Final Exam

Matthew Bentz

5/3/2022

Part 1 - Knowledge-Based Questions

1. a) Show that $P(A) = 1 - P(A^c)$.

We are going to use the fact $(A \cup A^c) = \mathcal{S}$ and that $A \cap A^c = \emptyset$ (they are mutually exclusive). With these two facts, we have that:

$$\begin{aligned} P(\mathcal{S}) &= 1 \Rightarrow \\ P(A \cup A^c) &= 1 \\ P(A) + P(A^c) &= 1 \\ P(A) &= 1 - P(A^c) \end{aligned}$$

b.) Show that $P(\emptyset) = 0$.

We again use the fact that $\mathcal{S} \cup \emptyset = \mathcal{S}$ (everything unioned with nothing means we still have everything), and the fact that $\mathcal{S} \cap \emptyset = \emptyset$ (everything and nothing have nothing in common) to have that:

$$\begin{aligned} P(\mathcal{S}) &= 1 \\ P(\mathcal{S} \cup \emptyset) &= 1 \\ P(\mathcal{S}) + P(\emptyset) &= 1 \\ 1 + P(\emptyset) &= 1 \\ P(\emptyset) &= 0 \end{aligned}$$

2. Show that events A and B cannot be both mutually exclusive and independent.

We define mutual exclusivity as $P(A \cup B) = P(A) + P(B)$ or $P(A \cap B) = 0$ and independence as $P(A \cap B) = P(A)P(B)$. If we assume the events are independent:

$$P(A \cap B) = P(A)P(B)$$

Then if the events were also mutually exclusive:

$$P(A \cap B) = 0$$

This shows that the events cannot be mutually exclusive if they are independent (and vice versa) because the probabilities of events A and B are greater than 0, $P(A)P(B) \neq 0$.

3. Find the value of c such that $f(y)$ is a probability distribution.

The total probability of a distribution must be equal to 1, shown by $P(\mathcal{S}) = 1$. We can use this fact to find c by setting $f(y) = 1$ and plugging in all values for y.

$$\begin{aligned} f(y) &= 1 \Rightarrow \\ cy &= 1 \\ c1 + c2 + c3 + c4 + c5 &= 1 \\ c(1 + 2 + 3 + 4 + 5) &= 1 \\ c(15) &= 1 \\ c &= \frac{1}{15} \end{aligned}$$

4. Find the value of c such that $f(x)$ is a probability distribution.

We can again use the fact that the total probability of $f(x)$ will be equal to 1. Because x is a continuous random variable, we will have to take the integral of $f(x)$ to find our c value.

$$f(x) = \begin{cases} ce^{-.15(x-.5)} & x \geq .5 \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} \int_{0.5}^{\infty} ce^{-.15(x-.5)} dx &= 1 \\ c(6.667) &= 1 \\ \Rightarrow c &= .15 \end{aligned}$$

5. a.) Find the maximum likelihood estimator for θ.

Step 1: Find the likelihood function $L(\theta) = \prod_{i=1}^{n} f(x_i)$.

$$L(\theta) = (\theta + 1)^{10} \prod_{i=1}^{10} (x_i)^{\theta}$$

Step 2: Take the log of the likelihood function.

$$\begin{aligned} ln(L) &= 10ln(\theta + 1) + ln \prod_{i=1}^{10} (x_i)^{\theta} \\ &= 10ln(\theta + 1) + \theta \sum_{i=1}^{10} ln(x_i) \end{aligned}$$

Step 3: Calculate the first derivative and set it equal to 0.

$$\frac{\partial ln(L)}{\partial\theta} = \frac{10}{\theta+1} + \sum_{i=1}^{10} ln(x_i) = 0$$

Step 4: Solve for the parameter θ.

$$\hat{\theta} = \frac{-10 - \sum_{i=1}^{10} ln(x_i)}{\sum_{i=1}^{10} ln(x_i)}$$

b.) Computing the point estimate with the given data.

$$\begin{aligned}\hat{\theta} &= \frac{-10 - \sum_{i=1}^{10} ln(x_i)}{\sum_{i=1}^{10} ln(x_i)}\\ &= \frac{-10 + 2.4295}{-2.4295}\\ &\Rightarrow 3.116\end{aligned}$$

Part II - Statistical Analysis of a Dataset

6. Accessing the data set.

```
library(MASS)
data(birthwt)
```

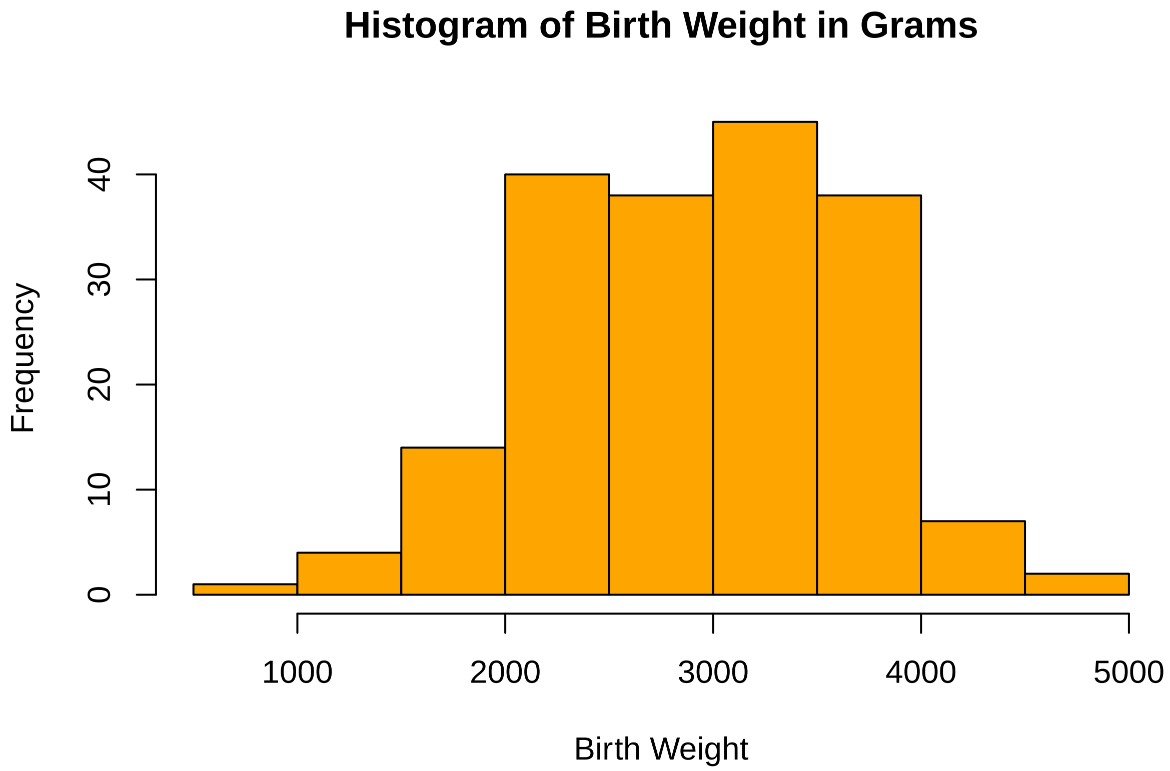
The data from the ‘birthwt’ data set consists of 189 samplings of a newborn and his/her mother to analyze the risk factors associated with an infant’s low birth weight. It is structured in a way that all points of data can be expressed numerically, and each of the 10 columns represents a recording of data in the 189 rows. There are three types of variables in the set - true/false indicators, placeholders, and measurements. The true/false indicators (0=false, 1=true) are the variables for the birth weight being less than 2.5kg (low), smoking status during pregnancy (smoke), history of hypertension (ht), and presence of uterine irritability (ui). The placeholder type variable is associated with the mother’s race (race); 1 = white, 2 = black, 3 = other. Finally, the measurements variables are the mother’s age in years (age), the mother’s weight in pounds at the last menstrual period (lwt), number of previous premature labours (ptl), number of physician visits during the first trimester (ftv), and the birth weight in grams (bwt).

7. Removing ‘low’ and coercing variables as factors.

```
birthdata <- subset(data.frame(birthwt), select = -c(1))
birthdataNoFactors = birthdata
birthdata$smoke = as.factor(birthdata$smoke)
birthdata$ht = as.factor(birthdata$ht)
birthdata$ui = as.factor(birthdata$ui)
birthdata$race = as.factor(birthdata$race)
```

8. Histogram for birth weights.

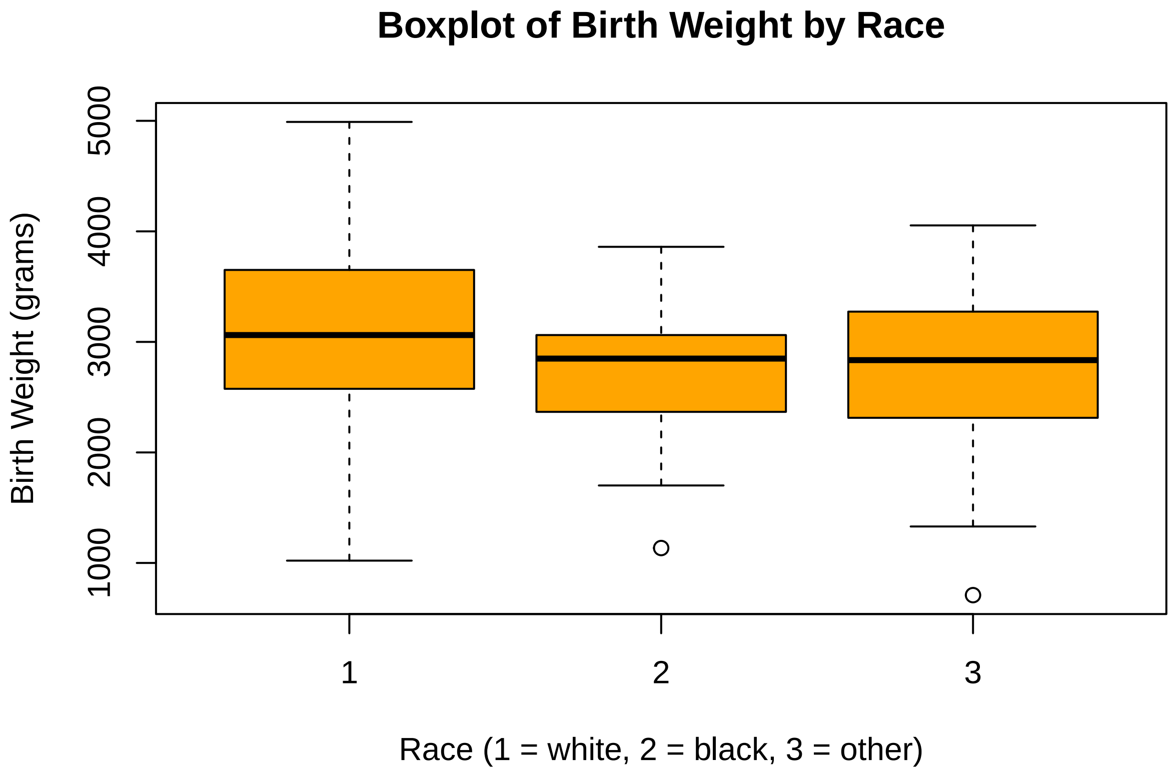
```
hist(birthdata$bwt,
     main = 'Histogram of Birth Weight in Grams',
     xlab = 'Birth Weight',
     ylab = 'Frequency',
     col = 'orange')
```



The histogram for birth weight resembles a normal distribution centered around 3000 grams. The histogram also show a unimodal and symmetric, bell-shaped curve with a negligible left skew.

9. Birth Weight boxplots by race.

```
boxplot(birthdata$bwt ~ birthdata$race,
        main = 'Boxplot of Birth Weight by Race',
        xlab = 'Race (1 = white, 2 = black, 3 = other)',
        ylab = 'Birth Weight (grams)',
        col = 'orange')
```



The medians are unexpectedly relatively the same, around 3000g (denoted by the black line). The spread of birth weight in white mothers is much more comparative to black and other races. White mothers also have a noticeably higher interquartile range (denoted by the orange box), which represents 25% to 75% of the samples taken. We could conclude from this boxplot that white mothers may have heavier infants but also have a greater variance when compared to infants of other races.

10. Two-sample t-tests for population mean birth weights among the three different races.

```
# First, we must find the standard deviations of each sample to
# determine whether or not to conduct pooled hypothesis tests.

bwtWhite <- subset(birthdata$bwt, birthdata$race == 1)
bwtBlack <- subset(birthdata$bwt, birthdata$race == 2)
bwtOther <- subset(birthdata$bwt, birthdata$race == 3)

sd(bwtWhite)
```

```
## [1] 727.8861
```

```
sd(bwtBlack)
```

```
## [1] 638.6839
```

```
sd(bwtOther)
```

```
## [1] 722.1944
```

Based on our results, we can say that $\sigma_{white} = \sigma_{other}$, but $\sigma_{black} \neq \sigma_{white}$ and $\sigma_{black} \neq \sigma_{other}$.

```
t.test(bwtWhite, bwtBlack, var.equal = F, conf.level = 0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  bwtWhite and bwtBlack
## t = 2.6301, df = 44.241, p-value = 0.0117
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##   -8.953359 775.006243
## sample estimates:
## mean of x mean of y
##  3102.719  2719.692
```

```
t.test(bwtWhite, bwtOther, var.equal = T, conf.level = 0.99)
```

```
##
##  Two Sample t-test
##
## data:  bwtWhite and bwtOther
## t = 2.5751, df = 161, p-value = 0.01092
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##   -3.647602 598.517938
## sample estimates:
## mean of x mean of y
##  3102.719  2805.284
```

```
t.test(bwtBlack, bwtOther, var.equal = F, conf.level = 0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  bwtBlack and bwtOther
## t = -0.55865, df = 51.19, p-value = 0.5788
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -495.4831  324.3005
## sample estimates:
## mean of x mean of y
##  2719.692  2805.284
```

We performed three two-sample t-tests to see whether there was a statistical significance in the population mean birth weights among the three difference races at a confidence of 99%. We got p-values that were all greater than our level of significance value, $\alpha = .01$, so we fail to reject the null hypothesis. That is, the population mean birth weights among the three races are not statistically different at a level of significance $\alpha = .01$.

11. Fitting an ANOVA model using race.

```
# We first need to alter the lengths of our data
# sets so that they can fit our data frame.

length(bwtBlack) <- length(bwtWhite)
length(bwtOther) <- length(bwtWhite)

suppressWarnings(library(tidyr))
bwtRace = data.frame(bwtWhite, bwtBlack, bwtOther)
bwtRace = gather(bwtRace, key = "Race", value = "Weight", bwtWhite, bwtBlack, bwtOther)
bwtRace$Race = as.factor(bwtRace$Race)

raceAOV = aov(bwtRace$Weight ~ bwtRace$Race)
summary(raceAOV)
```

```
##                 Df   Sum Sq Mean Sq F value  Pr(>F)
## bwtRace$Race   2  5015725 2507863   4.913 0.00834 **
## Residuals    186 94953931  510505
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 99 observations deleted due to missingness
```

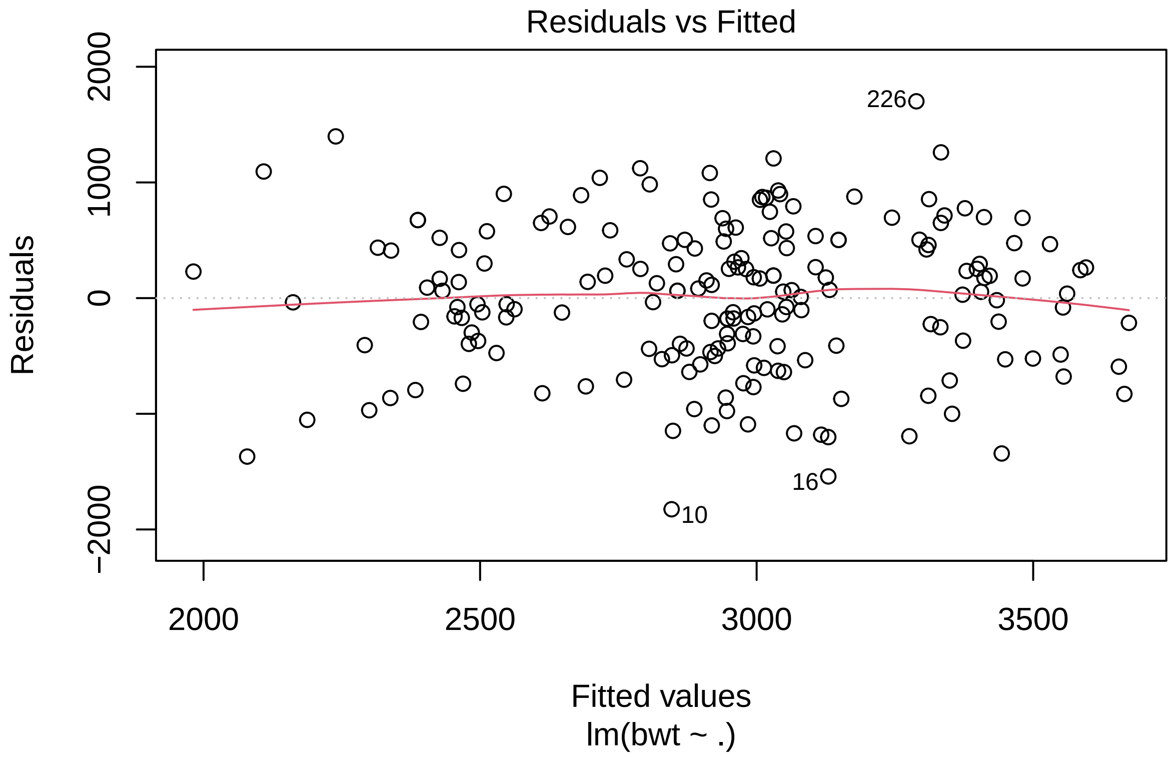
We performed a one-way ANOVA test for the population means of birth weight being statistically different between the three races. Our test statistic was $F = 4.913$ with $df_1 = 2$ and $df_2 = 186$, which corresponds to a p-value of 0.00834. Since $p < \alpha$ we can reject the null hypothesis that the mean birth weights between the three races are not statistically different. In other words, we can say that there are at least two races in which there are statistically significant differences in birth weights. The output from this ANOVA model does not match the previous question, in which we found no statistical difference in the population mean birth weights of the three races. We may consider fighting for the ANOVA model instead of the three separate hypothesis tests because it does not compare them individually which results in smaller room for error. When comparing the samples individually, we allow for a margin of error that compounds with each comparison.

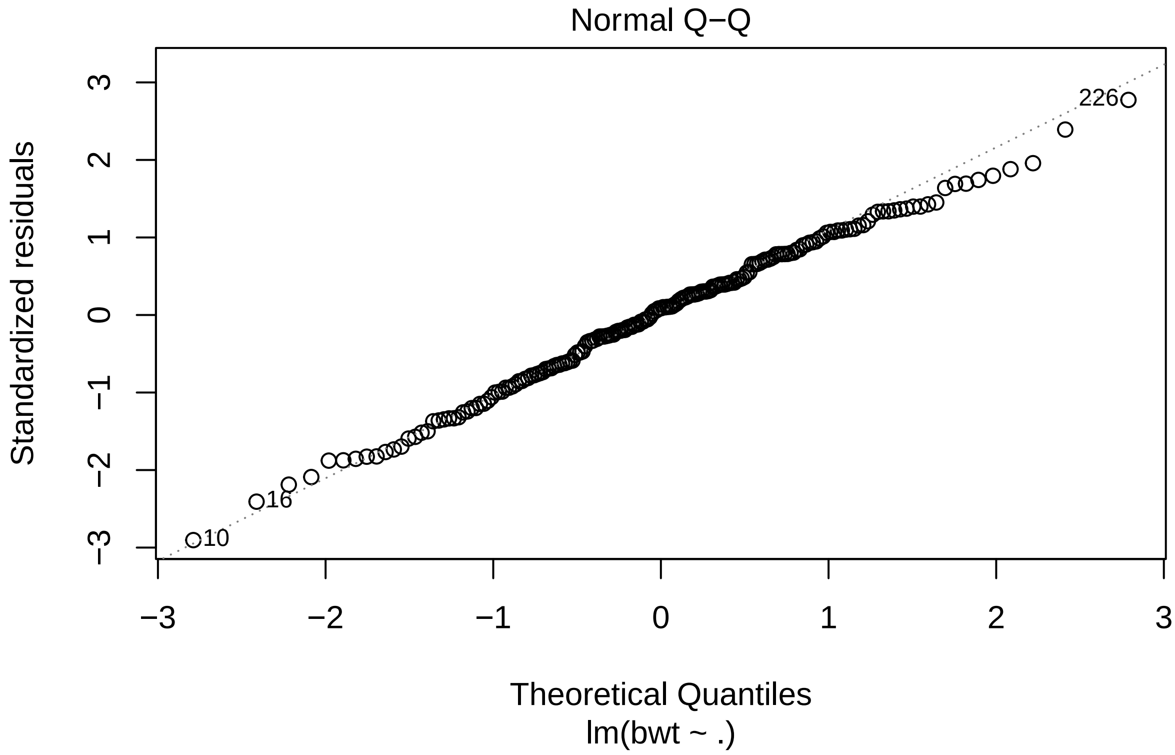
12. Fitting a regression model.

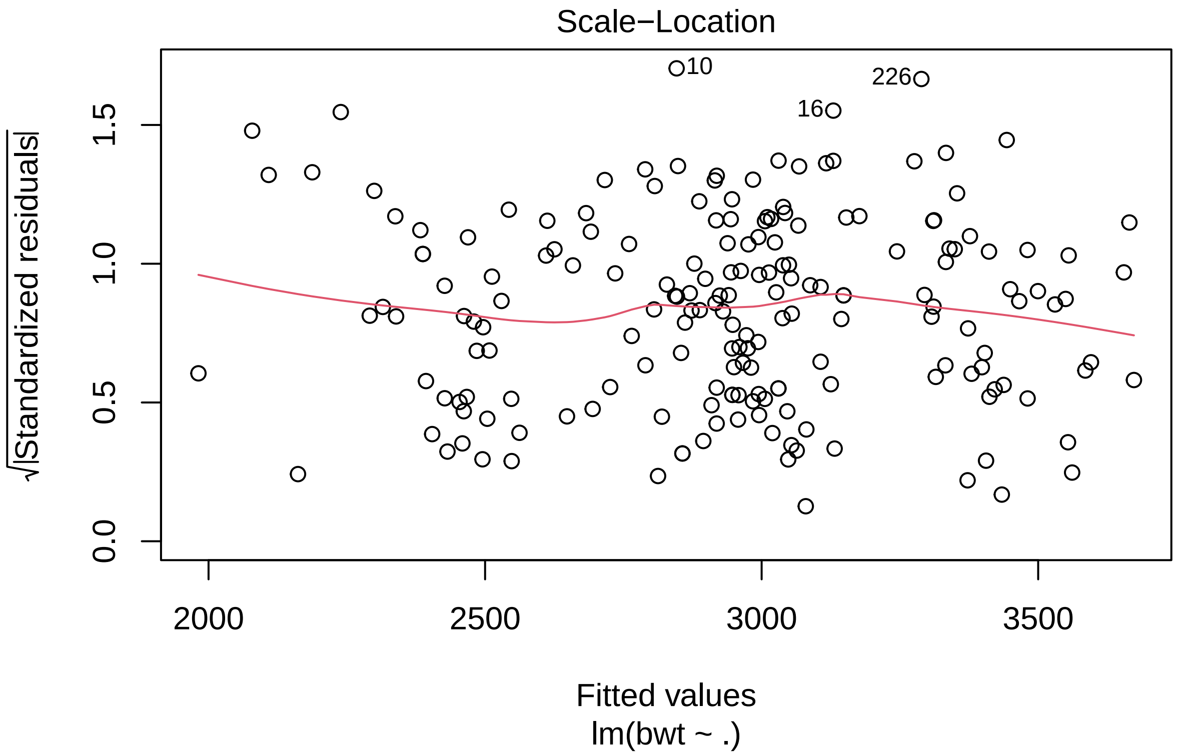
```
birthFit = lm(bwt ~ ., data = birthdata)
summary(birthFit)
```

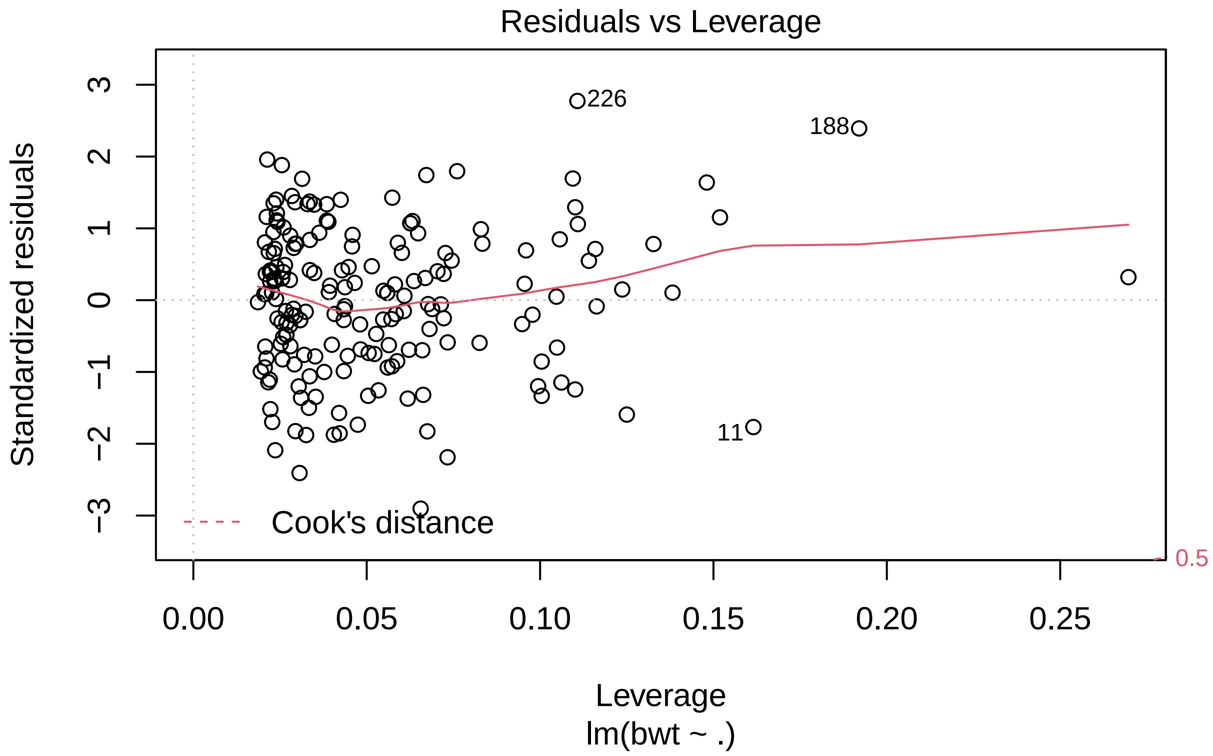
```
##
## Call:
## lm(formula = bwt ~ ., data = birthdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1825.26  -435.21    55.91   473.46  1701.20
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2927.962    312.904   9.357  < 2e-16 ***
## age           -3.570      9.620  -0.371 0.711012
## lwt            4.354      1.736   2.509 0.013007 *
## race2       -488.428    149.985  -3.257 0.001349 **
## race3       -355.077    114.753  -3.094 0.002290 **
## smoke1      -352.045    106.476  -3.306 0.001142 **
## ptl          -48.402    101.972  -0.475 0.635607
## ht1         -592.827    202.321  -2.930 0.003830 **
## ui1         -516.081    138.885  -3.716 0.000271 ***
## ftv          -14.058     46.468  -0.303 0.762598
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 650.3 on 179 degrees of freedom
## Multiple R-squared:  0.2427, Adjusted R-squared:  0.2047
## F-statistic: 6.376 on 9 and 179 DF,  p-value: 7.891e-08
```

```
plot(birthFit)
```




Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
226
16
10
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(bwt ~ .)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2000
2500
3000
3500
Fitted values
lm(bwt ~ .)
10
226
16



Variables *lwt*, *race*, *smoke*, *ht*, and *ui* are significant predictors of an infants birth weight. 24.27% of the total variability of birth weight is explained by the linear combination of these variables. No assumptions about the regression model are violated.

13. Finding the pairwise correlations.

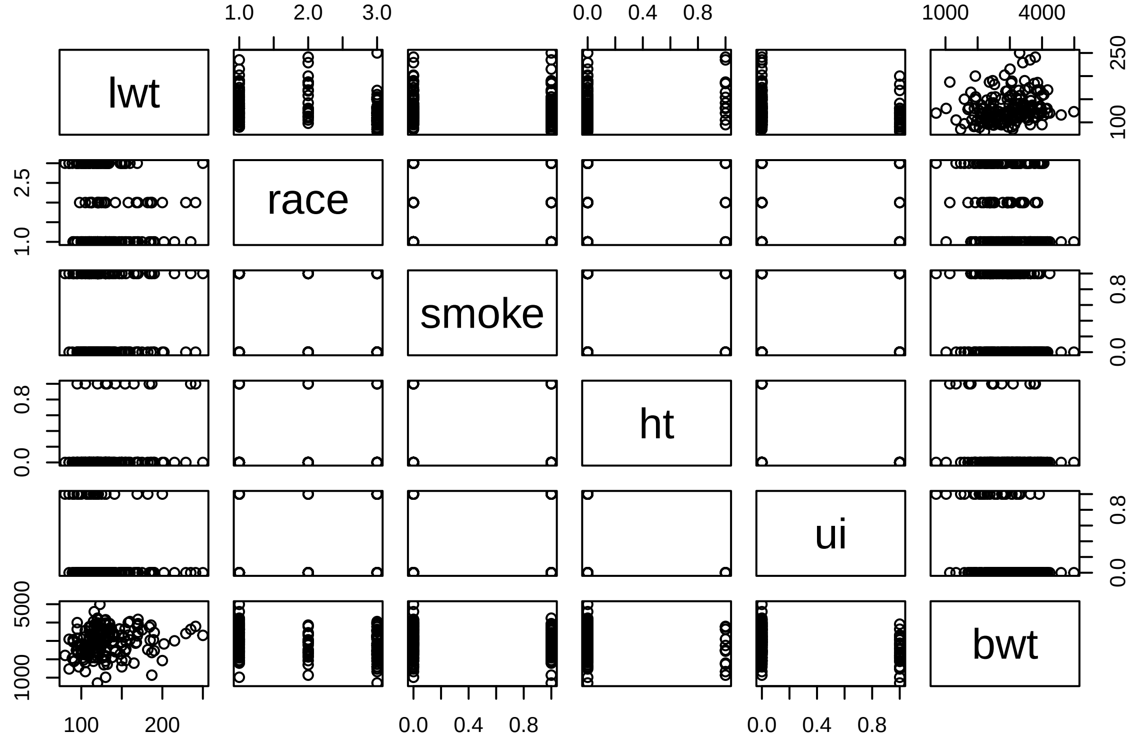
```
birthdataSig <- subset(birthdataNoFactors, select = c(2,3,4,6,7,9))
cor(birthdataSig)
```

```
##               lwt        race       smoke          ht          ui        bwt
## lwt    1.00000000 -0.16504854 -0.04417908  0.23636040 -0.15276317  0.1857333
## race  -0.16504854  1.00000000 -0.33903074  0.01992992  0.05360209 -0.1947135
## smoke -0.04417908 -0.33903074  1.00000000  0.01340704  0.06215900 -0.1904481
## ht     0.23636040  0.01992992  0.01340704  1.00000000 -0.10858506 -0.1459819
## ui    -0.15276317  0.05360209  0.06215900 -0.10858506  1.00000000 -0.2839274
## bwt    0.18573328 -0.19471349 -0.19044806 -0.14598189 -0.28392741  1.0000000
```

Based on our correlation values, we can see that there is no one single factor that has a strong linear relationship with birth weight. For all five factors we see a range of values from [-0.284, 0.186] which would be considered as very weak or no correlation to the birth weight.

14. Plotting the pairwise scatterplot matrix.

```r
pairs(birthdataSig)
```



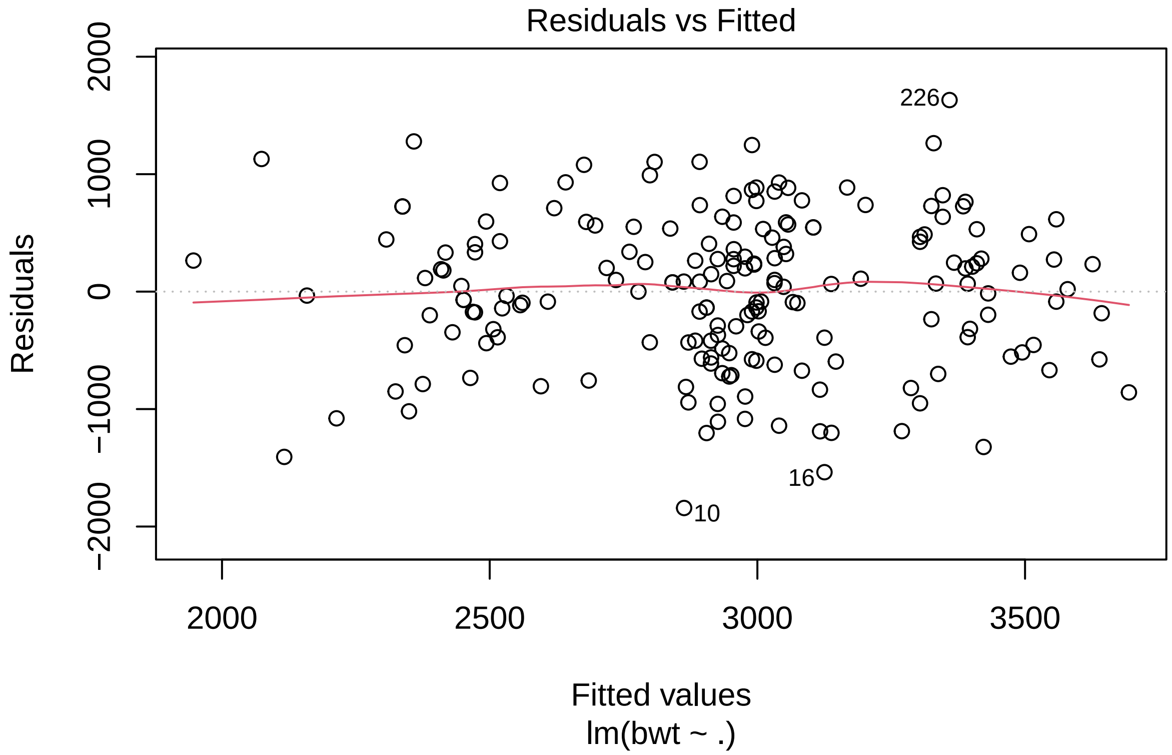
When looking at our scatterplot matrix, we again can see no linear relationship between any variable and infant's birth weight. You could argue that mother's weight (lwt) has some sort of relationship, but mostly due to the fact that the data points are measurements and not factors. It is interesting to note that none of the single variables has a noticeable effect, but combined they are decent predictors.

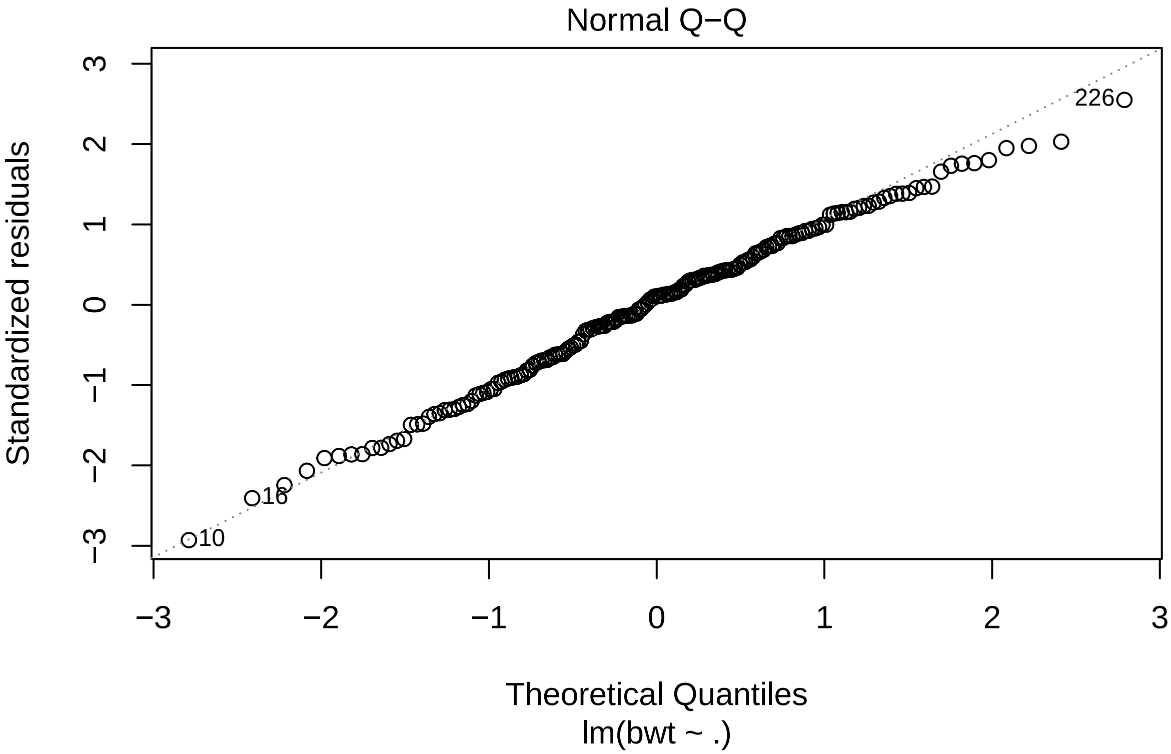
15. Fitting another regression model.

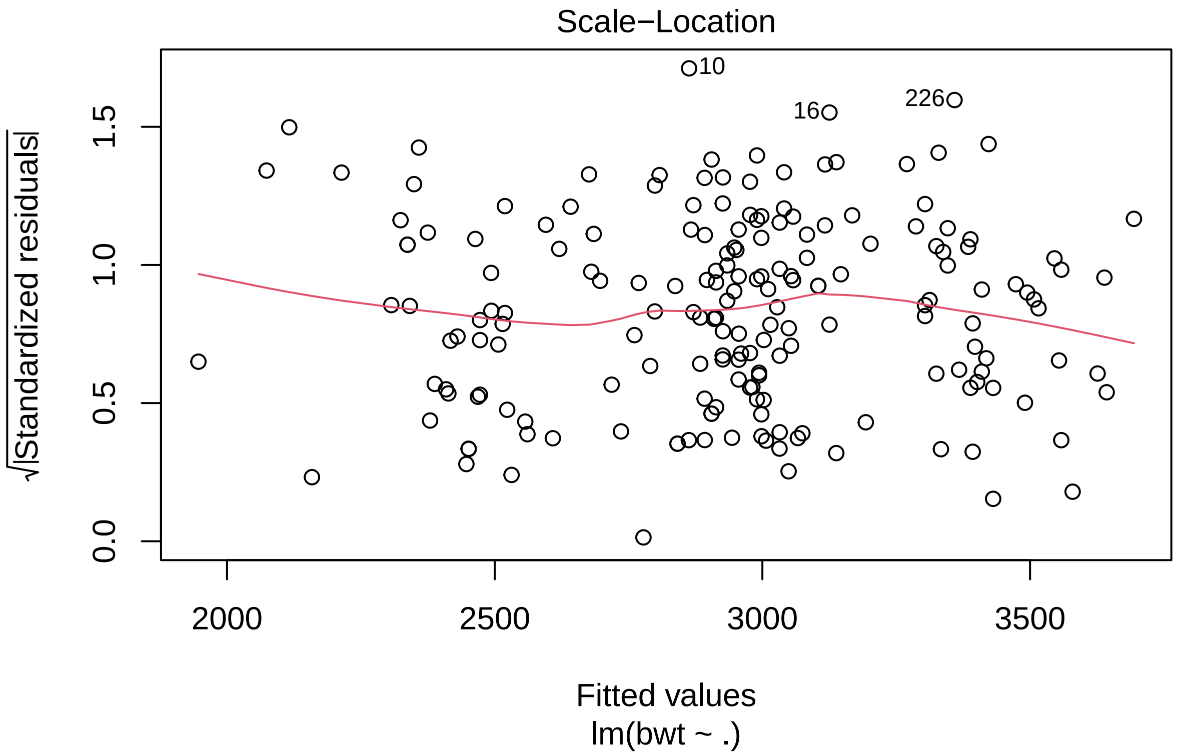
```r
birthdataSig <- subset(birthdata, select = c(2,3,4,6,7,9))
birthdataSigFit = lm(bwt ~ ., data = birthdataSig)
summary(birthdataSigFit)
```

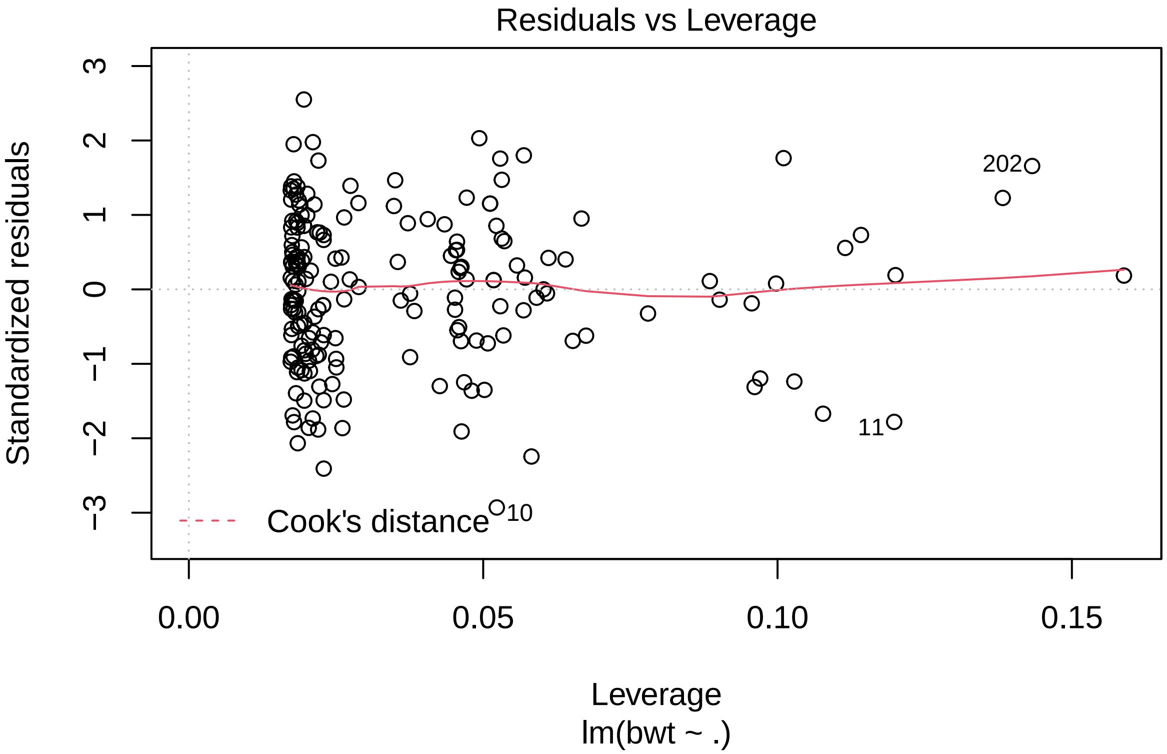
```
##
## Call:
## lm(formula = bwt ~ ., data = birthdataSig)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1842.14  -433.19    67.09   459.21  1631.03
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2837.264    243.676  11.644  < 2e-16 ***
## lwt            4.242      1.675   2.532 0.012198 *
## race2       -475.058    145.603  -3.263 0.001318 **
## race3       -348.150    112.361  -3.099 0.002254 **
## smoke1      -356.321    103.444  -3.445 0.000710 ***
## ht1         -585.193    199.644  -2.931 0.003810 **
## ui1         -525.524    134.675  -3.902 0.000134 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 645.9 on 182 degrees of freedom
## Multiple R-squared:  0.2404, Adjusted R-squared:  0.2154
## F-statistic:   9.6 on 6 and 182 DF,  p-value: 3.601e-09
```

```
plot(birthdataSigFit)
```




Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
226
16
10
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(bwt ~ .)


Scale−Location
10
16
226
0.0
0.5
1.0
1.5
√|Standardized residuals|
2000
2500
3000
3500
Fitted values
lm(bwt ~ .)



The two models are very similar to one another, however, I would say that this model containing only significant predictors is better because it is parsimonious. Both models have approximately the same coefficient of determination (R^2), but the second has a greater predictive power due to it's parsimony. This second model will be more immutable towards new data because insignificant variables are not included as predictors to birth weight since they are uncorrelated.