Assignment1

Yunzhi Chen 32051018

2023-04-05

Q1

```
dataA1 <- read_csv(here::here("dataA1_etc3580.csv"))
```

```
## Rows: 189 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): race, smoke
## dbl (2): bwt, age
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
dataA1 <- dataA1 %>% mutate(race = factor(race, levels = c("white", "black", "other"),
                                              labels = c("white", "black", "other")),
                            smoke = factor(smoke, levels = c("smoker", "nonsmoker"),
                                                labels = c("smoker", "nonsmoker")))
```

Q2

```
p1 <- dataA1 %>% ggplot(aes(x = age,
                            y = bwt)) +
  geom_point() +
  geom_smooth() +
  labs(title = bquote("Figure1"))

p2 <- dataA1 %>% ggplot(aes(x = race,
                            y = bwt)) +
  geom_boxplot() +
  labs(title = bquote("Figure2"))

p3 <- dataA1 %>% ggplot(aes(x = smoke,
                            y = bwt)) +
  geom_boxplot() +
  labs(title = bquote("Figure3"))

p1+p2+p3
```

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

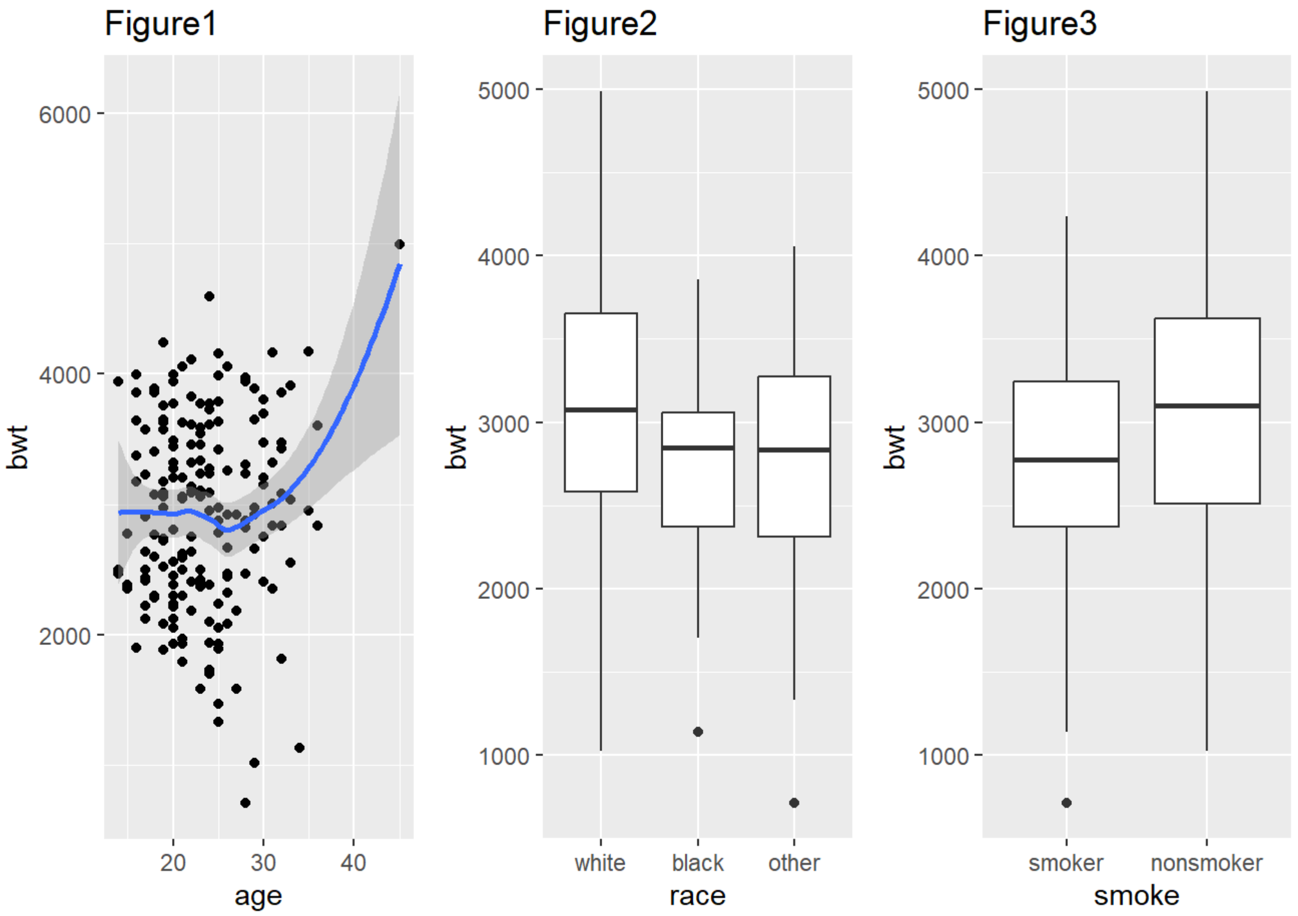


Figure1

Figure2

Figure3

Figure1 shows that overall, it seems that there is no relationship between maternal age and infant birth weight. A substantial variability in birth weight exists among each age group. As for ages 36 to 40 years and older, there are no data except one outlier, and this outlier can probably be ignored in the analysis.

Figure2 illustrates that there may be some differences in birth weight between racial groups. Specifically, white race has higher median birth weights than black and other races, while the rest groups of race have the similarly value of median birth weights of babies.

Figure3 boxplot shows that infants born to mothers who smoke tend to have lower birth weights than non-smoking mothers. In addition, the birth weights have more variation of nonsmoking mothers group.

Q3

```
lmod <- lm(bwt ~ age + race + smoke,
           data = dataA1)

summary(lmod)
```

```
##
## Call:
## lm(formula = bwt ~ age + race + smoke, data = dataA1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2321.93  -445.45    28.09   501.70  1615.09
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2859.407    247.319  11.562  < 2e-16 ***
## age               1.999      9.767   0.205 0.838071
## raceblack      -444.649    156.140  -2.848 0.004904 **
## raceother      -449.481    118.977  -3.778 0.000213 ***
## smokenonsmoker  425.556    109.951   3.870 0.000151 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 689.8 on 184 degrees of freedom
## Multiple R-squared:  0.1238, Adjusted R-squared:  0.1047
## F-statistic: 6.498 on 4 and 184 DF,  p-value: 6.512e-05
```

The estimated coefficient for age is 1.999, which suggests that, on average, birth weight increases by 1.999 grams for every additional year of age, holding all other factors constant, which indicates a very weak positive relationship.

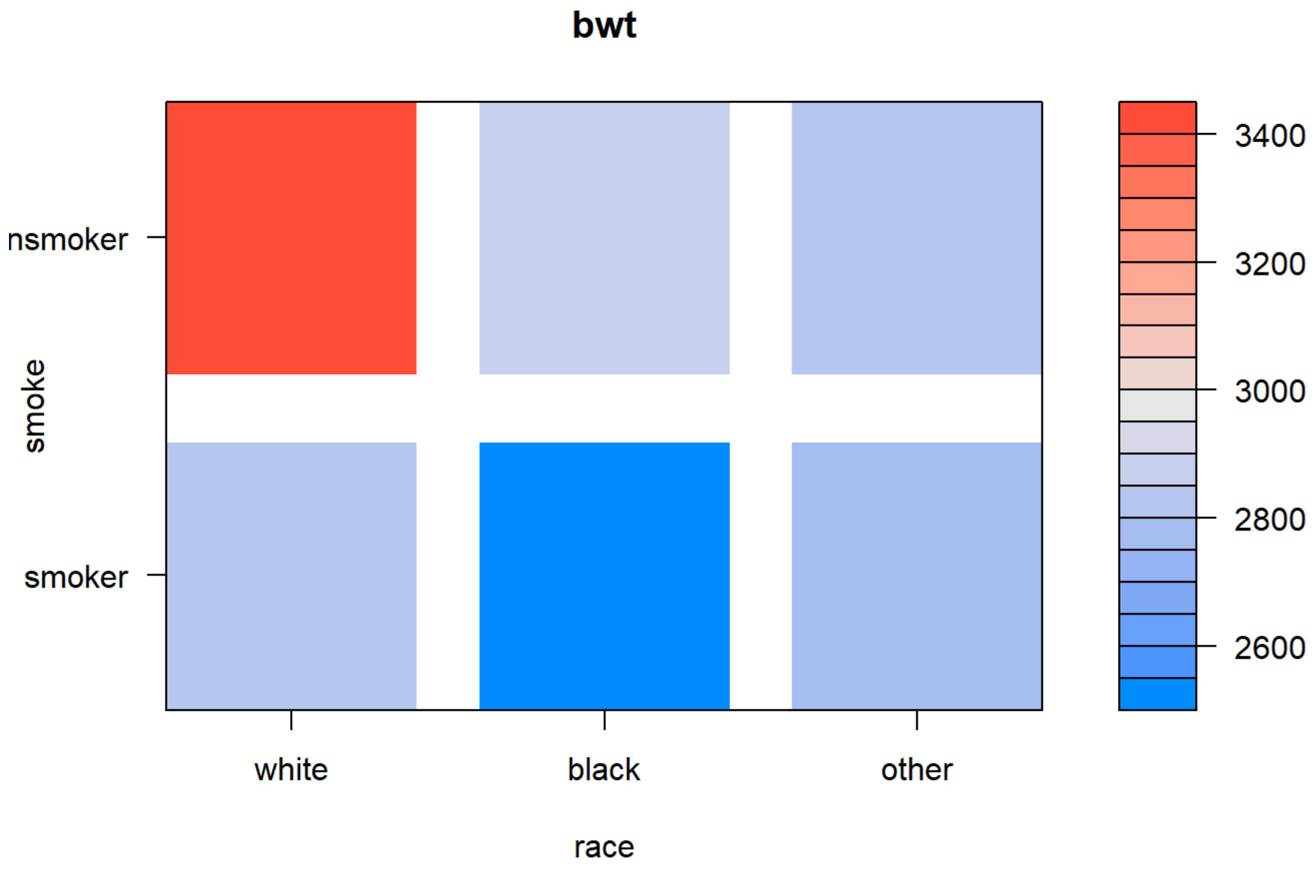
The estimated coefficient for nonsmokers is positive (425.556), indicates that, on average, infants born to mothers who do not smoke have birth weights that are 425.556 grams higher than those born to mothers who smoke, holding all other factors constant.

Q4

```
lmod2 <- lm(bwt ~ age + race + smoke + race * smoke,
           data = dataA1)
summary(lmod2)
```

```
##
## Call:
## lm(formula = bwt ~ age + race + smoke + race * smoke, data = dataA1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2404.13  -418.46    32.49   465.01  1584.35
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2855.165    247.195  11.550  < 2e-16 ***
## age                           -1.217      9.997  -0.122   0.9032
## raceblack                   -321.835    236.946  -1.358   0.1761
## raceother                    -70.616    219.469  -0.322   0.7480
## smokenonsmoker               605.255    143.948   4.205  4.1e-05 ***
## raceblack:smokenonsmoker -259.821    318.447  -0.816   0.4156
## raceother:smokenonsmoker -548.351    261.674  -2.096   0.0375 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 685.2 on 182 degrees of freedom
## Multiple R-squared:  0.1447, Adjusted R-squared:  0.1166
## F-statistic: 5.134 on 6 and 182 DF,  p-value: 6.74e-05
```

```
visreg2d(lmod2, "race", "smoke")
```



According to the color blocks, particularly pronounced for black infants born to smoking mothers, who have the lowest predicted birth weights overall, while white non-smoking mothers tend to have the highest birth weights babies.

Q5

```
anova(lmod, lmod2)
```

```
## Analysis of Variance Table
##
## Model 1: bwt ~ age + race + smoke
## Model 2: bwt ~ age + race + smoke + race * smoke
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1    184 87548473
## 2    182 85452800  2   2095673 2.2317 0.1103
```

H0: The interaction term does not significantly improve the model fit; H1: The interaction term does significantly improve the model fit.

By using the anova function to check if the interaction is significant. we can conclude from the output that the p-value is 0.1103, which is not less than 0.05, so we can not reject the null hypothesis, and the conclusion is that the interaction term is not significant.

Q6

```
fit <- lm(bwt ~ age + race + smoke,
           data = dataA1)
fit2 <- lm(bwt ~ I(age^2) + race + smoke,
           data = dataA1)

AIC(fit)
```

```
## [1] 3014.044
```

```
AIC(fit2)
```

```
## [1] 3013.759
```

As we can see from the result, fit 2 has slightly less value of AIC, which indicates the quadratic relation between age and bwt is better.

Q7

```
dataA1_prob <- dataA1 %>%
  mutate(prob = ifelse(bwt >= 2500, 0, 1))

glm <- glm(prob ~ age + race + smoke,
             family = binomial,
             data = dataA1_prob)
summary(glm)
```

```
## 
## Call:
## glm(formula = prob ~ age + race + smoke, family = binomial, data = dataA1_prob)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.4211  -0.9171  -0.5687   1.3687   2.0707  
## 
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)   
## (Intercept)     0.09300    0.82306   0.113  0.91004   
## age            -0.03488    0.03340  -1.044  0.29634   
## raceblack       1.01141    0.49342   2.050  0.04039 * 
## raceother       1.05673    0.40596   2.603  0.00924 **
## smokenonsmoker -1.10055    0.37195  -2.959  0.00309 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 218.86  on 184  degrees of freedom
## AIC: 228.86
## 
## Number of Fisher Scoring iterations: 4
```

```
pred_data <- data.frame(age = 30,
                        race = "black",
                        smoke = "nonsmoker")

prob_low_bwt <- predict(glm, pred_data, type = "response")
prob_low_bwt
```

```
##         1 
## 0.2606441
```

The predicted probability of low birth weight for a baby with a 30 year old black mother who does not smoke is 26.1%.

Q8

```
dataA1_agecat <- dataA1 %>% mutate(agecat = ifelse(age <= 20, "20 or younger",
                                                   ifelse(age <= 30, "age between 21-30",
                                                                     "older than 30")))
dataA1_groups <- dataA1_agecat %>%
  group_by(agecat,
           race,
           smoke) %>%
  summarise(lower_bwt = sum(bwt > 2500),
            higher_bwt = sum(bwt < 2500))
```

Q9

```
glm2 <- glm(cbind(lower_bwt, higher_bwt) ~ agecat + race + smoke,
            family = binomial,
            data = dataA1_groups)
summary(glm2)
```

```
## 
## Call:
## glm(formula = cbind(lower_bwt, higher_bwt) ~ agecat + race + 
##     smoke, family = binomial, data = dataA1_groups)
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -1.24690  -0.68978   0.05403   0.84686   1.83054  
## 
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)   
## (Intercept)               0.7204     0.3440   2.094  0.03623 * 
## agecatage between 21-30  -0.1745     0.3473  -0.503  0.61528   
## agecatolder than 30       0.8255     0.7123   1.159  0.24648   
## raceblack                -1.0928     0.4960  -2.203  0.02759 * 
## raceother                -1.0648     0.4065  -2.619  0.00881 **
## smokenonsmoker            1.1466     0.3760   3.050  0.00229 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 32.280  on 16  degrees of freedom
## Residual deviance: 15.222  on 11  degrees of freedom
## AIC: 62.315
## 
## Number of Fisher Scoring iterations: 4
```

```
# Odds ratios:
or <- round((exp(coef(glm2)[-1]) - 1) * 100, 1)
increase <- (or>0)
```

As we can see from the summary, the variables race and smoke are significant at 5% level. By analyzing the results using odds ratios, we can interpret the coefficients for race and smoke:

- odds of saying that the number of low birth weight babies is wrong decrease by 66.5% with the race of black when all other variables remain constant.
- odds of saying that the number of low birth weight babies is wrong decrease by 65.5% with other races when all other variables remain constant.
- odds of saying that the number of low birth weight babies is wrong increase by 214.7% with non-smoking mothers when all other variables remain constant.

Q10

Although the same age, race, and smoke are used as predictors in all three questions, the estimated coefficients have different signs across the estimated regressions for the following reasons.

- The sample size and variability of the data affect the signs of the estimated coefficients due to the slight difference in the data used in each model. Small sample sizes or high variability of the data can lead to unstable estimates and may result in different signs of the estimated coefficients in different regression models.
- Due to the different interactions between predictor variables in different models.
- The different treatment of the response variables: Q3 is modeled purely using a linear model, Q7 is the predicted likelihood, and Q9 is treating low-weight infants to the binary response, leading to different signs of the estimated coefficients.

Q11

```
glm3 <- glm(cbind(lower_bwt, higher_bwt) ~ agecat + race + smoke,
            family = quasibinomial,
            data = dataA1_groups)
summary(glm3)
```

```
##
## Call:
## glm(formula = cbind(lower_bwt, higher_bwt) ~ agecat + race +
##     smoke, family = quasibinomial, data = dataA1_groups)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.24690  -0.68978   0.05403   0.84686   1.83054
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.7204     0.3959   1.820   0.0961 .
## agecatage between 21-30  -0.1745     0.3997  -0.437   0.6708
## agecatolder than 30        0.8255     0.8197   1.007   0.3355
## raceblack                 -1.0928     0.5708  -1.914   0.0819 .
## raceother                 -1.0648     0.4678  -2.276   0.0438 *
## smokenonsmoker             1.1466     0.4326   2.650   0.0226 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 1.32428)
##
##     Null deviance: 32.280  on 16  degrees of freedom
## Residual deviance: 15.222  on 11  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

```
pchisq(deviance(glm3), df.residual(glm3))
```

```
## [1] 0.8274378
```

```
anova(glm2, glm3, test = "F")
```

```
## Warning: using F test with a 'binomial' family is inappropriate
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(lower_bwt, higher_bwt) ~ agecat + race + smoke
## Model 2: cbind(lower_bwt, higher_bwt) ~ agecat + race + smoke
##   Resid. Df Resid. Dev Df Deviance F Pr(>F)
## 1        11     15.222
## 2        11     15.222  0        0
```

By using Quasi-binomial to adjust for overdispersion in this question, we can see that the results are the same when comparing the generated model with Q9. I believe the two reasons why overdispersion may not be a problem are:

- Mean pi determines the variance pi(1 − pi).
- Percentage observations pi equal to 1 and 1 − pi to 0.