_c\ \widehat{SE\left(-d_{S,N}\right)},\ -\hat{d}_{S,N} + t_c\ \widehat{SE\left(-d_{S,N}\right)}\right]$$

$$\left[-\hat{d}_{S,N} - t_c\ \widehat{SE\left(d_{S,N}\right)},\ -\hat{d}_{S,N} + t_c\ \widehat{SE\left(d_{S,N}\right)}\right]$$

$$\left[-\hat{\beta}_2 - t_c\ \widehat{SE\left(\beta_2\right)},\ -\hat{\beta}_2 + t_c\ \widehat{SE\left(\beta_2\right)}\right]$$

$$[-(529.4\ dollars) - (2.483)(766.894\ dollars),\ -(529.4\ dollars) + (2.483)(766.894\ dollars)]$$

$$[-2,433.598\ dollars,\ 1,374.798\ dollars]$$

The 95-percent family confidence interval for the difference in mean response for PAY between teachers in the North region and the West region

$$\left[-\hat{d}_{W,N} - t_c\ \widehat{SE\left(d_{W,N}\right)},\ -\hat{d}_{W,N} + t_c\ \widehat{SE\left(d_{W,N}\right)}\right]$$

$$\left[-\hat{\beta}_3 - t_c\ \widehat{SE\left(\beta_3\right)},\ -\hat{\beta}_3 + t_c\ \widehat{SE\left(\beta_3\right)}\right]$$

$$[-(1,674\ dollars) - (2.483)(801.170\ dollars),\ -(1,674\ dollars) + (2.483)(801.170\ dollars)]$$

$$[-3,663.305\ dollars,\ 315.305\ dollars]$$

The 95-percent family confidence interval for the difference in mean response for PAY between teachers in the South region and the West region

$$\left[-\hat{d}_{W,S} - t_c\ \widehat{SE\left(d_{W,S}\right)},\ -\hat{d}_{W,S} + t_c\ \widehat{SE\left(d_{W,S}\right)}\right]$$

$$[-(1,144.6\ dollars) - (2.483)(861.118\ dollars),\ -(1,144.6\ dollars) + (2.483)(861.118\ dollars)]$$

$$[-3282.756\ dollars,\ 993.556\ dollars]$$

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

Because all three confidence intervals contain 0, we have insufficient evidence to reject a null hypothesis that the difference between any two mean annual teacher salaries for the same public school expenditure per student and different region are 0, per $M_{reduced}$. We conclude that the difference between any two mean annual teacher salaries are approximately 0 and that there is an insignificant difference between any two mean annual teacher salaries for the same public school expenditure per student and different region.