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",

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

# 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.

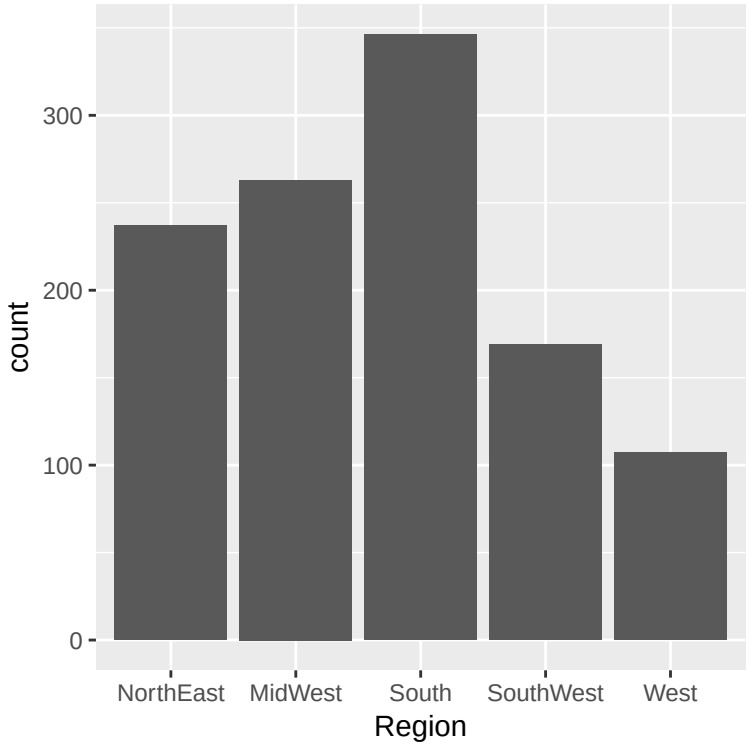
We created two new datasets to test and train our model.

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

(B) The Data Analysis

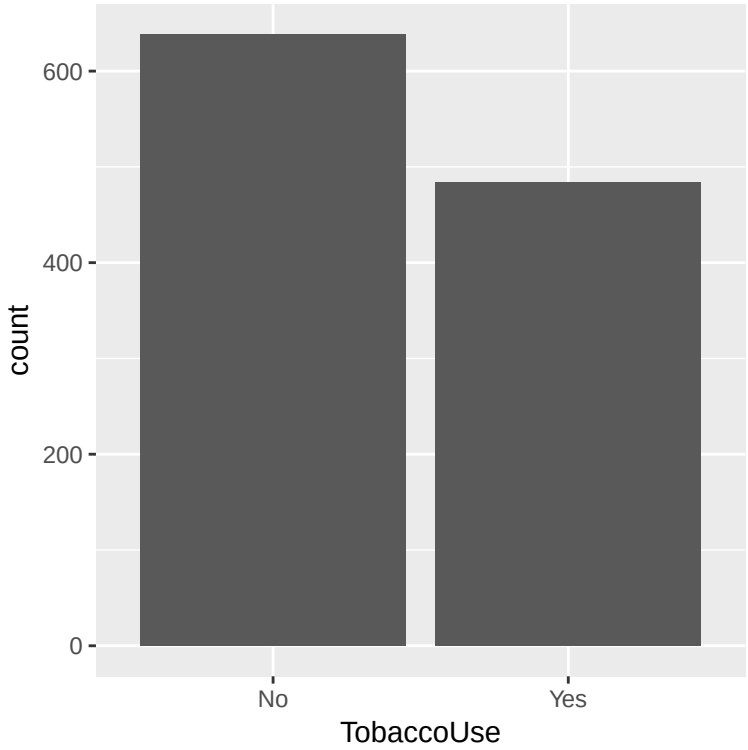
(i) barplots

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



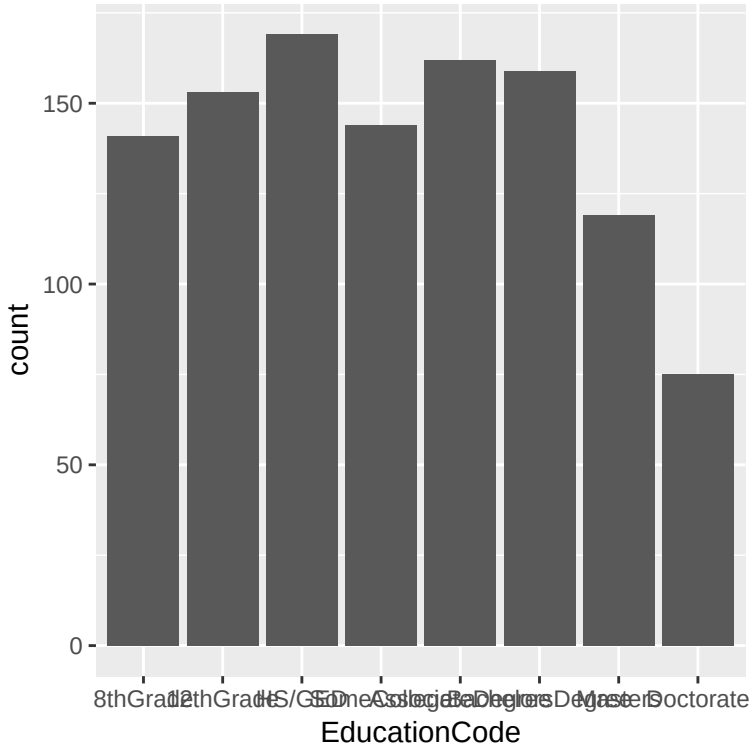
The above barplot indicates that there are the most entries from the South and the fewest numer of entries are from the West.

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



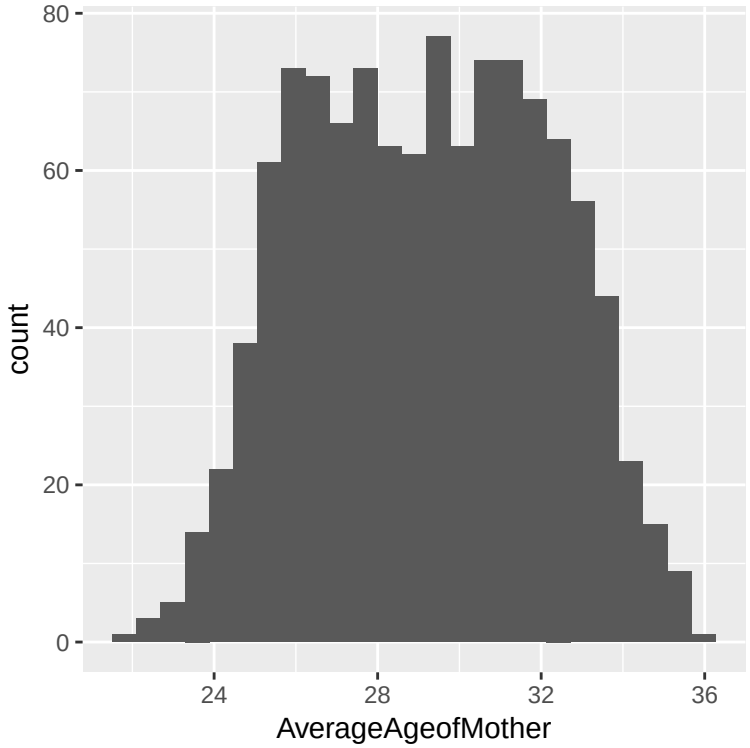
The above barplot indicates that proportionally less people use tobacco in our dataset.

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



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

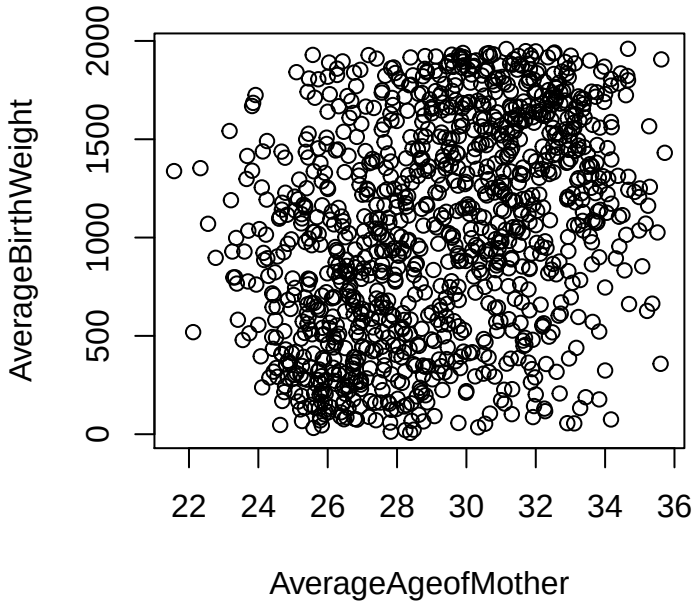
```
gf_histogram(~AverageAgeofMother, data = train_Natality)
```



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

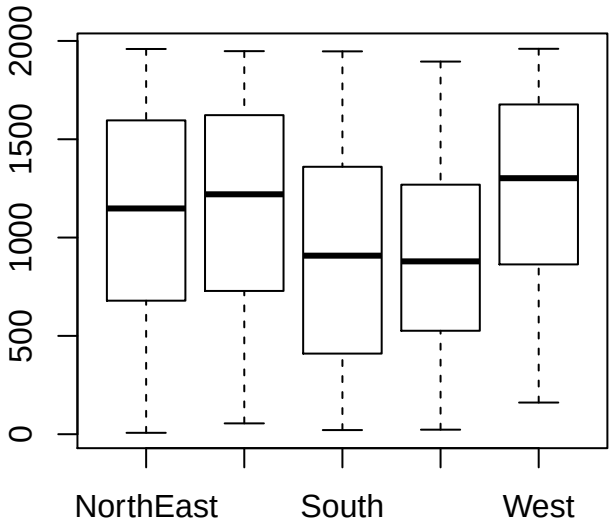
```
plot(AverageBirthWeight ~ AverageAgeofMother , data = train_Natality)
```



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)
```



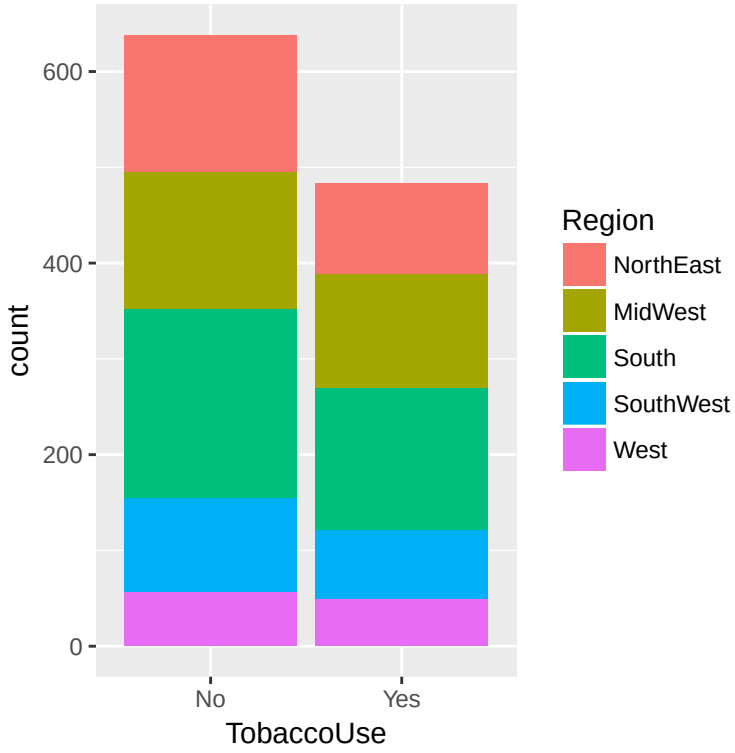
This boxplot allows us to see that the distribution of birth weight is different for each region; thus, region is an important variable to include in our model.

(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
## -1148.70  -178.46    -3.58   175.51  1278.28
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -432.083     88.630  -4.875 1.20e-06 ***
## RegionMidWest             104.540     24.017   4.353 1.44e-05 ***
## RegionSouth              -123.317     22.607  -5.455 5.73e-08 ***
```

```
## RegionSouthWest          -172.240     27.301  -6.309 3.70e-10 ***
## RegionWest                184.200     30.992   5.944 3.47e-09 ***
## DeliveryMethodNot Stated  120.196     46.603   2.579     0.01 *
## DeliveryMethodVaginal     391.160     16.549  23.637  < 2e-16 ***
## TobaccoUseYes            -699.047     16.478 -42.424  < 2e-16 ***
## AverageAgeofMother         54.699      2.808  19.479  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 307.8 on 1487 degrees of freedom
## Multiple R-squared:  0.6769, Adjusted R-squared:  0.6752
## F-statistic: 389.4 on 8 and 1487 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)
```

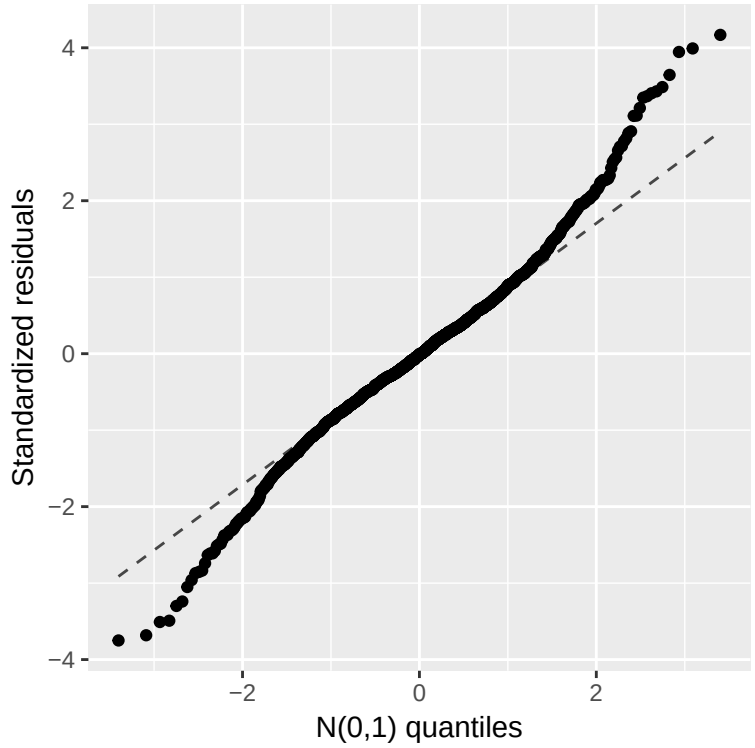
```
##   AverageBirthWeight    Region DeliveryMethod TobaccoUse
## 1               1607 NorthEast        Vaginal         No
## 2               1187 NorthEast       Cesarean         No
## 3                511 NorthEast        Vaginal        Yes
## 4               1097 NorthEast        Vaginal         No
## 5                461 NorthEast       Cesarean        Yes
## 6               1017 NorthEast       Cesarean         No
##   AverageAgeofMother   .fitted  .se.fit     .resid        .hat   .sigma
## 1              29.52 1573.7851 20.45240   33.21487 0.004414004 307.9443
## 2              31.25 1277.2535 20.48133  -90.25351 0.004426501 307.9366
## 3              26.33  700.2492 22.39051 -189.24917 0.005290201 307.9062
## 4              25.26 1340.7684 23.65519 -243.76844 0.005904693 307.8802
## 5              27.19  356.1296 22.88321  104.87036 0.005525582 307.9335
## 6              28.01 1100.0295 21.79031  -83.02954 0.005010384 307.9380
##        .cooksd .std.resid
## 1 5.760260e-06  0.1081348
## 2 4.265244e-05 -0.2938323
## 3 2.245170e-04 -0.6163934
## 4 4.162919e-04 -0.7942105
## 5 7.204406e-05  0.3416081
## 6 4.090732e-05 -0.2703931
```

(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.

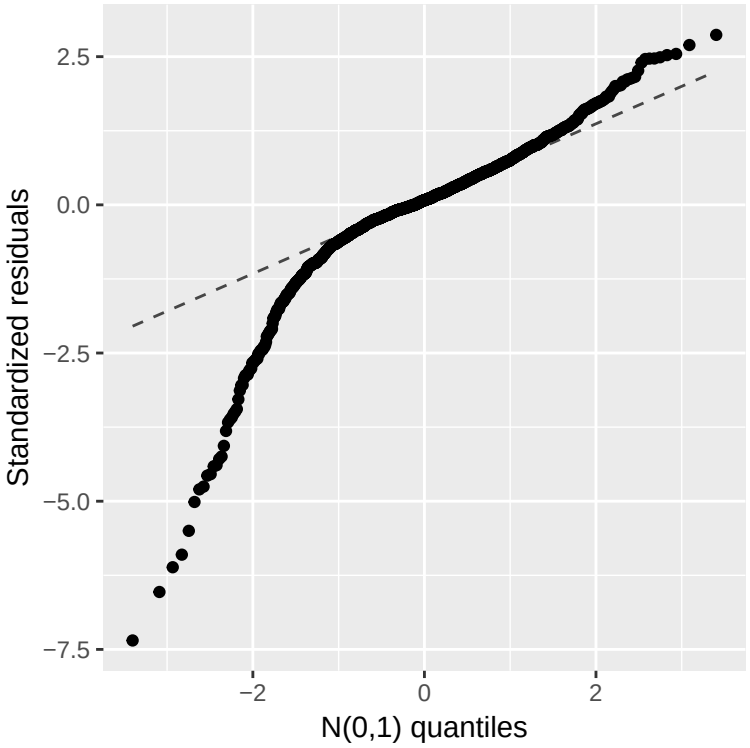
```
 gf_qq(~.std.resid, data = resid_mod_base) %>%
  gf_qqline() %>%
  gf_labs(x = "N(0,1) quantiles", y = "Standardized residuals" )
```




```
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
## 1              29.52 7.425865 0.03539966 -0.043740149 0.004414004
## 2              31.25 7.015687 0.03544974  0.063497341 0.004426501
## 3              26.33 6.240197 0.03875420 -0.003827497 0.005290201
## 4              25.26 7.139493 0.04094316 -0.139158681 0.005904693
## 5              27.19 5.771535 0.03960698  0.361862712 0.005525582
## 6              28.01 6.797883 0.03771536  0.126728922 0.005010384
##      .sigma      .cooksd  .std.resid
## 1 0.5330008 3.334476e-06 -0.08227315
## 2 0.5329994 7.047207e-06  0.11943625
## 3 0.5330020 3.065498e-08 -0.00720251
## 4 0.5329897 4.528495e-05 -0.26194708
## 5 0.5329188 2.863324e-04  0.68102694
## 6 0.5329918 3.181101e-05  0.23844254
```

```
  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.

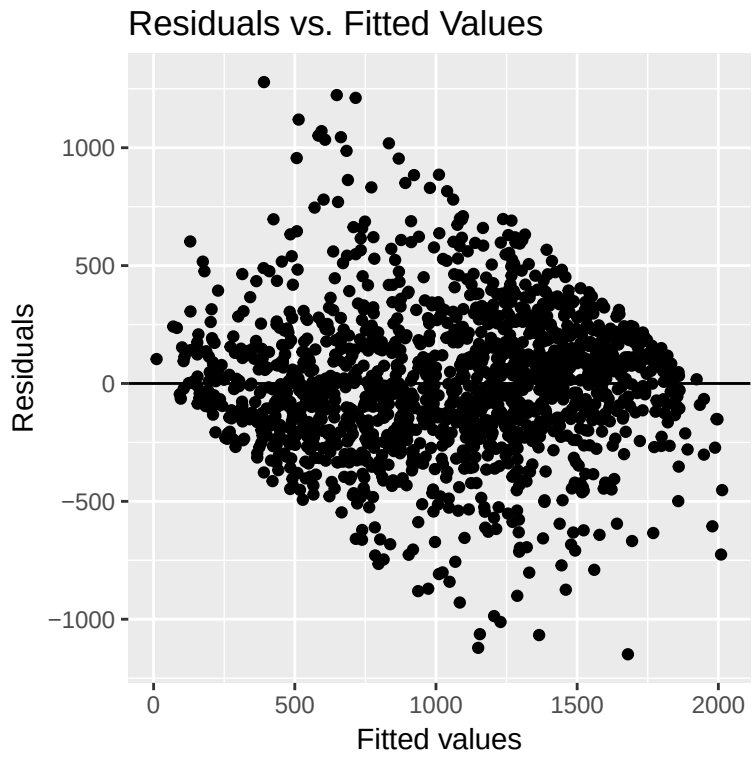
Furthermore, the QQ-plot also indicates heavy tails, particularly on the left side of the distribution of residuals; this means that our assumption of a normal distribution of residuals is not violated but that the distribution has a little more variance than usual. We could assume that the residuals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



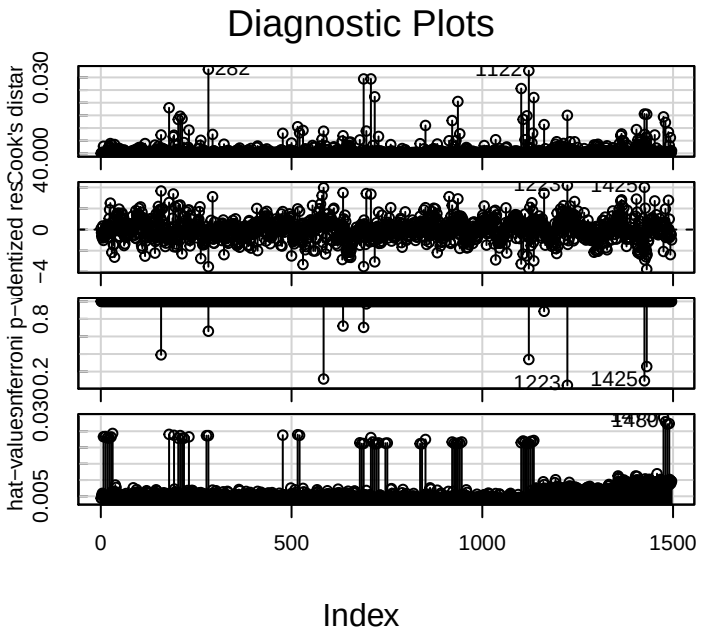
(iii) Independence

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               1.035228  4        1.004337
## DeliveryMethod       1.093301  2        1.022551
## TobaccoUse           1.045763  1        1.022625
## AverageAgeofMother 1.087144  1        1.042662
```

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   1448 126204382
## 2   1487 140918061 -39 -14713678 < 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=17199.69
## AverageBirthWeight ~ Region + DeliveryMethod + TobaccoUse + AverageAgeofMother
##
##                                     Df Sum of Sq       RSS   AIC
## + DeliveryMethod:TobaccoUse          1   2130649 138787411 17184
## + TobaccoUse:AverageAgeofMother      1   1416140 139501921 17192
## + Region:TobaccoUse                  4   3049122 137868938 17196
## + Region:AverageAgeofMother          4   3004545 137913516 17197
## <none>                                           140918061 17200
## + DeliveryMethod:AverageAgeofMother  2    393118 140524942 17210
## + Region:DeliveryMethod              7   1399669 139518391 17236
## - Region                             4  21493434 162411494 17383
## - AverageAgeofMother                 1  35958010 176876071 17532
## - DeliveryMethod                     2  53325844 194243905 17665
## - TobaccoUse                         1 170560606 311478667 18379
##
## Step:  AIC=17184.21
## AverageBirthWeight ~ Region + DeliveryMethod + TobaccoUse + AverageAgeofMother +
##     DeliveryMethod:TobaccoUse
##
##                                     Df Sum of Sq       RSS   AIC
## + Region:TobaccoUse                  4   3093781 135693630 17180
## + Region:AverageAgeofMother          4   2961020 135826391 17181
## + TobaccoUse:AverageAgeofMother      1    853067 137934344 17182
## <none>                                           138787411 17184
```

```
## + DeliveryMethod:AverageAgeofMother  2    246163 138541248 17196
## - DeliveryMethod:TobaccoUse          1   2130649 140918061 17200
## + Region:DeliveryMethod              7   1374903 137412508 17221
## - Region                             4  21502545 160289956 17371
## - AverageAgeofMother                 1  36623182 175410593 17527
##
## Step:  AIC=17179.73
## AverageBirthWeight ~ Region + DeliveryMethod + TobaccoUse + AverageAgeofMother +
##     DeliveryMethod:TobaccoUse + Region:TobaccoUse
##
##                                     Df Sum of Sq       RSS   AIC
## + TobaccoUse:AverageAgeofMother      1    674295 135019335 17180
## <none>                                           135693630 17180
## + Region:AverageAgeofMother          4   2453117 133240513 17182
## - Region:TobaccoUse                  4   3093781 138787411 17184
## + DeliveryMethod:AverageAgeofMother  2    256279 135437351 17192
## - DeliveryMethod:TobaccoUse          1   2175308 137868938 17196
## + Region:DeliveryMethod              7   1531516 134162114 17214
## - AverageAgeofMother                 1  36927081 172620711 17533
##
## Step:  AIC=17179.58
## AverageBirthWeight ~ Region + DeliveryMethod + TobaccoUse + AverageAgeofMother +
##     DeliveryMethod:TobaccoUse + Region:TobaccoUse + TobaccoUse:AverageAgeofMother
##
##                                     Df Sum of Sq       RSS   AIC
## <none>                                           135019335 17180
## - TobaccoUse:AverageAgeofMother      1    674295 135693630 17180
## + Region:AverageAgeofMother          4   2407555 132611780 17182
## - Region:TobaccoUse                  4   2915010 137934344 17182
## - DeliveryMethod:TobaccoUse          1   1650640 136669975 17191
## + DeliveryMethod:AverageAgeofMother  2    170018 134849316 17192
## + Region:DeliveryMethod              7   1512997 133506338 17214
```

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

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

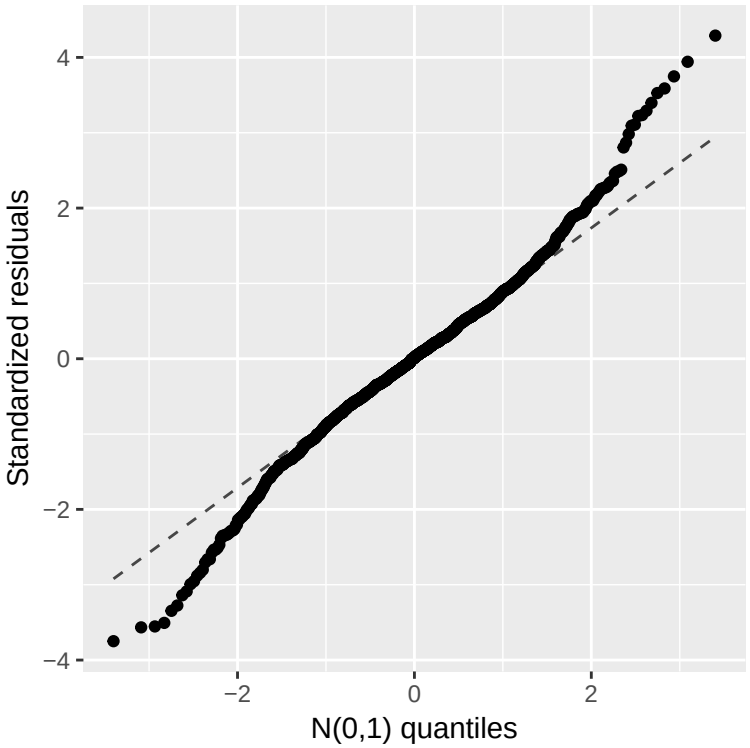
(E) Model Assumptions + Diagnostics for the Refined Model: mod.two.:

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

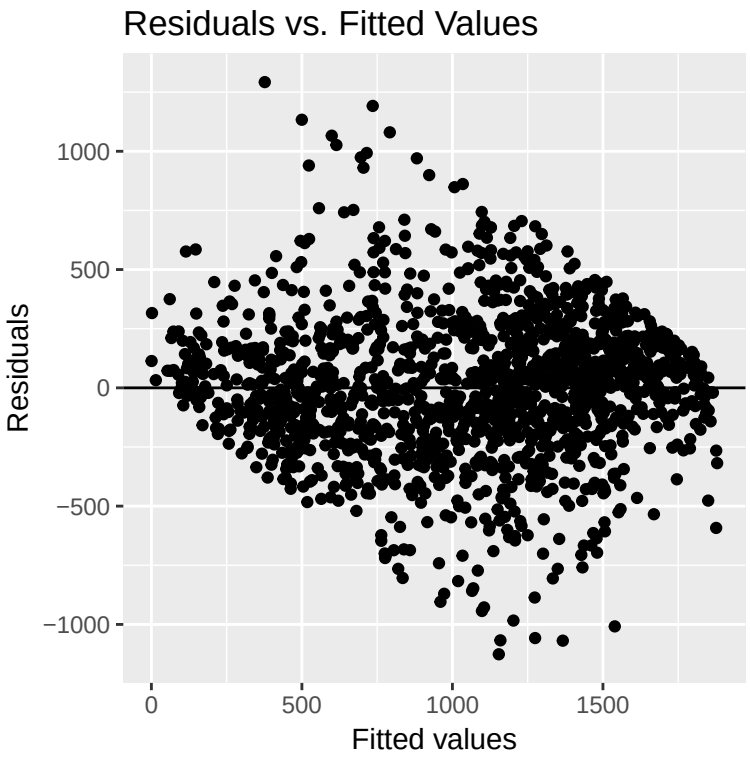
```
##   AverageBirthWeight    Region DeliveryMethod TobaccoUse
## 1               1607 NorthEast        Vaginal         No
## 2               1187 NorthEast       Cesarean         No
## 3                511 NorthEast        Vaginal        Yes
## 4               1097 NorthEast        Vaginal        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

```
## 1              29.52 1595.8789 23.76885   11.12105 0.006126206 303.7798
## 2              31.25 1287.8652 23.89935 -100.86518 0.006193657 303.7685
## 3              26.33  638.0293 30.50686 -127.02932 0.010091832 303.7618
## 4              25.26 1395.1956 28.57746 -298.19564 0.008855688 303.6802
## 5              27.19  306.2834 29.90600  154.71657 0.009698216 303.7530
## 6              28.01 1135.2328 25.45303 -118.23280 0.007025122 303.7642
##         .cooksd  .std.resid
## 1 5.941117e-07  0.03673397
## 2 4.941663e-05 -0.33317922
## 3 1.287168e-04 -0.42043035
## 4 6.208664e-04 -0.98632579
## 5 1.833485e-04  0.51196539
## 6 7.714369e-05 -0.39071168
```

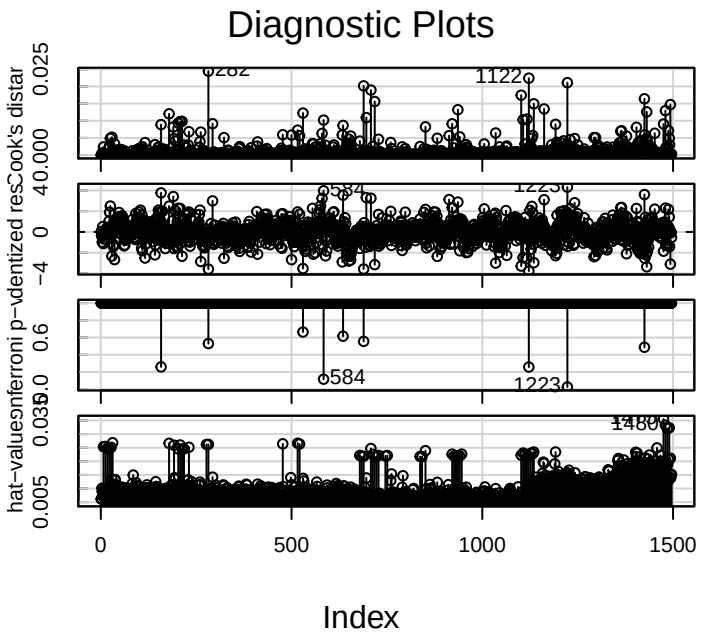
```
gf_qq(~.std.resid, data = resid_mod_two) %>%
 gf_qqline() %>%
 gf_labs(x = "N(0,1) quantiles", y = "Standardized residuals" )
```




```
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)
```

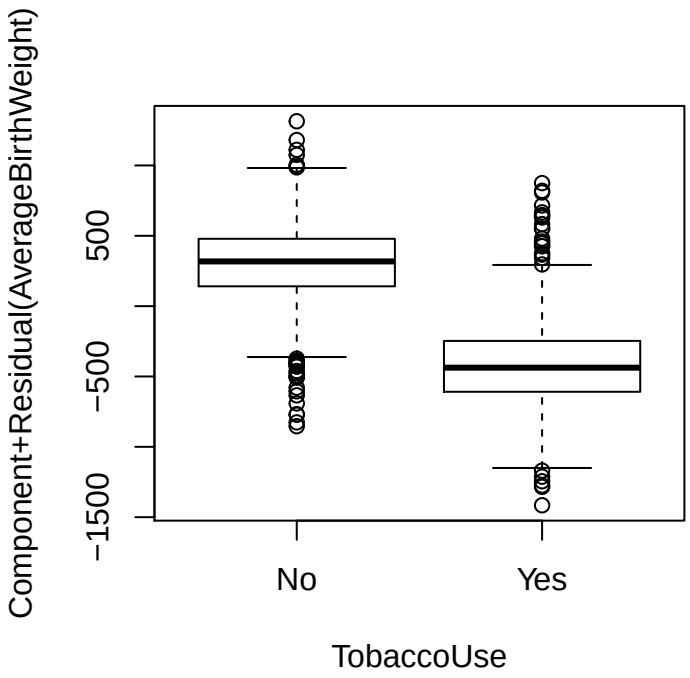
```
##                                     GVIF Df GVIF^(1/(2*Df))
## Region                          9.244888  4        1.320498
## DeliveryMethod                  1.108856  2        1.026169
## TobaccoUse                    112.090525  1       10.587281
## AverageAgeofMother              1.762498  1        1.327591
## TobaccoUse:AverageAgeofMother 102.966378  1       10.147235
## Region:TobaccoUse              30.055122  4        1.530170
```

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.

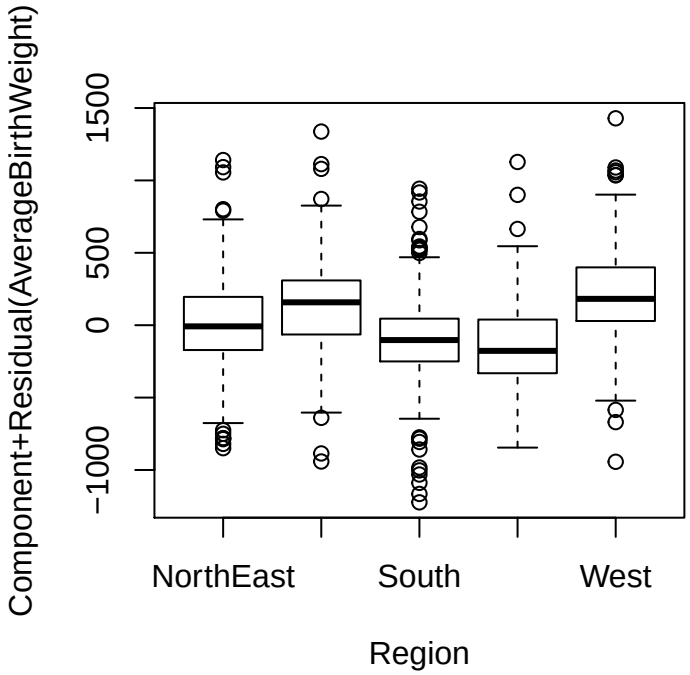
Further proof of linearity for our chosen model:

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.

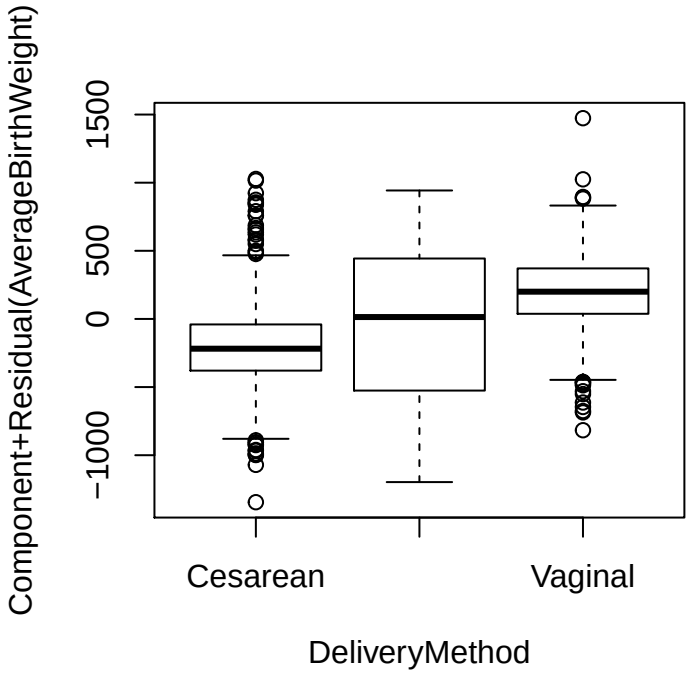
```
crPlot(mod.base, variable = "TobaccoUse")
```



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




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



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