Final Project

Kyle Duplessis, Dean Gladish & Kavie Yu

6/02/2018

After installing some prerequisite packages, we are ready to begin our exploratory data analysis.

To begin with, we added to the dataset a binary version of the TobaccoUse variable, a factored Region variable indicating general location, a factored version of the EducationCode variable (shortened), and numericized the AverageBirthWeight variable.

```
# Turns Tobacco Use into 1s and 0s - yes and no
Natality <- mutate(Natality, TobaccoUseCodeBinary = TobaccoUseCode - 1)

# Removes all table entries where TobaccoUse is Not Stated
Natality <- Natality[!(Natality$TobaccoUse == "Not Stated"),]

# Removes all table entries where Education is not stated
Natality <- Natality[!(Natality$Education == "Unknown/Not on certificate"),]

Natality$EducationCode <- factor(Natality$EducationCode, labels = c("8thGrade", "12thGrade", "HS/GED",

# Removes all table entries where DeliveryMethod is not stated.

Natality <- Natality[!(Natality$DeliveryMethod == "Not Stated"),]

Natality$Region <- factor(Natality$Region, labels = c("NorthEast", "MidWest", "South", "SouthWest", "We

# Turns the Average Birth Weight column into a numeric.

Natality <- mutate(Natality, AverageBirthWeight = as.numeric(AverageBirthWeight))
```

(A) The Creation of New Data

Since our goal for this study is to predict an age range for mothers who are interested in having a baby with a "healthy weight", we will examine three models and select our preferred model. We will also set aside 25% of our data to test the efficacy of our chosen model.

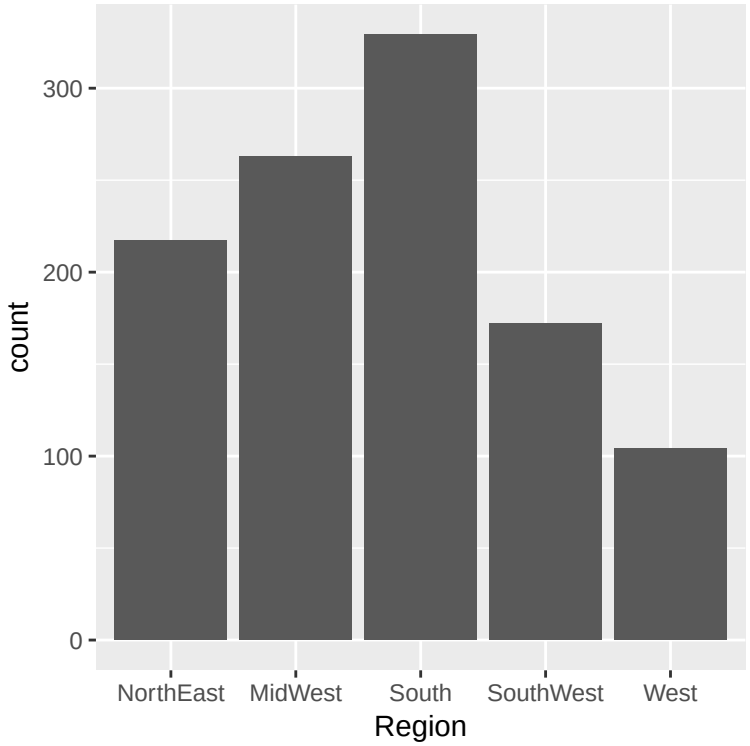
We created two new datasets to test and train our model.

```
index <- sample(nrow(Natality), size = nrow(Natality)*0.75)
train_Natality <- Natality[index,]
test_Natality <- Natality[-index,]
```

(B) The Data Analysis

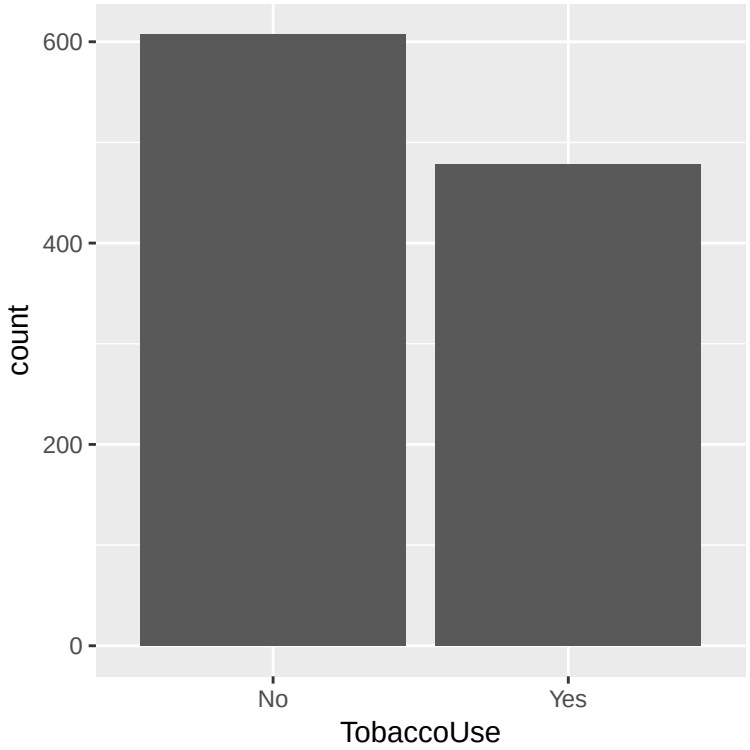
(i) barplots

```
gf_bar(~Region, data = train_Natality)
```



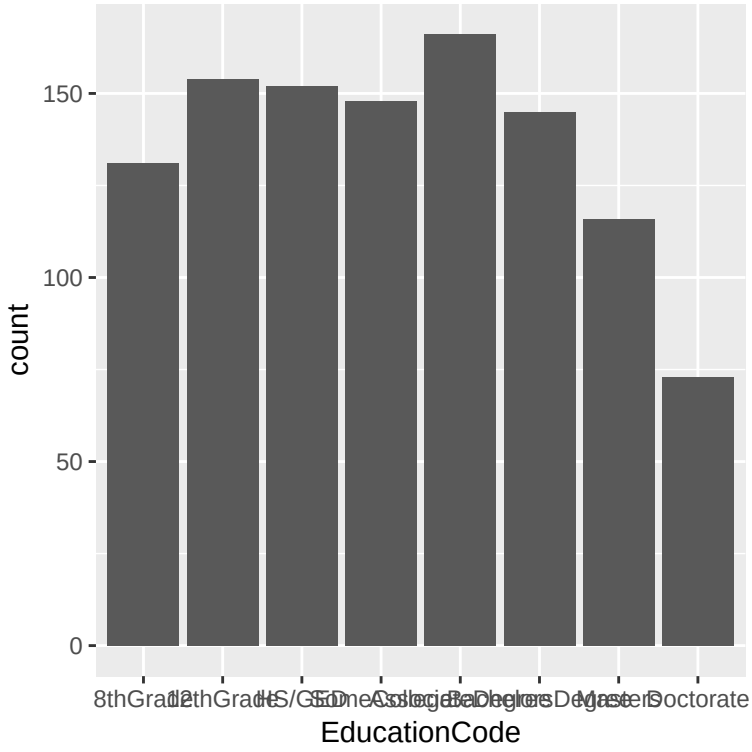
The above barplot indicates that there are the most entries from the South and the fewest numer of entries are from the West.

```
gf_bar(~TobaccoUse, data = train_Natality)
```



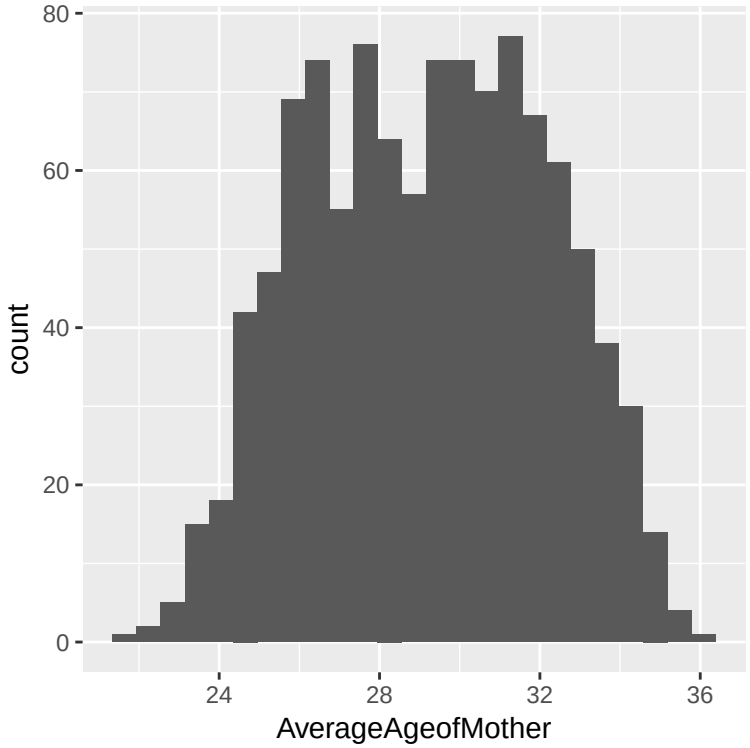
The above barplot indicates that proportionally less people use tobacco in our dataset.

```
gf_bar(~EducationCode, data = train_Natality)
```



This plot shows that there is a drop-off in frequency as we approach higher education levels.

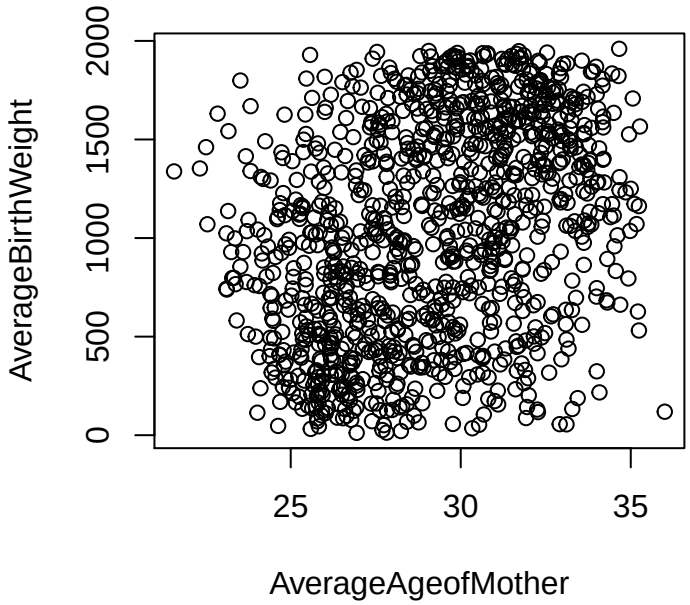
```
gf_histogram(~AverageAgeofMother, data = train_Natality)
```



This histogram demonstrates that the average age of the mother is generally normally distributed; while it reaches a plateau in the middle of the graph, it also has tails that diminish very quickly and as such the average age of the mother is normally distributed.

(ii) Scatterplot

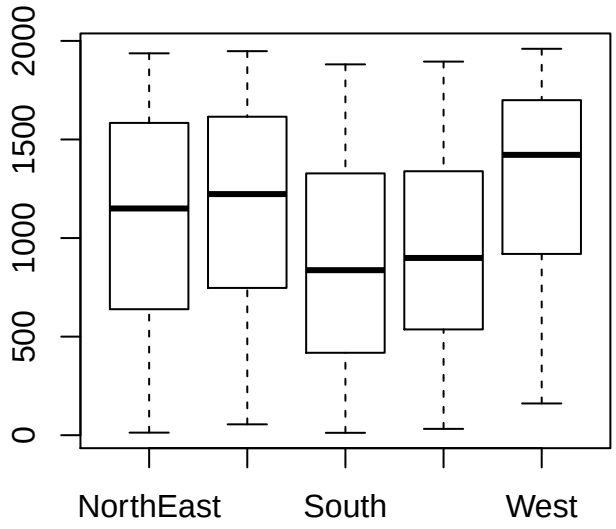
```
plot(AverageBirthWeight ~ AverageAgeofMother , data = train_Natality)
```



This scatterplot shows that despite a large amount of variability, there is a discernably linear and positive association between average age of mother and average birth weight.

(iii) Boxplots

```
boxplot(AverageBirthWeight ~ Region, data = train_Natality)
```



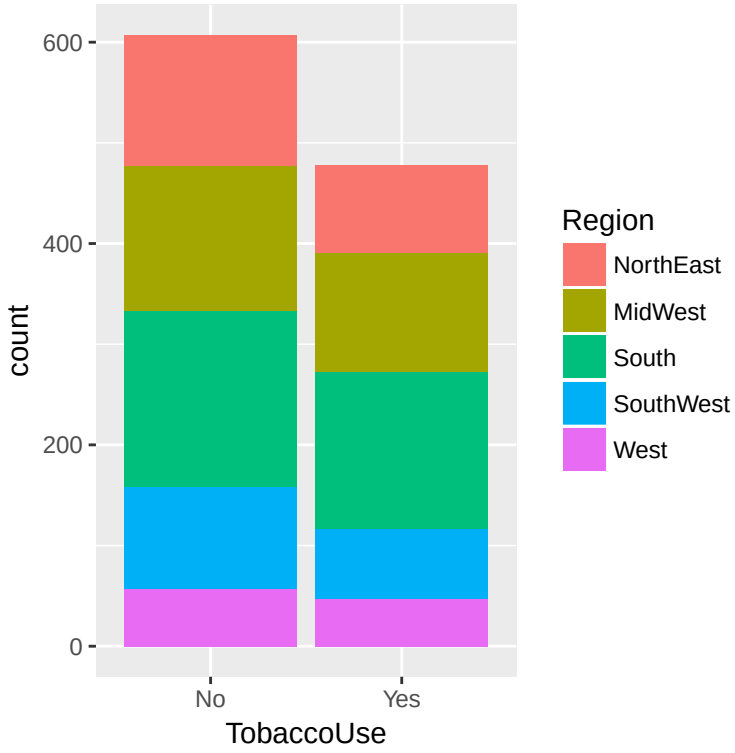
This boxplot allows us to see that the distribution of birth weight is different for each region; thus, region is an important variable to include in our model.

(iv) Barplots

```
library(viridisLite)
```

```
## Warning: package 'viridisLite' was built under R version 3.4.4
```

```
gf_bar( ~ TobaccoUse, data = train_Natality, fill = ~Region)
```



From this barplot of frequency count and TobaccoUse against Region, we can see for example that the West contributes the least count to both groups.

Our Model:

We created two models - mod.base is our linear regression model on all important variables with no interaction terms included. The only difference between mod.base and mod.plus is that mod.plus assumes that log(AverageBirthWeight) has a linear relationship with our variables instead of AverageBirthWeight having a linear relationship with the rest of the variables in our model.

```
library(MASS)
mod.base <- lm(AverageBirthWeight ~ Region + DeliveryMethod + TobaccoUse + AverageAgeofMother, data = Na
mod.plus <- lm(log(as.numeric(AverageBirthWeight)) ~ Region + DeliveryMethod + TobaccoUse + AverageAgeo
summary(mod.base)
```

```
##
## Call:
## lm(formula = AverageBirthWeight ~ Region + DeliveryMethod + TobaccoUse +
##     AverageAgeofMother, data = Natality)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1150.80  -173.87    -7.91   170.54  1281.48
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -460.811     85.454  -5.392 8.11e-08 ***
## RegionMidWest          109.799     23.217   4.729 2.47e-06 ***
## RegionSouth           -109.610     22.092  -4.962 7.82e-07 ***
```

```
## RegionSouthWest       -165.125     26.211  -6.300 3.95e-10 ***
## RegionWest             189.590     30.047   6.310 3.71e-10 ***
## DeliveryMethodVaginal  392.063     15.794  24.824  < 2e-16 ***
## TobaccoUseYes         -698.729     15.722 -44.443  < 2e-16 ***
## AverageAgeofMother      55.421      2.707  20.470  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 293.7 on 1439 degrees of freedom
## Multiple R-squared:  0.7023, Adjusted R-squared:  0.7008
## F-statistic: 484.9 on 7 and 1439 DF,  p-value: < 2.2e-16
```

Then, we find the residuals of our mod.base:

```
library(broom)
resid_mod_base <- augment(mod.base)
head(resid_mod_base)
```

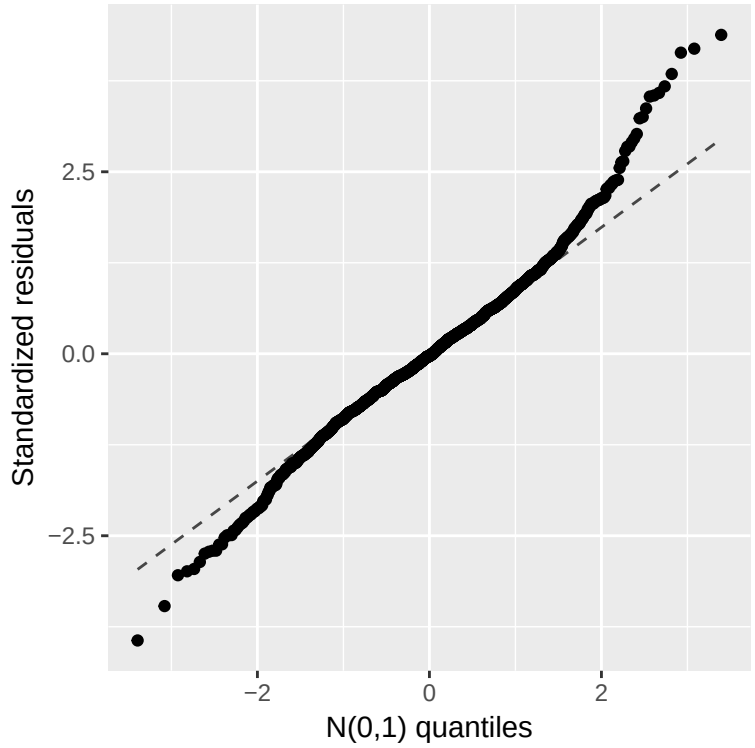
```
##   AverageBirthWeight    Region DeliveryMethod TobaccoUse
## 1               1607 NorthEast        Vaginal         No
## 2               1187 NorthEast       Cesarean         No
## 3                511 NorthEast        Vaginal        Yes
## 4               1097 NorthEast        Vaginal         No
## 5                461 NorthEast       Cesarean        Yes
## 6               1017 NorthEast       Cesarean         No
##   AverageAgeofMother   .fitted  .se.fit     .resid        .hat   .sigma
## 1              29.52 1567.2706 19.72714   39.72944 0.004512555 293.7657
## 2              31.25 1271.0857 19.75535  -84.08572 0.004525470 293.7592
## 3              26.33  691.7496 21.58892 -180.74964 0.005404508 293.7287
## 4              25.26 1331.1783 22.81304 -234.17834 0.006034767 293.7023
## 5              27.19  347.3488 22.06035  113.65121 0.005643117 293.7522
## 6              28.01 1091.5226 21.01623  -74.52263 0.005121579 293.7610
##        .cooksd .std.resid
## 1 1.041790e-05  0.1355944
## 2 4.680068e-05 -0.2869817
## 3 2.587154e-04 -0.6171650
## 4 4.855297e-04 -0.7998494
## 5 1.068530e-04  0.3881057
## 6 4.165282e-05 -0.2544194
```

(C) Model Assumptions Check

(i) QQ-plot for checking constant variance

The next few segments of code generate qq-plots that will allow us to determine if the log-transformation was necessary.

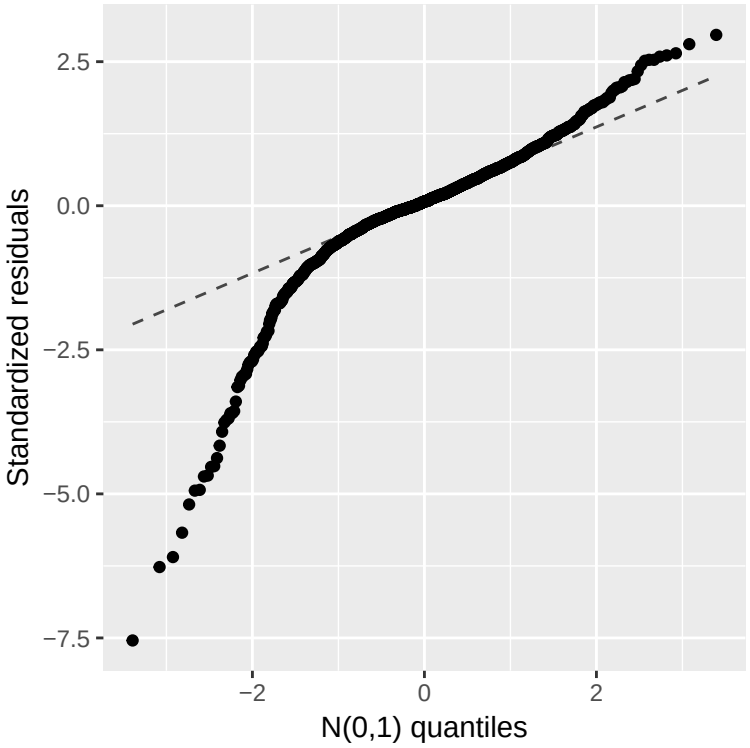
```
 gf_qq(~.std.resid, data = resid_mod_base) %>%
  gf_qqline() %>%
  gf_labs(x = "N(0,1) quantiles", y = "Standardized residuals" )
```




```
library(broom)
resid_mod_plus <- augment(mod.plus)
head(resid_mod_plus)
```

```
##   log.as.numeric.AverageBirthWeight..    Region DeliveryMethod TobaccoUse
## 1                            7.382124 NorthEast        Vaginal         No
## 2                            7.079184 NorthEast       Cesarean         No
## 3                            6.236370 NorthEast        Vaginal        Yes
## 4                            7.000334 NorthEast        Vaginal         No
## 5                            6.133398 NorthEast       Cesarean        Yes
## 6                            6.924612 NorthEast       Cesarean         No
##   AverageAgeofMother  .fitted    .se.fit      .resid        .hat    .sigma
## 1              29.52 7.413073 0.03473484 -0.03094817 0.004512555 0.5172548
## 2              31.25 7.003482 0.03478451  0.07570216 0.004525470 0.5172515
## 3              26.33 6.224059 0.03801300  0.01231087 0.005404508 0.5172553
## 4              25.26 7.121580 0.04016838 -0.12124542 0.006034767 0.5172455
## 5              27.19 5.754938 0.03884308  0.37845981 0.005643117 0.5171586
## 6              28.01 6.781784 0.03700463  0.14282880 0.005121579 0.5172416
##        .cooksd  .std.resid
## 1 2.039025e-06 -0.05998780
## 2 1.223550e-05  0.14673683
## 3 3.871170e-07  0.02387323
## 4 4.198080e-05 -0.23519369
## 5 3.821870e-04  0.73399743
## 6 4.935115e-05  0.27693427
```

```
  gf_qq(~.std.resid, data = resid_mod_plus) %>%
   gf_qqline() %>%
   gf_labs(x = "N(0,1) quantiles", y = "Standardized residuals" )
```



Transformation does not seem to be necessary. In fact, it seems to be detrimental as our data deviates more from the normal distribution when we use the log-transformation.

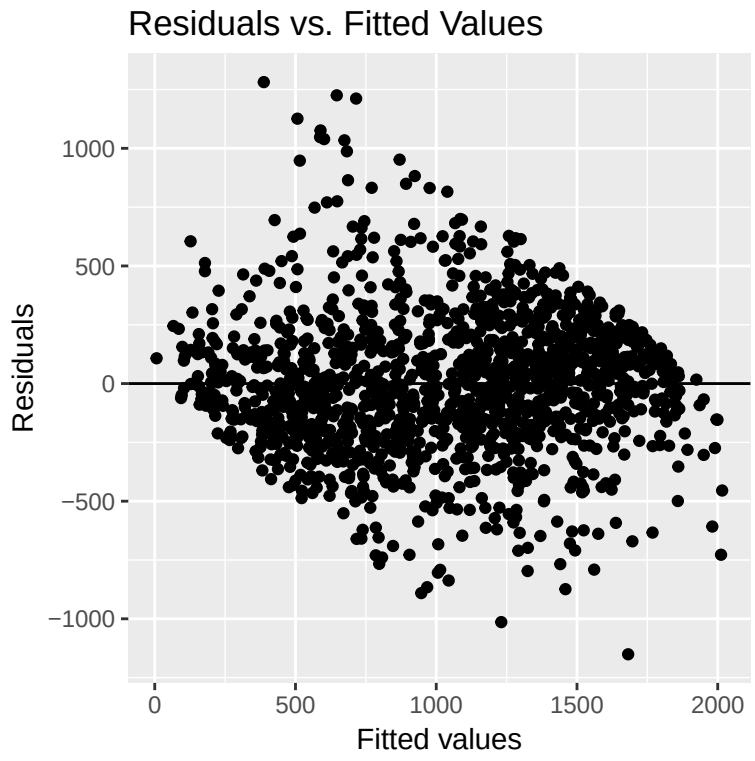
Furthermore, the QQ-plot also indicates heavy tails, particularly on the left side of the distribution of residuals; this means that our assumption of a normal distribution of residuals is not violated but that the distribution has a little more variance than usual. We could assume that the residuals follow a t-distribution.

In essence, we have determined our preference for mod.base instead of mod.plus.

(ii) Residual plot for checking linearity

It appears that the linearity is satisfied by this original model, as the points are randomly dispersed around the horizontal line.

```
gf_point(.resid ~ .fitted, data = resid_mod_base) %>%
  gf_hline(yintercept = 0, col = "blue", lty = 2) %>%
  gf_labs(x = "Fitted values", y = "Residuals", title = "Residuals vs. Fitted Values")
```



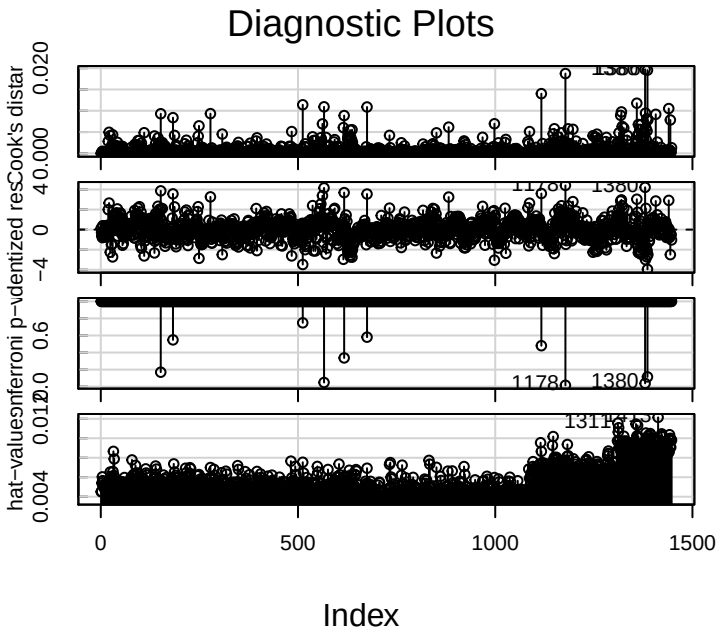
(iii) Independence

The location of any response variable in relation to its mean cannot be predicted from the knowledge of the explanatory variable.

(D) Model Diagnostics

(i) Check for influential points and high-leverages

```
library(car)
influenceIndexPlot(mod.base, id.n = 3)                    # row numbers of high 3 cases
```



The studentized residuals are too high for data points 63 (with an average birthweight of 2446.04 grams), 657 (with an average birthweight of 2515.94 grams) and 1122 (with an average birthweight of 2683.07 grams). Our reference number 4, since this is a large dataset.

Data point 1486 (with an average birthweight of 3191.61 grams)has a large hat value.

There appears to be no influential point, as there is no Cook's Distance that is close to 1.

The reference hat-value we use is $3((p+1)/n) = 0.018$.

After examining the aberrant data points, we still decided to keep the original model, as the Cook's Distance plot did not show any influential points.

(ii) Check for multicollinearity

Now we must check for multicollinearity:

```
vif(mod.base)
```

```
##                          GVIF Df GVIF^(1/(2*Df))
## Region                1.021370  4        1.002647
## DeliveryMethod        1.046111  1        1.022796
## TobaccoUse            1.020267  1        1.010083
## AverageAgeofMother    1.086192  1        1.042205
```

Multicollinearity is not suspected. All the VIF values are much smaller than 5.

Given these factors, we select mod.base and now need to make sure that interaction terms truly are not necessary.

(D) Investigate whether our model is sufficient.

(i) We included all the interaction terms in a new model called mod.one.

Looking at the ANOVA chi-square test, we can compare the two models.

```
mod.one <- lm(AverageBirthWeight ~ Region * DeliveryMethod * TobaccoUse * AverageAgeofMother, data = Na
anova(mod.one, mod.base, test = "Chisq")
```

```
## Analysis of Variance Table
##
## Model 1: AverageBirthWeight ~ Region * DeliveryMethod * TobaccoUse * AverageAgeofMother
## Model 2: AverageBirthWeight ~ Region + DeliveryMethod + TobaccoUse + AverageAgeofMother
##   Res.Df       RSS  Df Sum of Sq  Pr(>Chi)
## 1   1407 110968195
## 2   1439 124098522 -32 -13130327 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our low p-value indicates that interaction terms indeed do have a significant effect and might be included in our model.

In order to assess which interaction terms might be necessary in a hypothetical refined model of mod.one (mod.forwardstep),

(ii) We used StepAIC to determine an interaction-based model with the lowest AIC criterion using stepwise elimination.

```
mod.forwardstep <- stepAIC(mod.base, scope = list(lower = ~1, upper = ~Region + DeliveryMethod + Tobacc
```

```
## Start:  AIC=16495.18
## AverageBirthWeight ~ Region + DeliveryMethod + TobaccoUse + AverageAgeofMother
##
##                                     Df Sum of Sq       RSS   AIC
## + DeliveryMethod:TobaccoUse          1   2139964 121958558 16477
## + Region:TobaccoUse                  4   3223766 120874756 16486
## + Region:AverageAgeofMother          4   3223417 120875105 16486
## + TobaccoUse:AverageAgeofMother      1   1316499 122782023 16487
## <none>                                           124098522 16495
## + DeliveryMethod:AverageAgeofMother  1    223732 123874790 16500
## + Region:DeliveryMethod              4    668374 123430148 16517
## - Region                             4  20077597 144176119 16683
## - AverageAgeofMother                 1  36134725 160233247 16858
## - DeliveryMethod                     1  53141290 177239812 17004
## - TobaccoUse                         1 170337315 294435837 17738
##
## Step:  AIC=16477.29
## AverageBirthWeight ~ Region + DeliveryMethod + TobaccoUse + AverageAgeofMother +
##     DeliveryMethod:TobaccoUse
##
##                                     Df Sum of Sq       RSS   AIC
## + Region:TobaccoUse                  4   3271337 118687221 16467
## + Region:AverageAgeofMother          4   3179903 118778655 16468
## + TobaccoUse:AverageAgeofMother      1    769141 121189417 16475
## <none>                                           121958558 16477
```

```
## + DeliveryMethod:AverageAgeofMother  1     71181 121887377 16484
## - DeliveryMethod:TobaccoUse          1   2139964 124098522 16495
## + Region:DeliveryMethod              4    642669 121315889 16499
## - Region                             4  20087604 142046162 16669
## - AverageAgeofMother                 1  36808948 158767506 16852
##
## Step:  AIC=16467.05
## AverageBirthWeight ~ Region + DeliveryMethod + TobaccoUse + AverageAgeofMother +
##     DeliveryMethod:TobaccoUse + Region:TobaccoUse
##
##                                     Df Sum of Sq       RSS   AIC
## + Region:AverageAgeofMother          4   2566676 116120545 16465
## <none>                                           118687221 16467
## + TobaccoUse:AverageAgeofMother      1    562931 118124289 16468
## + DeliveryMethod:AverageAgeofMother  1     55110 118632111 16474
## - Region:TobaccoUse                  4   3271337 121958558 16477
## - DeliveryMethod:TobaccoUse          1   2187535 120874756 16486
## + Region:DeliveryMethod              4    627906 118059315 16489
## - AverageAgeofMother                 1  37267276 155954497 16855
##
## Step:  AIC=16464.53
## AverageBirthWeight ~ Region + DeliveryMethod + TobaccoUse + AverageAgeofMother +
##     DeliveryMethod:TobaccoUse + Region:TobaccoUse + Region:AverageAgeofMother
##
##                                     Df Sum of Sq       RSS   AIC
## <none>                                           116120545 16465
## + TobaccoUse:AverageAgeofMother      1    513824 115606720 16465
## - Region:AverageAgeofMother          4   2566676 118687221 16467
## - Region:TobaccoUse                  4   2658110 118778655 16468
## + DeliveryMethod:AverageAgeofMother  1     56736 116063809 16471
## - DeliveryMethod:TobaccoUse          1   2147148 118267693 16484
## + Region:DeliveryMethod              4    380756 115739788 16489
```

(iii) This is our refined model: mod.two.

```
mod.two <- lm(AverageBirthWeight ~ Region + DeliveryMethod + TobaccoUse + AverageAgeofMother + TobaccoU
```

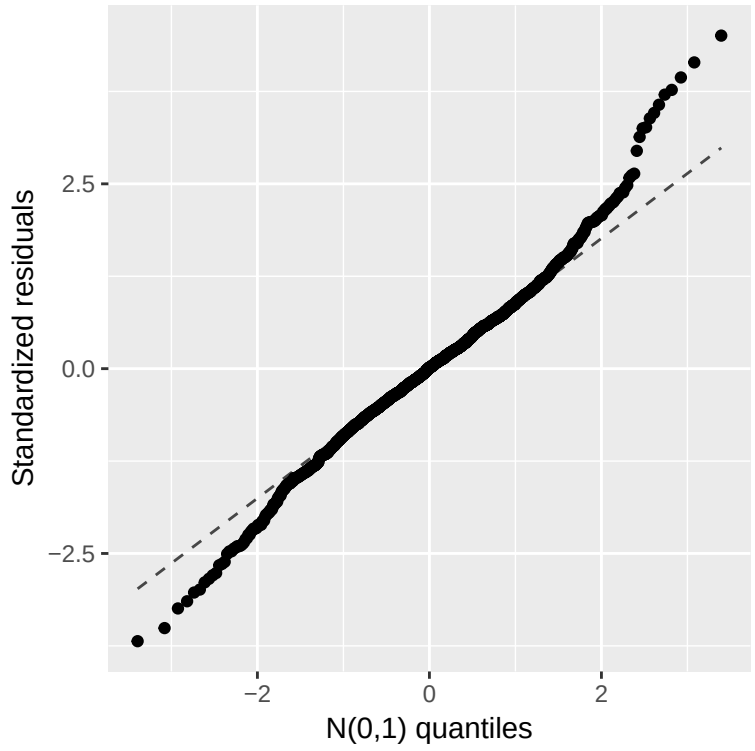
(E) Model Assumptions + Diagnostics for the Refined Model: mod.two.:

```
library(broom)
resid_mod_two <- augment(mod.two)
head(resid_mod_two)
```

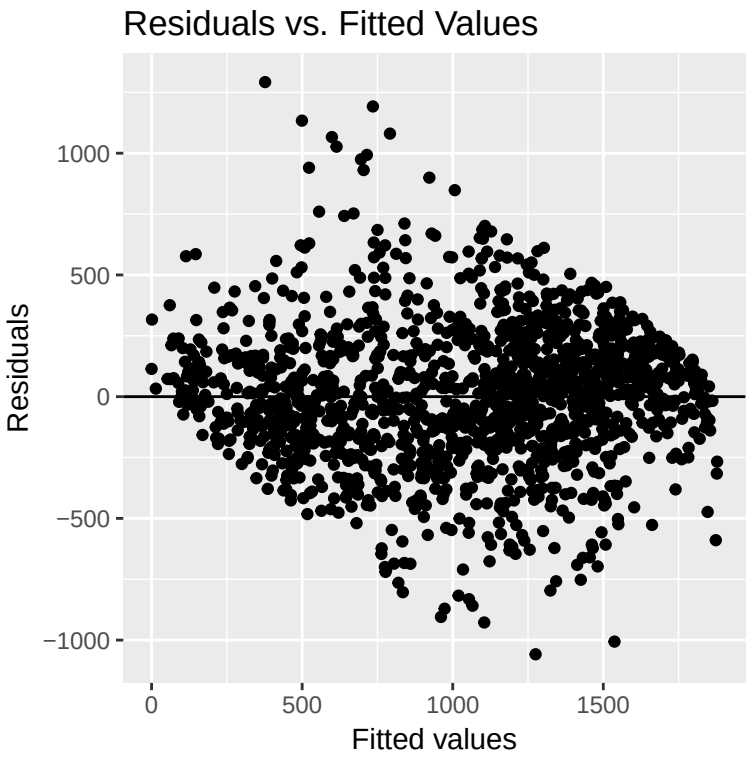
```
##   AverageBirthWeight    Region DeliveryMethod TobaccoUse
## 1               1607 NorthEast        Vaginal         No
## 2               1187 NorthEast       Cesarean         No
## 3                511 NorthEast        Vaginal        Yes
## 4               1097 NorthEast        Vaginal         No
## 5                461 NorthEast       Cesarean        Yes
## 6               1017 NorthEast       Cesarean         No
##   AverageAgeofMother   .fitted  .se.fit     .resid        .hat    .sigma
```

```
## 1              29.52 1584.7320 23.15276   22.26804 0.006415649 289.1566
## 2              31.25 1277.7152 23.27532  -90.71523 0.006483752 289.1472
## 3              26.33  638.4012 29.03881 -127.40117 0.010092358 289.1374
## 4              25.26 1379.3289 27.80373 -282.32892 0.009252114 289.0601
## 5              27.19  305.7598 28.46766  155.24018 0.009699258 289.1278
## 6              28.01 1121.4932 24.78499 -104.49320 0.007352118 289.1439
##        .cooksd  .std.resid
## 1 2.966788e-06  0.07728532
## 2 4.976550e-05 -0.31485469
## 3 1.539011e-04 -0.44298960
## 4 6.916987e-04 -0.98127614
## 5 2.194341e-04  0.53968213
## 6 7.500489e-05 -0.36283389
```

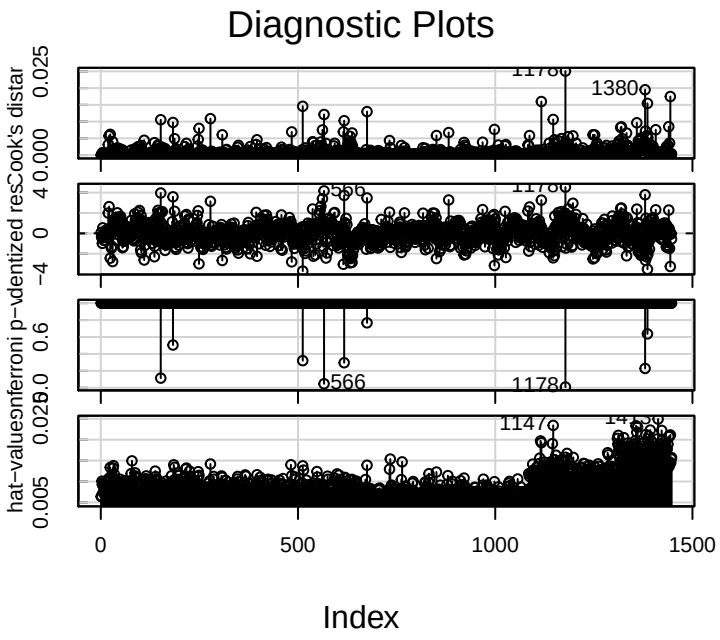
```
gf_qq(~.std.resid, data = resid_mod_two) %>%
 gf_qqline() %>%
 gf_labs(x = "N(0,1) quantiles", y = "Standardized residuals" )
```




```
gf_point(.resid ~ .fitted, data = resid_mod_two) %>%
  gf_hline(yintercept = 0, col = "blue", lty = 2) %>%
  gf_labs(x = "Fitted values", y = "Residuals", title = "Residuals vs. Fitted Values")
```




```
library(car)
influenceIndexPlot(mod.two, id.n = 3)                    # row numbers of high 3 cases
```



```
vif(mod.two)
```

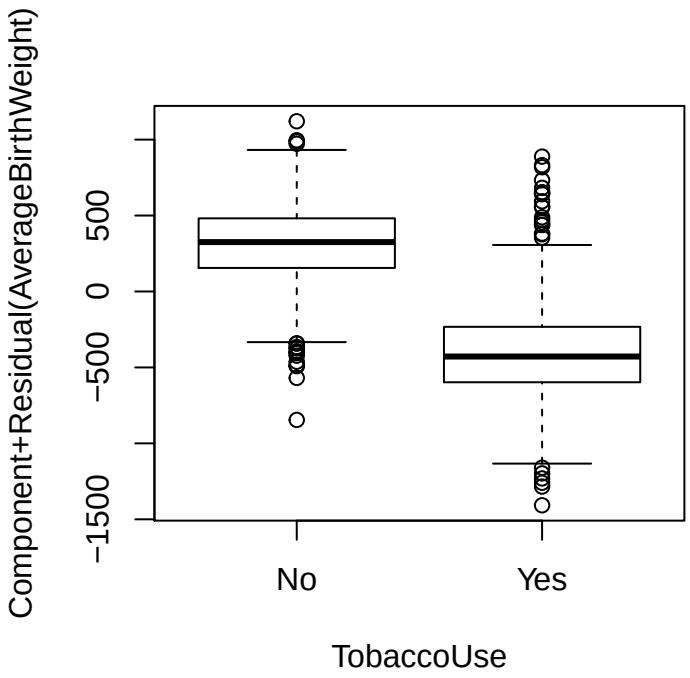
```
##                                     GVIF Df GVIF^(1/(2*Df))
## Region                          9.813612  4        1.330389
## DeliveryMethod                  1.046920  1        1.023191
## TobaccoUse                    110.926938  1       10.532186
## AverageAgeofMother              1.785413  1        1.336193
## TobaccoUse:AverageAgeofMother 101.828735  1       10.091022
## Region:TobaccoUse              31.777143  4        1.540864
```

On the basis of high variance inflation factors for the coefficients of all of the interaction terms that remain after the stepAIC elimination process, we must reject the notion that interaction terms are needed in our model. Therefore, our model (mod.base) describes the average birth weight of an infant as a linear combination of TobaccoUse, Region, DeliveryMethod, and AverageAgeofMother. We believe that these are all significant and linear predictors of birth weight.

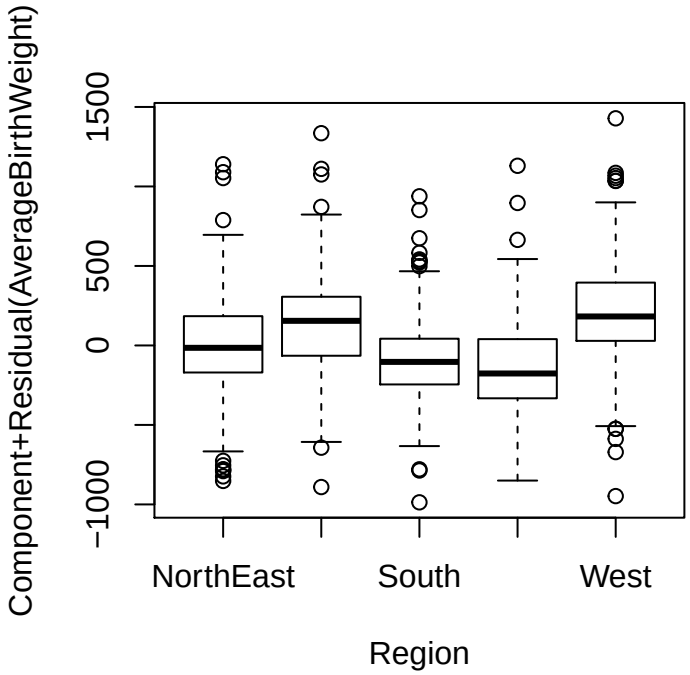
Further proof of linearity for our chosen model:

In order to assess whether our variables are linearly correlated average birth weight, we have created the following (component +) partial residual plots for our chosen model.

```
crPlot(mod.base, variable = "TobaccoUse")
```



```
crPlot(mod.base, variable = "Region")
```




```
crPlot(mod.base, variable = "DeliveryMethod")
```

```
crPlot(mod.base, variable = "AverageAgeofMother")
```