Homework8

Dima Mikhaylov

10/28/2021

1. You will use the birthwt data set from the MASS package for this question. The response variable is bwt, the weight of the baby at birth in grams.

```
library(MASS)
#data(package = 'birthwt')
head(birthwt)
```

```
##    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
## 91   0  21 124    3     0   0  0  0   0 2622
```

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.

First, check the data type of the categorical var `race`. If `integer`, it needs to be changed to levels…

```
class(birthwt$race)
```

```
## [1] "integer"
```

Use `factor()` function to change it to categorical levels:

```
birthwt$race <- factor(birthwt$race)
class(birthwt$race)
```

```
## [1] "factor"
```

Now, `race` can be used to split observations to show possible interaction effect:

```
library(tidyverse)
```

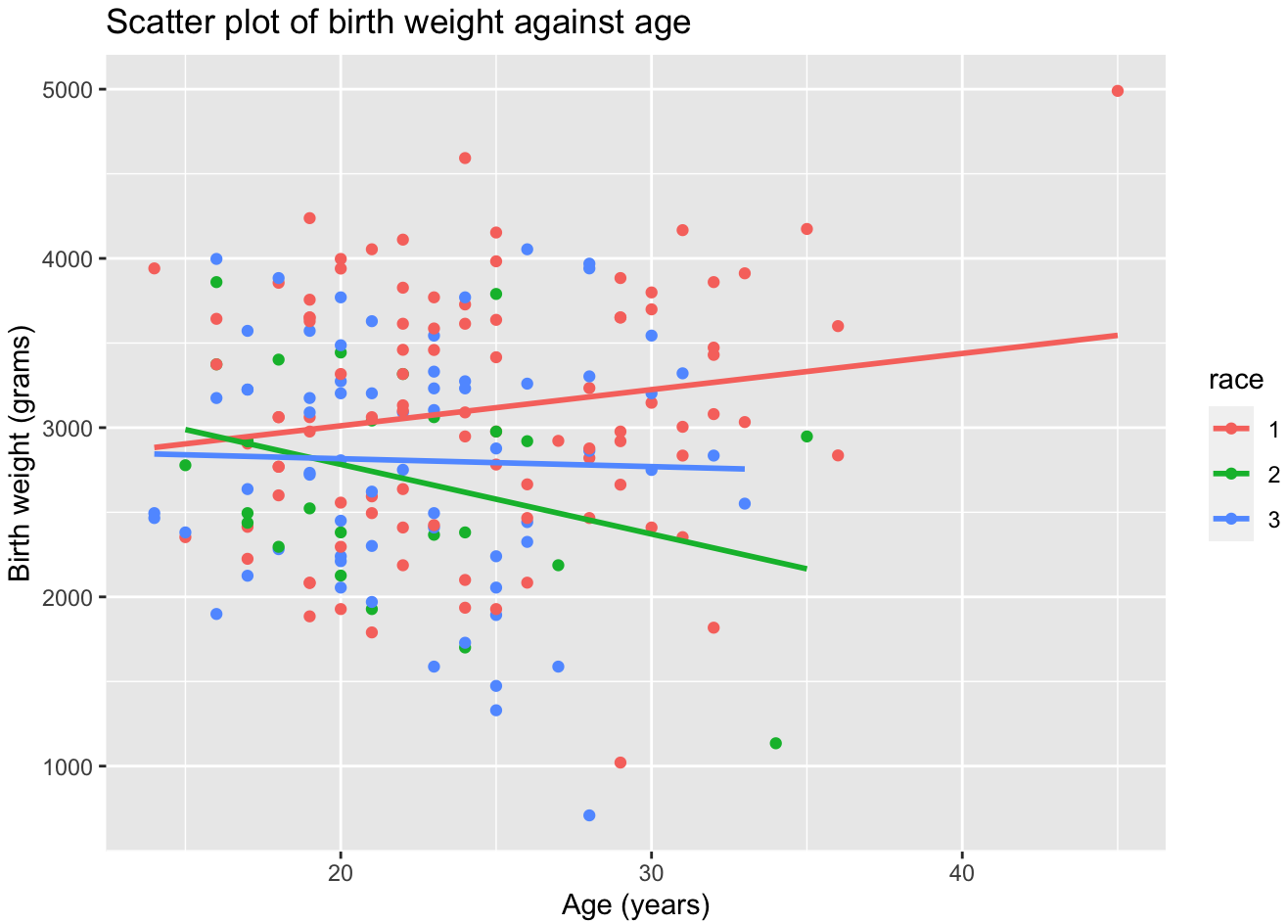
```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.4     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.3     ✓ stringr 1.4.0
## ✓ readr   2.0.1     ✓ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
```

```
ggplot(birthwt, aes(x=age, y=bwt, color=race))+
  geom_point()+
  geom_smooth(method = "lm", se=FALSE)+
  labs(x="Age (years)",
       y="Birth weight (grams)",
       title="Scatter plot of birth weight against age")
```

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



Conclusion: from the plot above, all three races appear to have very different slopes of regression lines - learly not parallel. This is an idicator of strong interaction effect. Parallel lines would suggest additivity, but different slopes suggest interaction.

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?

First, review the set up of the dummy variables coding:

```
contrasts(birthwt$race)
```

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

Next, run regression with interaction of age and weight:

```
model = lm(bwt ~ age*race, data=birthwt)
summary(model)
```

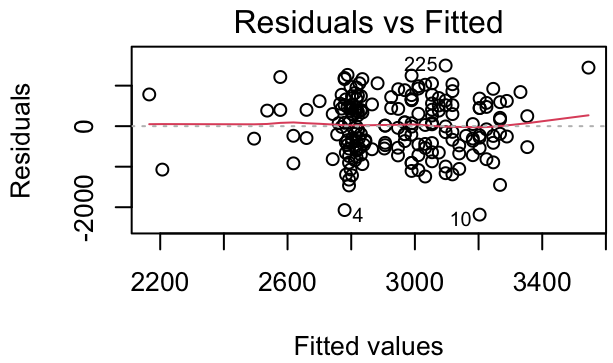
```
##
## Call:
## lm(formula = bwt ~ age * race, data = birthwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2182.35  -474.23    13.48   523.86  1496.51
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2583.54     321.52   8.035 1.11e-13 ***
## age            21.37      12.89   1.658   0.0991 .
## 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
```

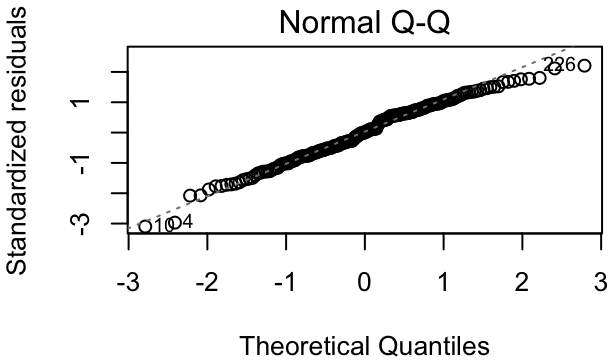
Estimated model:

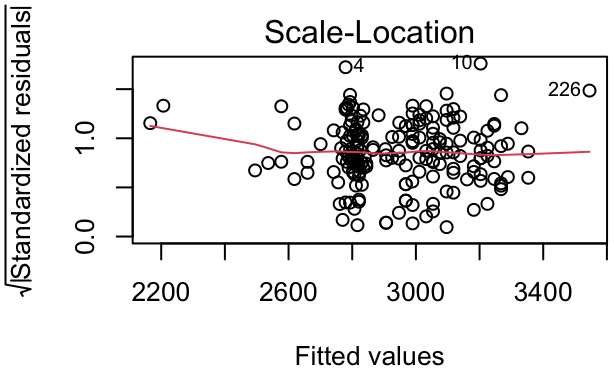
bwt= 2583.54+21.37*(Age)*+*1022.79*(Race2)+326.05*(Race3)*-*62.54*(Age*Race2)*-*26.03*(Age*Race3)

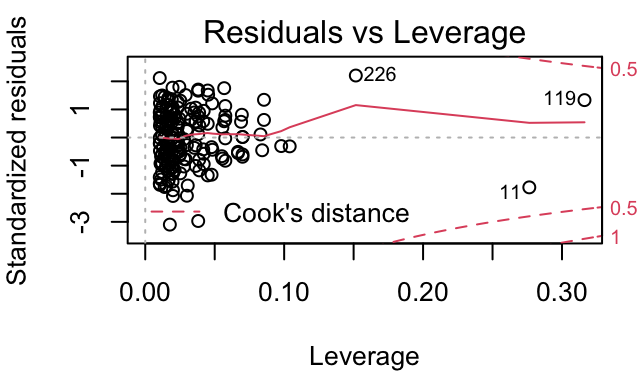
The nodel produces very low R-squared, check linear regression assumptions:

```
par(mfrow=c(2,2))
plot(model)
```



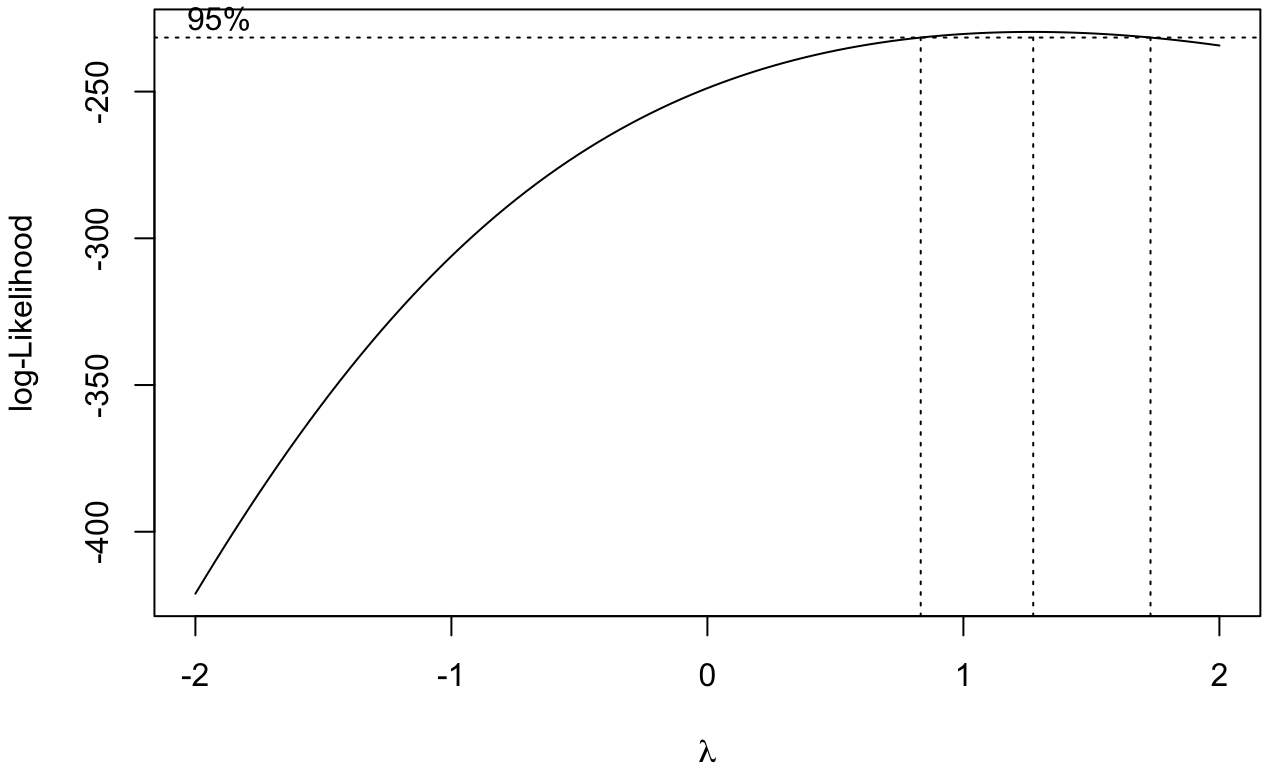






Check if 1 is in the CI of Box Cox plot to confirm that the variance is constant and hence we do not need to transform the respones variable.

```
boxcox(model)
```



Conclusion: assumptions seem to be met, but the model is likely suffering from a very noisy data.

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.

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

It seems that mean pay may be influenced by AREA, as the pay is higher in the West ($26K), somewhat lower in the North ($24K), and the lowest in the South ($23).

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

Public school expenditure (per student) also seems to be related to teacher pay, as it tends to be higher in the area where teacher pay is higher ($3,919 in the West) and lower in the area where teached pay is lower ($3,274 in the South).

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.

It is possible that public school expenditure interact with area and thus an interection term can better explain variance in the response variable teachers' pay. Multiple linear regression allows for additional high-order intercation terms

3. This question is a continuation of question 2:

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

```
SpendArea_MS = 9720281 / 2
F_stat = SpendArea_MS / 5266633
p_value = 1-pf(F_stat, 2, 45, lower.tail=F)
cat("p_value of", p_value, "is very high what suggests statistically insignificant slope of the interaction term"
)
```

```
## p_value of 0.5952134 is very high what suggests statistically insignificant slope of the interaction term
```

Conclusion: fail to reject H0, in other words the interection term is not statistically significant.

b. Regardless of your answer from part 3a, suppose the interaction terms are dropped. What is the reference class for this model?

Reference class of variable `AREA` appears to be "North". This class has value of 0 for all indicators.

c. What is the estimate of B2? Give an interpretation of this value.

From the model specification, B2 is a slope coefficient associated with I2 parameter, one the dummy codes for `AREA` categorical predictor. It is only meaningful, when I2 takes a value of 1, as opposed to 0, and this is the case of `AREA` being South.

From the summary table, B2 is estimated to be 529.4, positive increase in the response variable that is attributable to one unit of increase in SPEND in the South region (i.e. when controlling for SPEND).

d. Using the Bonferroni procedure, compute the 95% family confidence intervals for the difference in mean response for PAY between teachers in the North region and the South region, North region and the West region, South region and the West region.

First, compute the multiplier for 95% family confidence intervals:

```
g = 3 # number of pairwise comparisons
a = 0.05 # original alpha
p = 4 # params after the interaction terms were dropped
n = 51 # number of observations
df = n-p # degrees of freedom
multiplier = qt(1-(a/2*g), df)
```

So, B2 of 529.4, B2_se=766.9, is the difference between South and North. Note Estimate < Error, this will produce a very wide range.

```
B2 = 529.4
B2_se = 766.9
cat("B2 (North v South) 95% interval is [", B2-multiplier*B2_se,",", B2+multiplier*B2_se, "]")
```

```
## B2 (North v South) 95% interval is [ -592.9212 , 1651.721 ]
```

B3 of 1674, B3_se=801.2, is the difference between West and North:

```
B3 = 1674
B3_se=801.2
cat("B3 (North v West) 95% interval is [", B3-multiplier*B3_se,",", B3+multiplier*B3_se, "]")
```

```
## B3 (North v West) 95% interval is [ 501.4824 , 2846.518 ]
```

To compare the mean response for two non-reference classes, region South and West, we need to look at the difference in the coefficients associated with these two classes.

```
Dif = 1.674e+03 - 5.294e+02 # Difference, B2-B3
Dif_var = 6.418738e+05 + 588126.71689 - 2 * 2.442380e+05 #Variance of the Difference
Dif_se = sqrt(Dif_var)
cat("Diff. (South v West) 95% interval is [", Dif-multiplier*Dif_se,",", Dif+multiplier*Dif_se, "]")
```

```
## Diff. (South v West) 95% interval is [ -115.605 , 2404.805 ]
```

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

It seems that results are not conclusive because the ranges are very large, and even include 0 when comparing North vs South or South vs West. This is due to a very large standard error of South slope estimate. Noteworthy that it also has small t value and insignificant p-vlaue.