MTH-245 Final project Part 2
Fall 2022

Name:

```
library("tidyverse")
library("xtable")
library("ggplot2")
library("patchwork")
library("bestglm")
library("EnvStats")
library("car")
library("GGally")
library("olsrr")
library("gridExtra")
library("boot")
source("https://cipolli.com/students/code/plotResiduals.R")
```

1 Part 1: Abstract

Background: According to The Washington Post, the average birth weight of American infants has dropped 453.592 grams between 1990 and 2013, making the average birth weight 3247.721 grams. While, this drop in weight may not seem significant, it brings us closer to an average low birth weight which is classified as 2,500 grams or less. Stanford University released a study on how low birth weights can impose health issues on children. Such issues include infection, breathing problems and immature lungs, nervous system problems, bleeding inside the brain, sudden infant death syndrome, and other long term complications such as cerebral palsy, blindness, deafness, developmental delay. Clearly these are extremely high risks and the same study from Stanford listed some social factors of the mothers that influence birth weight such as smoking, not gaining enough weight during pregnancy, African-American background, and the age of the mother being less than 17 or more than 35 years. Awareness of what social factors that influence low birth weights can help with prenatal guidance and care to avoid the potential risks listed above. The purpose of this study is to identify key predictors influencing low birth weights using a sample of infant birth weights and other information collected from North Carolina. **Methods:** We will use a linear regression to model the relationship between whether the mother was a smoker, weight gained by the mother, the mother's age, ... and the infant's birth weight. **Findings:** After making adjustments to the initial model such as transformations, centering, and interactions, we determined our final best model to predict which factors have the most influence on birth weight. Our final model had *an 83.6 increase in precision and a 7 increase in predictive ability as compared to the first order additive linear model*.

2 Part 2: Introduction

According to The World Health Organization, the average weight of a baby born at 37–40 weeks ranges from 5 lb 8 oz to 8 lb 13 oz. This is 2,500 grams to 4,000 grams. Birth weight is something that we don't typically consider when we are projecting the health of our future population, but it plays an extremely important role in influencing the expectancy, quality and health of a person's life. If an infant is born with a low birth weight, they could face immediate and long term health issues. If a child is inflicted with long term health issues, they will require medical care and resources for the rest of their lives. These considerations are important population-wise, because as the population grows and the birth weight continues to decrease, there may be a strain on medical care and some resources available to those with long term health issues. Those who work in the healthcare industry regarding women and children's health, especially Obstetricians, should be informed on what social factors and behaviors within the population strongly influence birth weight so that they can provide the correct medical care and advice for each patient, accordingly. Our data is called NCbirths and comes from the Stat2Data package in R datasets. It was collected by statistician John Holcomb at Cleveland State University, from the North Carolina State Center for Health and Environmental Statistics. NCbirths contains data from births in North Carolina in 2001, with 1450 observations on 15 variables that include social and behavioral characteristics of the mother. The response variable of our study was BirthWeightGM, which is the baby's birth weight in grams. We hypothesized that the following

variables would be the most predictive, after our background research using the Low Birth Weight study published by Stanford University: race of mother, gestation period (weeks), sex of the infant, whether just a single infant was delivered or more than one, if the mother smoked while pregnant, weight gained by mother, the mother's age. We hypothesize that gestation period will be very influential, but our study will determine which other variables are influential. The following code imports our data and alters the type of each variable. We also renamed the levels within our categorical variables, and treated them all as factors. We centered and scaled all of our quantitative variables and created more variables for each transformation conducted on the quantitaive variables.

```
prepData <- function() {
  births <- read_csv("~/GitHub/Mth245Final/dataset/NCbirths.csv")

  births <- births %>% mutate(Sex = case_when(Sex == 1 ~ "Male",
                                              Sex == 2 ~ "Female"),
                              Marital = case_when(Marital == 1 ~ "Married",
                                                  Marital == 2 ~ "Unmarried"),
                              RaceMom = case_when(RaceMom == 1 ~ "White",
                                                  RaceMom == 2 ~ "Black",
                                                  RaceMom == 3 ~ "Am. Indian",
                                                  RaceMom == 4 ~ "Chinese",
                                                  RaceMom == 5 ~ "Japanese",
                                                  RaceMom == 6 ~ "Hawaiian",
                                                  RaceMom == 7 ~ "Filipino",
                                                  RaceMom == 8 ~ "Other Asian / PI"),
                              Smoke = case_when(Smoke == 1 ~ "Yes",
                                                Smoke == 0 ~ "No"),
                              Premie = case_when(Premie == 1 ~ "Yes",
                                                 Premie == 0~ "No"))
  births$Sex <- as.factor(births$Sex)
  births$Marital <- as.factor(births$Marital)
  births$Premie <- as.factor(births$Premie)
  births$Smoke <- as.factor(births$Smoke)
  births$RaceMom <- as.factor(births$RaceMom)
  births$Plural <- as.factor(births$Plural)

  births <- births %>% mutate(MomAgeSC = scale(MomAge, center=T, scale=T),
                              MomAgeSq = MomAgeSC^2,
                              WeeksSC = scale(Weeks, center=T, scale=T),
                              WeeksSq = WeeksSC ^2,
                              GainedSC = scale(Gained, center=T, scale=T),
                              GainedSq = I(GainedSC^2))

  births <- births %>% filter(!is.na(GainedSC) & !is.na(Smoke))

  # Part 1: First-Order Model and Determinations of Necessary Transformations

  births <- births %>% mutate(WeightGmLog = log(BirthWeightGm))
  births <- births %>% mutate(WeightGmSqrt = BirthWeightGm^.5)
  births <- births %>% mutate(WeightGmS = BirthWeightGm^2)
  births <- births %>% mutate(WeightGmSLog = log(BirthWeightGm)^2)
  births <- births %>% mutate(WeightGmLogLog = log(log(BirthWeightGm)))
  births <- births %>% mutate(WeightGmLogSqr = log(BirthWeightGm^2))
  births <- births %>% mutate(WeightGmLogQuad = log(BirthWeightGm)^4)
  births <- births %>% mutate(WeightGmSqrtLog = log(BirthWeightGm)^.5)
  births <- births %>% mutate(WeightGmInverse = 1/(BirthWeightGm))
  births <- births %>% mutate(WeightGmSC = scale(BirthWeightGm, center=T, scale=T))
  births <- births %>% mutate(WeightLogSC = scale(WeightGmLog, center=T, scale=T))
```

```
  births <- births %>% mutate(Twin = (as.character(Plural) == "2"))
  births <- births %>% mutate(Triplet = (as.character(Plural) == "3"))
  births <- births %>% mutate(Filipino = (RaceMom == "Filipino"))
  births <- births %>% mutate(Black = (RaceMom == "Black"))

  births$Twin = as.factor(births$Twin)
  births$Triplet = as.factor(births$Triplet)
  births$Fllipino = as.factor(births$Filipino)
  births$Black = as.factor(births$Black)

  births

}

births <- prepData()
```

```
## Rows:  1450 Columns:  15
## -- Column specification --------------------------------------------------------
## Delimiter:  ","
## chr  (2):  HispMom, MomRace
## dbl (13):  ID, Plural, Sex, MomAge, Weeks, Marital, RaceMom, Gained, Smoke, B...
##
## 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.
```

3 Part 3: Exploratory Data Analysis

a. **Graphically summarize the variables in your dataset.**

First we visualized our quantitative variables, including our response variable.

And we specifically created a boxplot for the birth weights recorded for mothers who are Black, since our research from the Stanford study indicated that race is influential in birth weight.

```
violin.BirthWeightGm <- ggplot(births, aes(x=BirthWeightGm, y=""))+
  geom_violin(fill = "lightblue",
              trim = FALSE)+
  geom_boxplot(width = .3,
               fill = "white") +
  theme_bw()+
  xlab("Birth Weights")+
  ylab(" ")+
  ggtitle("Distribution of Birth Weights",
          subtitle = "NCBirths Data")


violin.GestationPeriod <- ggplot(births, aes(x=Weeks, y=""))+
  geom_violin(fill = "lightblue",
              trim = FALSE)+
  geom_boxplot(width = .3,
               fill = "white") +
  theme_bw()+
  xlab("Weeks")+
  ylab(" ")+
  ggtitle("Distribution of Gestation Period",
          subtitle = "NCBirths Data")
```

```
violin.MomAge <- ggplot(births, aes(x=MomAge, y=""))+
  geom_violin(fill = "lightblue",
              trim = FALSE)+
  geom_boxplot(width = .3,
               fill = "white") +
  theme_bw()+
  xlab("Age (years)")+
  ylab(" ")+
  ggtitle("Distribution of Mothers' Ages",
          subtitle = "NCBirths Data")

violin.MomAge + violin.BirthWeightGm + violin.GestationPeriod
```

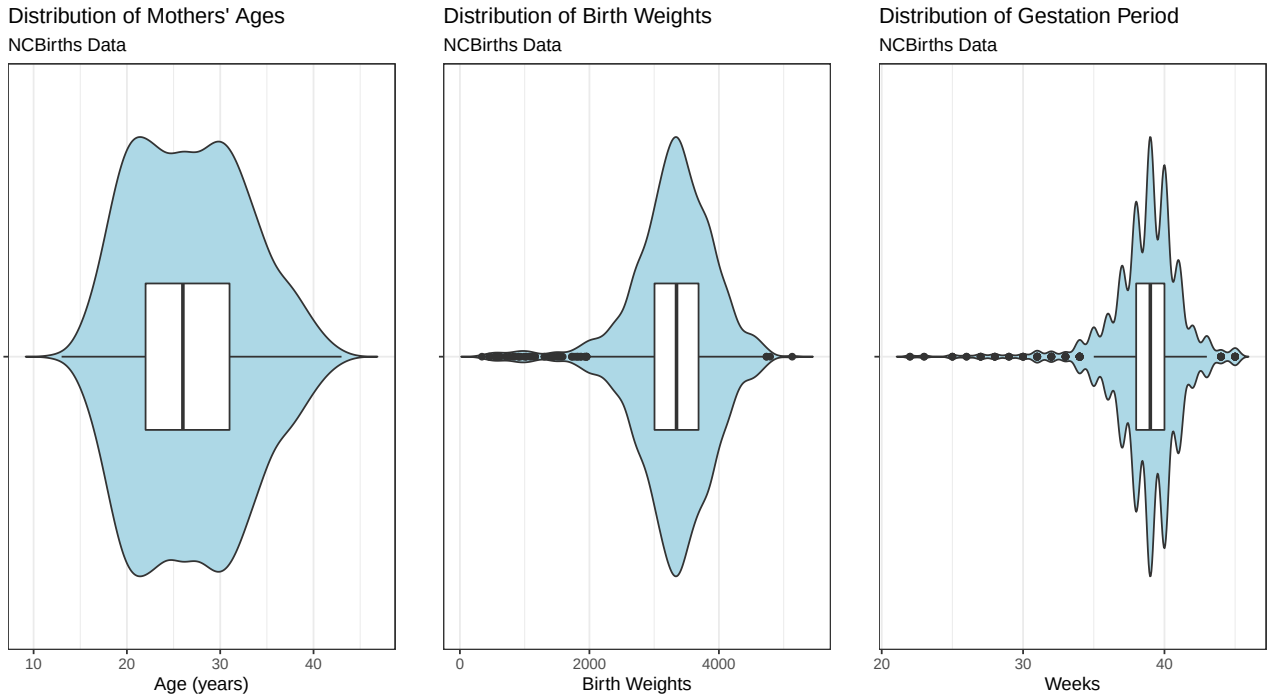


Figure 1: Violin plots of each variable.

the shape of the distribution of each variable any unusual looking observations

From 1 We can see that there is variability in almost all of the quantitative variables and they all have and many odd observations.

We also wanted to visualize the distributions of these variables.

```
histogram.BirthWeight<- ggplot(births, aes(x=BirthWeightGm))+
  geom_histogram(fill = "lightblue",
                 color = "black",
                 bins = 5) +
  theme_bw() +
  xlab("Birth Weights")+
  ylab("Count of Weight(gm)")+
  ggtitle("Frequencies of Birth Weights")
```

```
histogram.Gestation<- ggplot(births, aes(x=Weeks))+
  geom_histogram(fill = "lightblue",
                 color = "black",
                 bins = 5) +
  theme_bw() +
  xlab("Gestation Period")+
  ylab("Count of Weeks")+
  ggtitle("Frequencies of Gestation Periods")

histogram.MomAge <- ggplot(births, aes(x=MomAge))+
  geom_histogram(fill = "lightblue",
                 color = "black",
                 bins = 5) +
  theme_bw() +
  xlab("Ages of Mothers(years)")+
  ylab("Count of Ages")+
  ggtitle("Frequencies of Ages")

histogram.BirthWeight + histogram.Gestation + histogram.MomAge
```

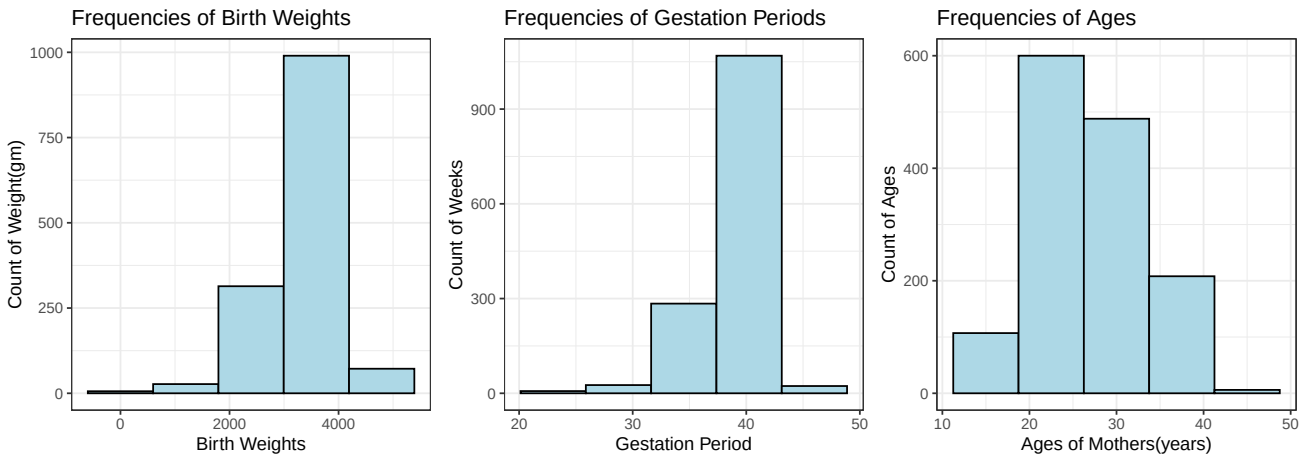


Figure 2: Grid of histograms for the quantitative variables.

2 shows that the quantitative variables do not follow normal distributions and are all skewed.

We expected the distribution for weights and weeks to be similar in shape because typically babies that are born prematurely have low birth weights. We did find it interesting though, that the distributions showed more preemies that we initially expected.

We weren't sure if the preemie weights in the data set were outlier instances that were heavily skewing the distribution, or if there were just more preemies that we expected. To further investigate this, we created a bootstrap confidence interval to see whether a baby born with a median weight falls within the interval.

```
## Bootstrapping for median weights
median(births$BirthWeightGm)
```

```
## [1] 3345.3
```

```
set.seed(23)
alpha <- 0.05
n <- nrow(births)
R <- 10000
```

```
boot.stats <- rep(NA, R)
for (i in 1:R){
  boot.data <- sample(x = births$BirthWeightGm, size = n, replace = TRUE)
  boot.stats[i] <- median(boot.data)
}

quantile(boot.stats, probs = c(alpha/2, 1 - alpha/2))
```

```
##    2.5%   97.5%
## 3316.95 3373.65
```

```
samp.boot.med <- function(data, indicies){
  median(data[indicies])
}

boot.medians <- boot(data = births$BirthWeightGm, statistic = samp.boot.med, R = 10000)
boot.ci(boot.medians, conf = 0.95)
```

```
## Warning in boot.ci(boot.medians, conf = 0.95):  bootstrap variances needed for studentized
intervals
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot.medians, conf = 0.95)
##
## Intervals :
## Level      Normal              Basic
## 95%   (3314, 3381 )   (3317, 3374 )
##
## Level     Percentile            BCa
## 95%   (3317, 3374 )   (3289, 3345 )
## Calculations and Intervals on Original Scale
## Some BCa intervals may be unstable
```

The median birth weight of our dataset was 3,345.3 grams. Our 95% percentile confidence interval was 3316.95 g. to 3373.65 g. This range contained our median birth weight of 3,345.3 grams, but more than that, our confidence interval was extremely close to the reported statistics from the World Health Organization of median birth weight of 3.3 kg (The WHO did not include any additional significant figures, so we assume that this could range from values anywhere to 3.25 kg to 3.35 kg.). This seems to lend support to our decision to assume that the collected data was representative of the population, despite the lack of specific knowledge of the collection techniques. Further, it indicates to us that we have a reasonable sample even with its non-normal distribution, that was caused by its skew from data primarily relating to babies born prematurely.

b. **Numerically summarize the variables in your dataset.**

```
#add in weeks
(sumstats <- births %>% summarize(meanW=mean(BirthWeightGm),
                                  medianW = median(BirthWeightGm),
                                   varianceW=var(BirthWeightGm),
                                  meanA=mean(MomAge),
                                  medianA=median(MomAge),
                                  varianceA = var(MomAge),
                                  meanWeeks = mean(Weeks),
                                  medianWeeks = median(Weeks),
                                  varianceWeeks = var(Weeks)))
```

```
## # A tibble: 1 x 9
##   meanW medianW varianceW meanA medianA varianceA meanWeeks medianWeeks varian~1
```

```
##   <dbl>   <dbl>     <dbl> <dbl>   <dbl>     <dbl>     <dbl>       <dbl>    <dbl>
## 1 3301.   3345.   393454.  26.8      26      37.1      38.6          39     7.04
## # ... with abbreviated variable name 1: varianceWeeks
```

Tableu visual

c. **Create a scatterplot matrix and table of correlations.**

```
library(GGally)
#ADD IN CORRECT VARIABLE NAMES
correlationsmatrix <- ggpairs(births, columns = c())

correlationsmatrix
```

```
## Error in data.frame(plotType = plotType, xVar = xVar, yVar = yVar, posX = posX, :
                  arguments imply differing number of rows:  0, 1
     ## Error in eval(expr, envir, enclos):  object 'correlationsmatrix' not found
```

Figure 3: Matrix of ScatterPlots and Correlations for the variables.

d. **Other interesting plots.**

Plot significant correlations - weeks and birth weight

e. **Comment on...**

the shape of the distribution of each variable

the relationship between the response and the quantitative predictors

any unusual looking observations

any other interesting takeaways

4 Part 4: First-Order Model and Model Selection.

First we created some functions that...

```
#create functions
#Calculates the residuals departure from the theoretical quantiles, returning the mean abs() value of the
quantDepart <- function(model) {
  residuals <- sort(scale(model$residuals, scale=T))
  i <- 1:length(residuals)
  fi <- (i - 0.5) / length(residuals)
  x.norm <- qnorm(fi)

  mean(abs(residuals - x.norm))
}

#Displays a variety of summary stats and graphics of the model tailored based off of the specific options
modelSummary <- function(model, coef=T, stat=T, plot=T){
  if(coef) {
    print(round(summary(model)$coefficients,10))
  }
  if(stat) {
    print(paste("R-squared:", summary(model)$r.squared))
    print(paste("Adjusted R-Squared:", summary(model)$adj.r.squared))
    print(paste("Sigma:", summary(model)$sigma))
    print(paste("AIC:", AIC(model)))
```

```
    print(paste("BIC:", BIC(model)))
    print(paste("Quantile Departure:", quantDepart(model)))
  }
  if(plot) {
    plotResiduals(model)
  }
}

#Calculates the  R squared of predicted vs actual values
r_squared <- function(actual, predicted) {
  cor(actual, predicted)^2
}

#If provided with a model and a number of subsets of data, this will generate summary statistics for those
predictForSubsets <- function(model, class.attr, ...) {
  x <- list(...)
  i <- 1
  predPlots = list()
  residPlots = list()
  for (v in x) {
    v$predict <- predict(model, v)
    v$resid <- v[[class.attr]] - v$predict


    print(paste("Subset", i, "R-Squared:", r_squared(v$predict, v[[class.attr]])))
    print(paste("Subset", i, "Mean Abs. Error:", mean(abs(v$resid))))

    ggplot(data=v, aes(x=predict, y=resid)) +
      geom_point(size=1,
                 shape=16)+
      theme_bw()+
      xlab("Predicted")+
      ylab("Residuals")+
      ggtitle("Predicted versus Residuals") -> residuals
    ggplot(data=v, aes(x=predict, y=get(class.attr))) +
      geom_point(size=1,
                 shape=16)+
      theme_bw()+
      xlab("Predicted")+
      ylab("Actual")+
      ggtitle("Predicted versus Actual") -> predictions
    print(predictions + residuals)
    i <- i + 1
  }
}
```

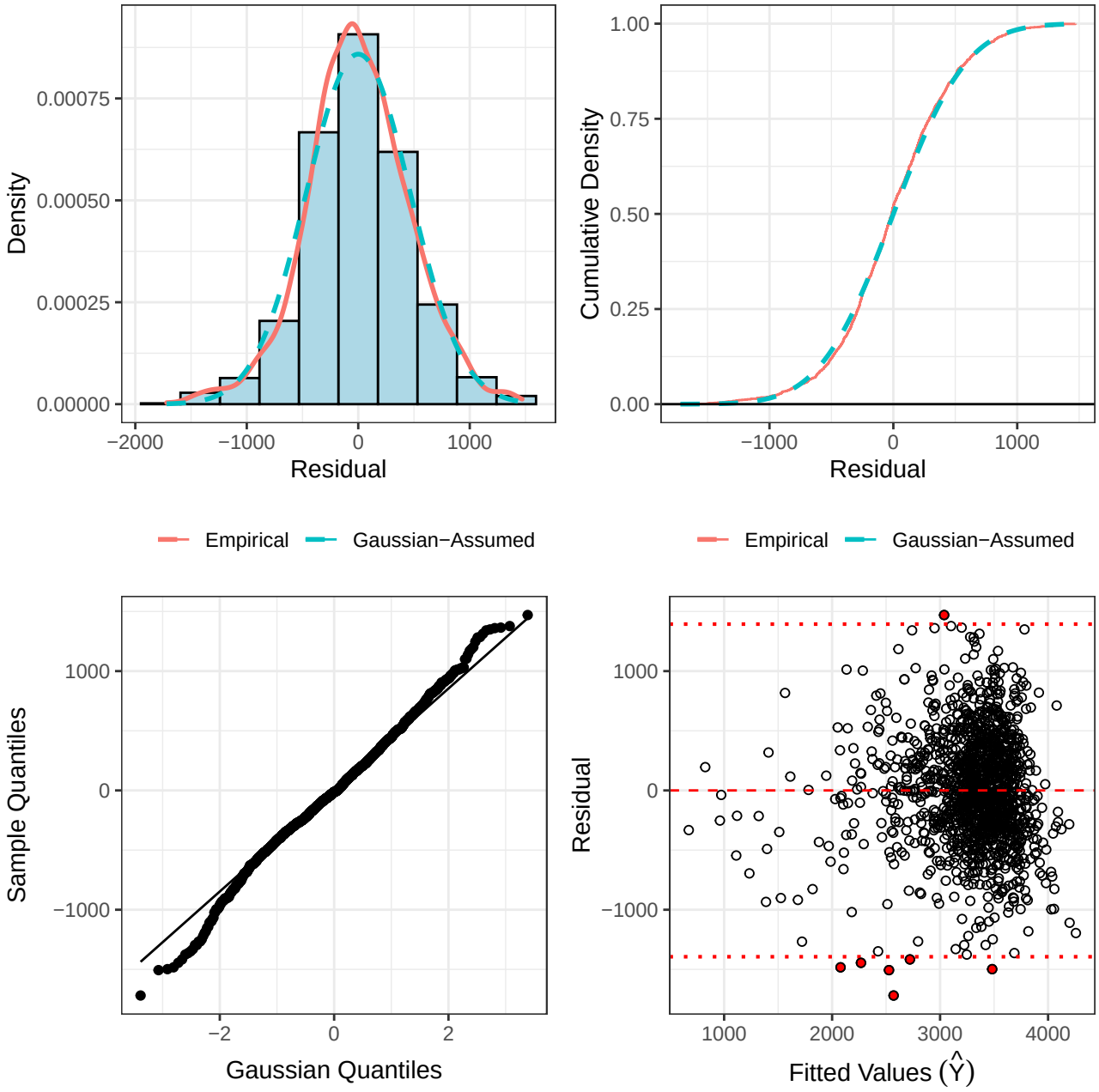
4.1 First Order Model

First, we fit a first-order linear model with all of our predictors. The estimated linear regression equation is

$$\hat{y} = -964.29 + -704.71 \cdot \text{I}(\text{Plural} = 2) + -932.13 \cdot \text{I}(\text{Plural} = 3) + 93.90 \cdot \text{I}(\text{Sex} = \text{Male}) + 10.88 \cdot \text{MomAge} + 97.81 \cdot \text{Weeks} - 76.88 \cdot \text{I}(\text{RaceMom} = \text{Black}) + 81.31 \cdot \text{I}(\text{RaceMom} = \text{Chinese}) - 860.40 \cdot \text{I}(\text{RaceMom} = \text{Filipino}) + 29.04 \cdot \text{I}(\text{RaceMom} = \text{Japanese}) - 58.43 \cdot \text{I}(\text{RaceMom} = \text{Other Asian/PI}) + 30.78 \cdot \text{I}(\text{RaceMom} = \text{White}) - 48.07 \cdot \text{I}(\text{Marital} = \text{Unmarried}) + 7.87 \cdot \text{Gained} - 203.49 \cdot \text{I}(\text{Smoke} = \text{Yes}) - 217.62 \cdot \text{I}(\text{Premie} = \text{Yes})$$

```r
#first model = model.1
lm(BirthWeightGm ~ Plural + Sex + MomAge + Weeks + RaceMom +
     Marital + Gained + Smoke + Premie, births) -> model.1

modelSummary(model.1, coef = TRUE)
```

```
##                           Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)             -964.28919 294.9607224 -3.2692122 0.0011048007
## Plural2                 -704.70670  76.5727750 -9.2030973 0.0000000000
## Plural3                 -932.12784 237.3884857 -3.9265925 0.0000903705
## SexMale                   93.89649  24.8249532  3.7823432 0.0001619382
## MomAge                    10.88077   2.3232147  4.6834980 0.0000030955
## Weeks                     97.81144   6.8023799 14.3790033 0.0000000000
## RaceMomBlack             -76.87702 102.7336877 -0.7483137 0.4543973043
## RaceMomChinese            81.30745 344.1690322  0.2362428 0.8132790072
## RaceMomFilipino         -860.40322 476.1359772 -1.8070536 0.0709696081
## RaceMomJapanese           29.03632 106.7611068  0.2719747 0.7856817781
## RaceMomOther Asian / PI  -58.43107 140.9347082 -0.4145967 0.6785009952
## RaceMomWhite              30.78400 101.2713293  0.3039755 0.7611919583
## MaritalUnmarried         -48.07301  32.1293957 -1.4962315 0.1348198655
## Gained                     7.86655   0.9098714  8.6457814 0.0000000000
## SmokeYes                -203.48734  36.5920140 -5.5609767 0.0000000321
## PremieYes               -217.62136  53.3527133 -4.0789183 0.0000478089
## [1] "R-squared: 0.45706523101896"
## [1] "Adjusted R-Squared: 0.451218840828927"
## [1] "Sigma: 464.671921644458"
## [1] "AIC: 21322.7497301962"
## [1] "BIC: 21412.0105338985"
## [1] "Quantile Departure: 0.0472456539674342"
```




```
xtable(model.1, caption="Predictor Statistics and Significance", label="First.Order.Table")
```

```
## % latex table generated in R 4.2.1 by xtable 1.8-4 package
## % Wed Dec  7 17:31:29 2022
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrr}
##   \hline
##  & Estimate & Std. Error & t value & Pr($>$$|$t$|$) \\
##   \hline
## (Intercept) & -964.2892 & 294.9607 & -3.27 & 0.0011 \\
##   Plural2 & -704.7067 & 76.5728 & -9.20 & 0.0000 \\
##   Plural3 & -932.1278 & 237.3885 & -3.93 & 0.0001 \\
##   SexMale & 93.8965 & 24.8250 & 3.78 & 0.0002 \\
```

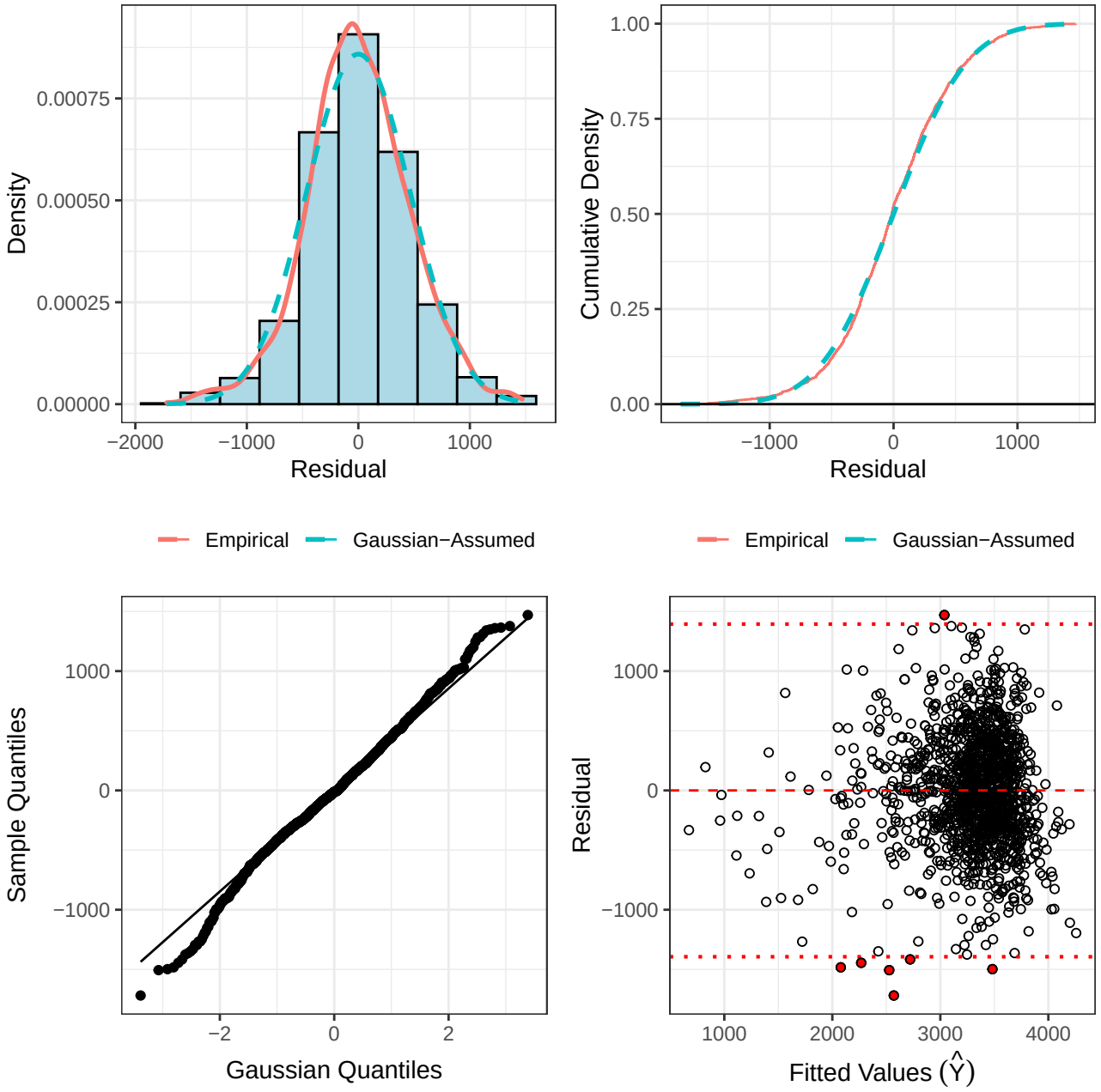
```
##   MomAge & 10.8808 & 2.3232 & 4.68 & 0.0000 \\
##   Weeks & 97.8114 & 6.8024 & 14.38 & 0.0000 \\
##   RaceMomBlack & -76.8770 & 102.7337 & -0.75 & 0.4544 \\
##   RaceMomChinese & 81.3075 & 344.1690 & 0.24 & 0.8133 \\
##   RaceMomFilipino & -860.4032 & 476.1360 & -1.81 & 0.0710 \\
##   RaceMomJapanese & 29.0363 & 106.7611 & 0.27 & 0.7857 \\
##   RaceMomOther Asian / PI & -58.4311 & 140.9347 & -0.41 & 0.6785 \\
##   RaceMomWhite & 30.7840 & 101.2713 & 0.30 & 0.7612 \\
##   MaritalUnmarried & -48.0730 & 32.1294 & -1.50 & 0.1348 \\
##   Gained & 7.8665 & 0.9099 & 8.65 & 0.0000 \\
##   SmokeYes & -203.4873 & 36.5920 & -5.56 & 0.0000 \\
##   PremieYes & -217.6214 & 53.3527 & -4.08 & 0.0000 \\
##    \hline
## \end{tabular}
## \caption{Predictor Statistics and Significance}
## \label{First.Order.Table}
## \end{table}
```

R-Squared	Adjusted R-Squared	RSE	AIC	BIC
0.4570	0.4512	464.6719	21322.7497	21412.0105

Table 1: Summary of first order regression model R-squared, Adj R-squared, RSE, AIC, and BIC.

As Table **??** indicates, Plural2 (Twins), Plural3 (Triplets), Sex = Male, MomAge, Weeks, RaceMom = Filipino, Gained, Smoke = Yes, and Premie = Yes are all significant predictors. Table 1 demonstrates that the R-squared is is fairly high, meaning the model is quite predictive; however, AIC and BIC values are also extremely high, meaning that the number of predictor variables is not predictive enough to justify the variety of parameters we are using.

```
plotResiduals(model.1)
```



Before eliminating the non-significant variables from our model we wanted to test transformations of BirthWeight, to improve residual distribution. The model statistics for each transformation are listed in the table below.

```
# Log of Weight
lm(WeightGmLog ~ Plural + Sex + MomAgeSC + WeeksSC + RaceMom +
     Marital + GainedSC + Smoke + Premie, births) -> model.log

# Square Root of Weight
lm(WeightGmSqrt ~ Plural + Sex + MomAgeSC + WeeksSC + RaceMom +
     Marital + GainedSC + Smoke + Premie, births) -> model.sqrt

# Squared of Weight
lm(WeightGmS ~ Plural + Sex + MomAgeSC + WeeksSC + RaceMom +
     Marital + GainedSC + Smoke + Premie, births) -> model.s
```

Transformation	R-Squared	Adjusted R-Squared	RSE	AIC	BIC
First-order Model	0.4570	0.4512	464.6719	21322.7497	21412.0105
Log of Weight	0.5243	0.5192	0.1756	-885.2705	-796.0097
Square Root of Weight	0.4994	0.4940	4.3160	8137.3786	8226.6394
Weight Squared	0.3699	0.3631	3044641.3454	46086.0996	46175.3604
Log of Weight Squared	0.5208	0.5156	2.6968	6812.1540	6901.4149
Log of the Log of Weight	0.5249	0.5198	0.0231	-6597.7982	-6508.5374
Log of the Square Root of Weight	0.5243	0.5192	0.3512	1068.0181	1157.2789
Square Root of the Log of Weight	0.5250	0.5199	0.0318	-5699.0284	-5609.7676
Inverse of the Weight	0.5250	0.5199	0.0318	-5699.0284	-5609.7676

Table 2: Summary of all adjusted first order regression model R-squared, Adj R-squared, RSE, AIC, and BIC.

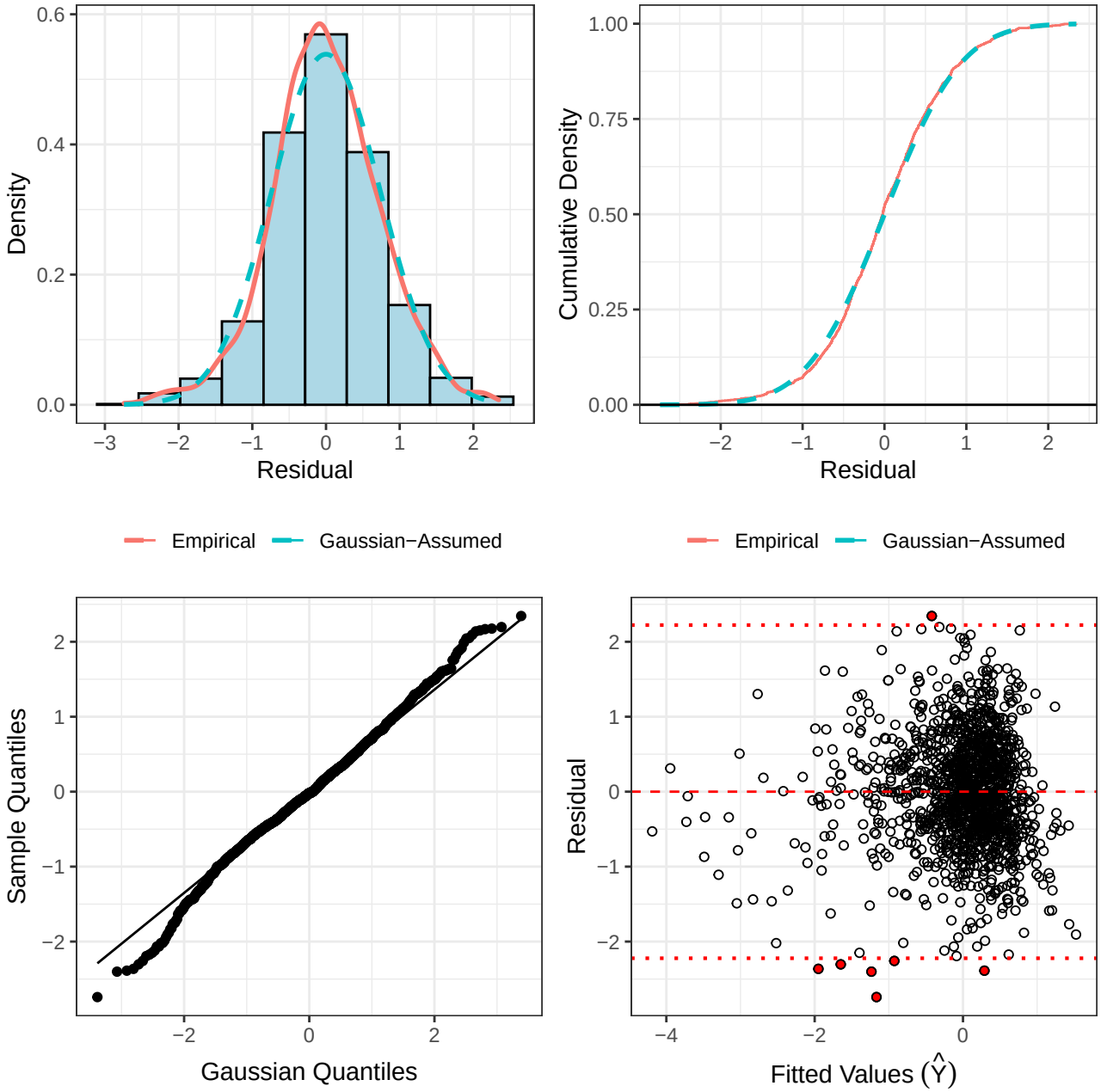
```
# Square of the Log of Weight
lm(WeightGmSLog ~ Plural + Sex + MomAgeSC + WeeksSC + RaceMom +
     Marital + GainedSC + Smoke + Premie, births) -> model.slog

# Log of the Log of Weight
lm(WeightGmLogLog ~ Plural + Sex + MomAgeSC + WeeksSC + RaceMom +
     Marital + GainedSC + Smoke + Premie, births) -> model.loglog

# Log of the Square Root of Weight
lm(WeightGmLogSqr ~ Plural + Sex + MomAgeSC + WeeksSC + RaceMom +
     Marital + GainedSC + Smoke + Premie, births) -> model.logsqr

# Square Root of the Log of Weight
lm(WeightGmSqrtLog ~ Plural + Sex + MomAgeSC + WeeksSC + RaceMom +
     Marital + GainedSC + Smoke + Premie, births) -> model.sqrtlog

# Inverse of the Weight
lm(WeightGmInverse ~ Plural + Sex + MomAgeSC + WeeksSC + RaceMom +
     Marital + GainedSC + Smoke + Premie, births) -> model.inverse
```

At this point, we have two potential "best models": the un-transformed weight, which better satisfies the assumptions behind linear regression modeling, namely the equal distribution of residuals. We also have the log weight model, which has higher r-squared values and lower sigma, AIC, and BIC values than most other models, meaning it is more accurate. This accuracy is, however, gained only at the expense of making residuals less equally distributed, and thus satisfies the assumptions necessary to do things like conduct statistical inference on parameters. Thus, we will be proceeding with two models, the assumptions model and the accuracy model, where one will seek to minimize departure from assumptions and the other will seek to maximize accuracy. Assumption models will use the scaled and centered gram weight as their target, while accuracy models will use the scaled and centered log of gram weight.

Below is our current First-Order Model with all appropriate transformations, and numeric attributes scaled and centered.

```
#First-Order Model with All Appropriate Transformations, and Numeric Attributes Scaled & Centered
lm(WeightGmSC ~ Plural + Sex + MomAgeSC + WeeksSC + RaceMom +
     Marital + GainedSC + Smoke + Premie, births) -> model.2.assu

modelSummary(model.2.assu)

##                            Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)              0.07028229 0.16262733  0.4321678 0.6656863816
## Plural2                 -1.12347022 0.12207523 -9.2030973 0.0000000000
## Plural3                 -1.48603366 0.37845375 -3.9265925 0.0000903705
```

```
## SexMale                   0.14969336 0.03957688  3.7823432 0.0001619382
## MomAgeSC                  0.10576700 0.02258291  4.6834980 0.0000030955
## WeeksSC                   0.42088560 0.02927085 14.3790033 0.0000000000
## RaceMomBlack             -0.12256027 0.16378195 -0.7483137 0.4543973043
## RaceMomChinese            0.12962343 0.54868736  0.2362428 0.8132790072
## RaceMomFilipino          -1.37168754 0.75907408 -1.8070536 0.0709696081
## RaceMomJapanese           0.04629081 0.17020262  0.2719747 0.7856817781
## RaceMomOther Asian / PI -0.09315303 0.22468347 -0.4145967 0.6785009952
## RaceMomWhite              0.04907703 0.16145060  0.3039755 0.7611919583
## MaritalUnmarried         -0.07663983 0.05122190 -1.4962315 0.1348198655
## GainedSC                  0.17403975 0.02013002  8.6457814 0.0000000000
## SmokeYes                 -0.32440725 0.05833638 -5.5609767 0.0000000321
## PremieYes                -0.34694024 0.08505692 -4.0789183 0.0000478089
## [1] "R-squared: 0.457065231018961"
## [1] "Adjusted R-Squared: 0.451218840828928"
## [1] "Sigma: 0.740797650624698"
## [1] "AIC: 3170.99900079143"
## [1] "BIC: 3260.25980449371"
## [1] "Quantile Departure: 0.0472456539674346"
```



$$\begin{aligned}\hat{y} = {} & 0.0703 - 1.1235 \cdot \text{Twin} - 1.4860 \cdot \text{Triplet} + 0.1497 \cdot \text{I(Sex = Male)} \\ & + 0.1058 \cdot \text{MomAgeSC} + 0.4209 \cdot \text{WeeksSC} \\ & - 0.1226 \cdot \text{I(RaceMom = Black)} + 0.1296 \cdot \text{I(RaceMom = Chinese)} \\ & - 1.3717 \cdot \text{I(RaceMom = Filipino)} + 0.0463 \cdot \text{I(RaceMom = Japanese)} \\ & - 0.0932 \cdot \text{I(RaceMom = Other Asian/PI)} + 0.0491 \cdot \text{I(RaceMom = White)} \\ & - 0.0766 \cdot \text{I(Marital = Unmarried)} + 0.1740 \cdot \text{GainedSC} \\ & - 0.3244 \cdot \text{I(Smoke = Yes)} - 0.3469 \cdot \text{I(Premie = Yes)}\end{aligned}$$

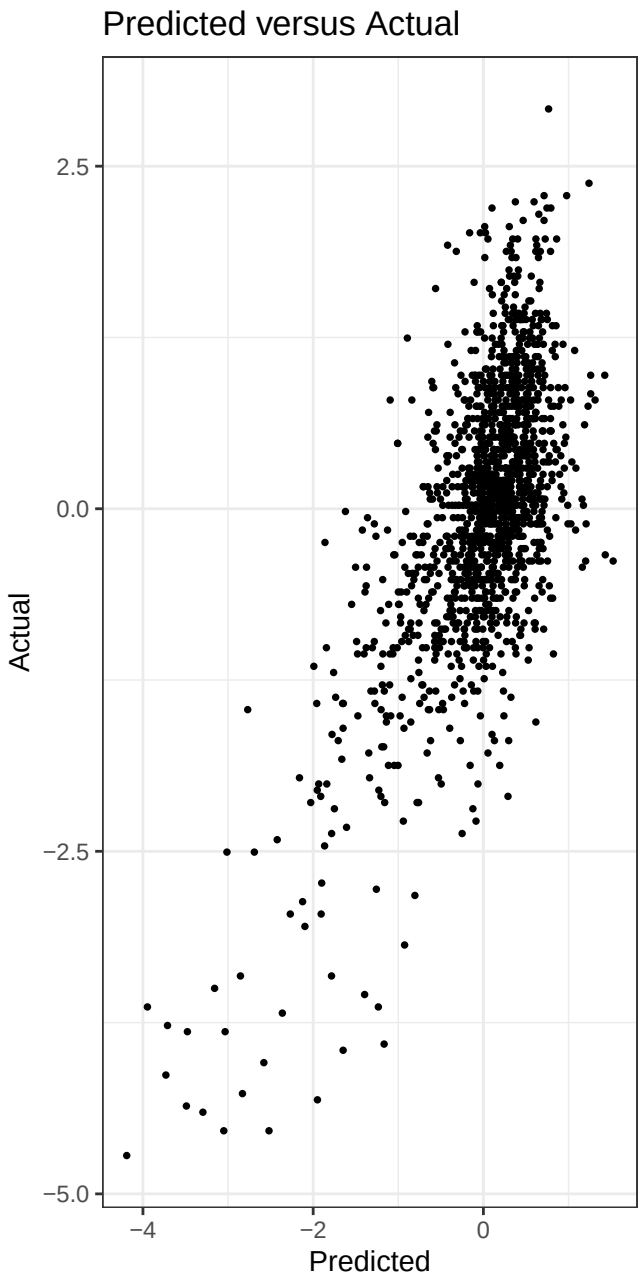
After obtaining our new model, we were curious as to how some of the predictors influenced the data set. Preemie means that the baby was born weeks before the predicted due date, and its commonly known that preemies have

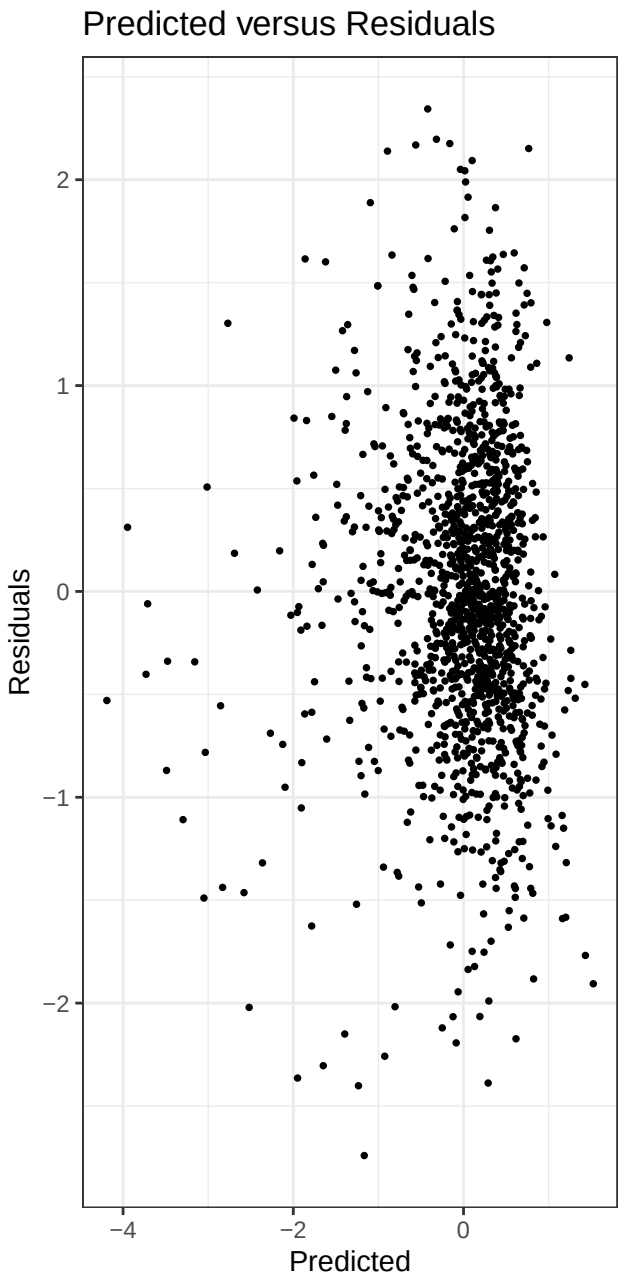
low birth weights.

The code below tests the model on the data subset that only includes preemies, and the data subset that excludes preemies. We then ran the model on both and compared the results.

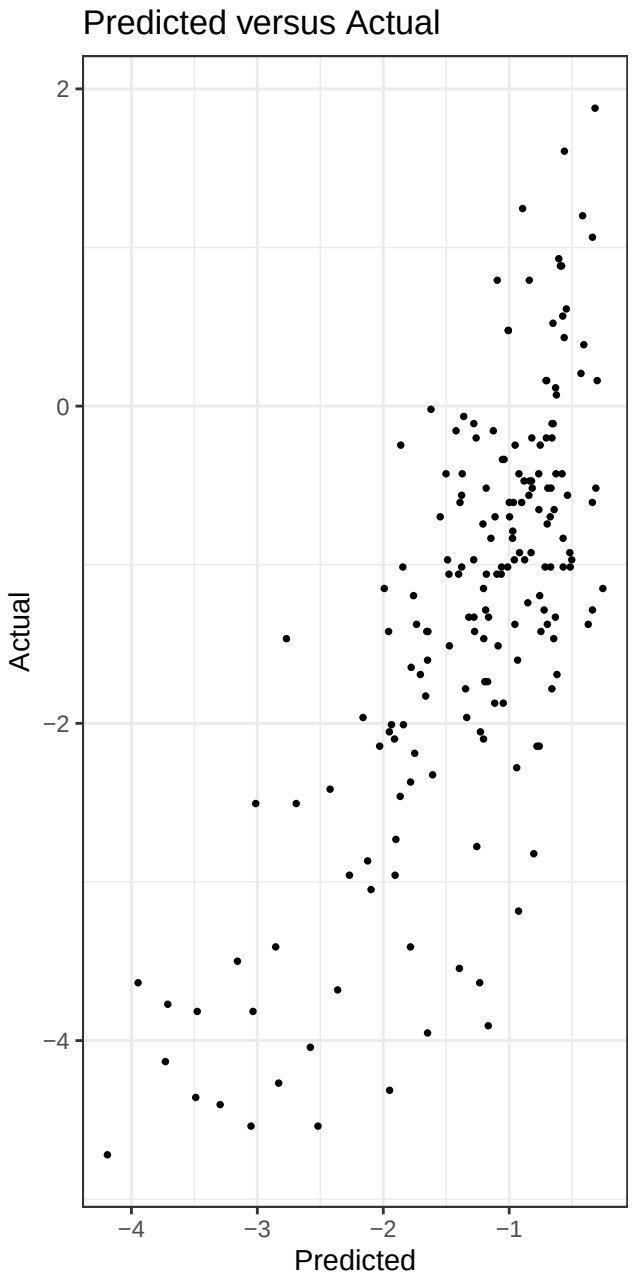
```
predictForSubsets(model.2.assu, "WeightGmSC",
                  births,
                  births %>% filter(Premie=="Yes"),
                  births %>% filter(Premie=="No"))
```

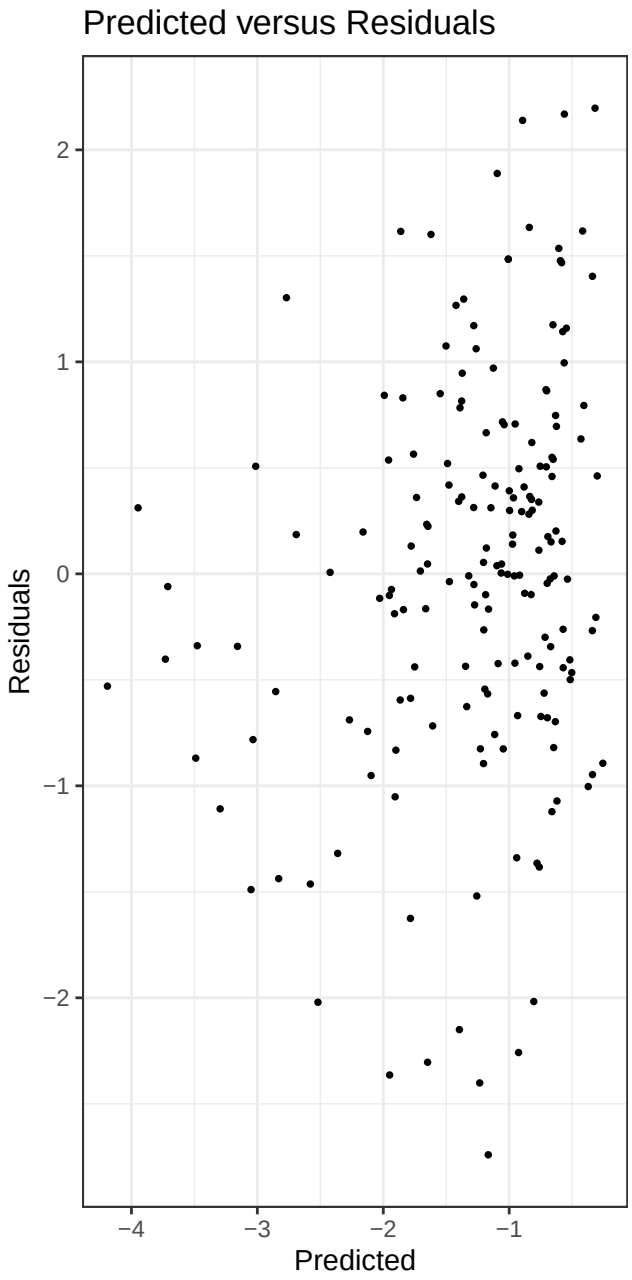
```
## [1] "Subset 1 R-Squared: 0.457065231018961"
## [1] "Subset 1 Mean Abs. Error: 0.571443438631823"
```



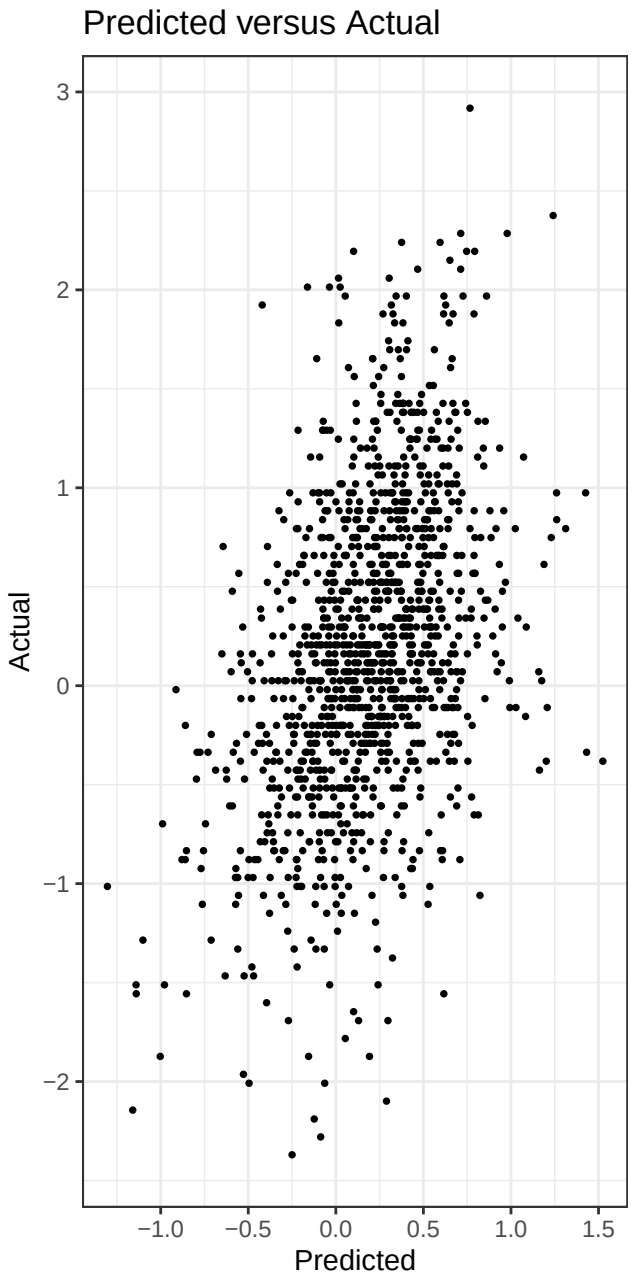


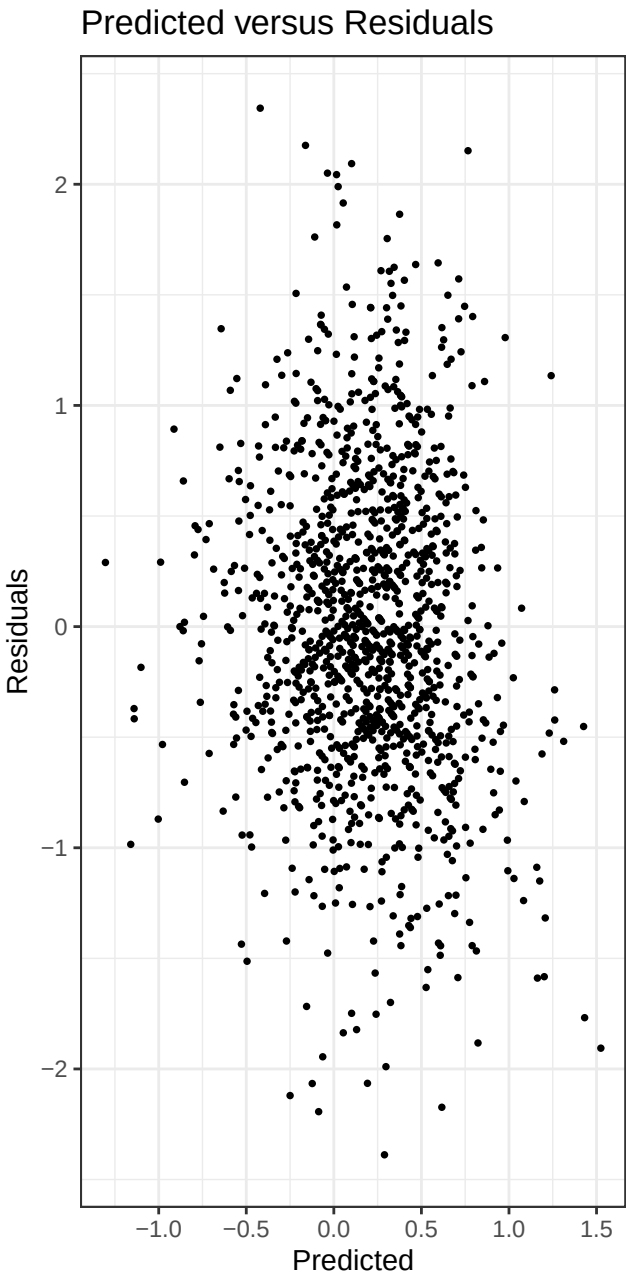

```
## [1] "Subset 2 R-Squared: 0.551006218232442"
## [1] "Subset 2 Mean Abs. Error: 0.709020485309315"
```






```
## [1] "Subset 3 R-Squared: 0.175512160348786"
## [1] "Subset 3 Mean Abs. Error: 0.551165388592225"
```

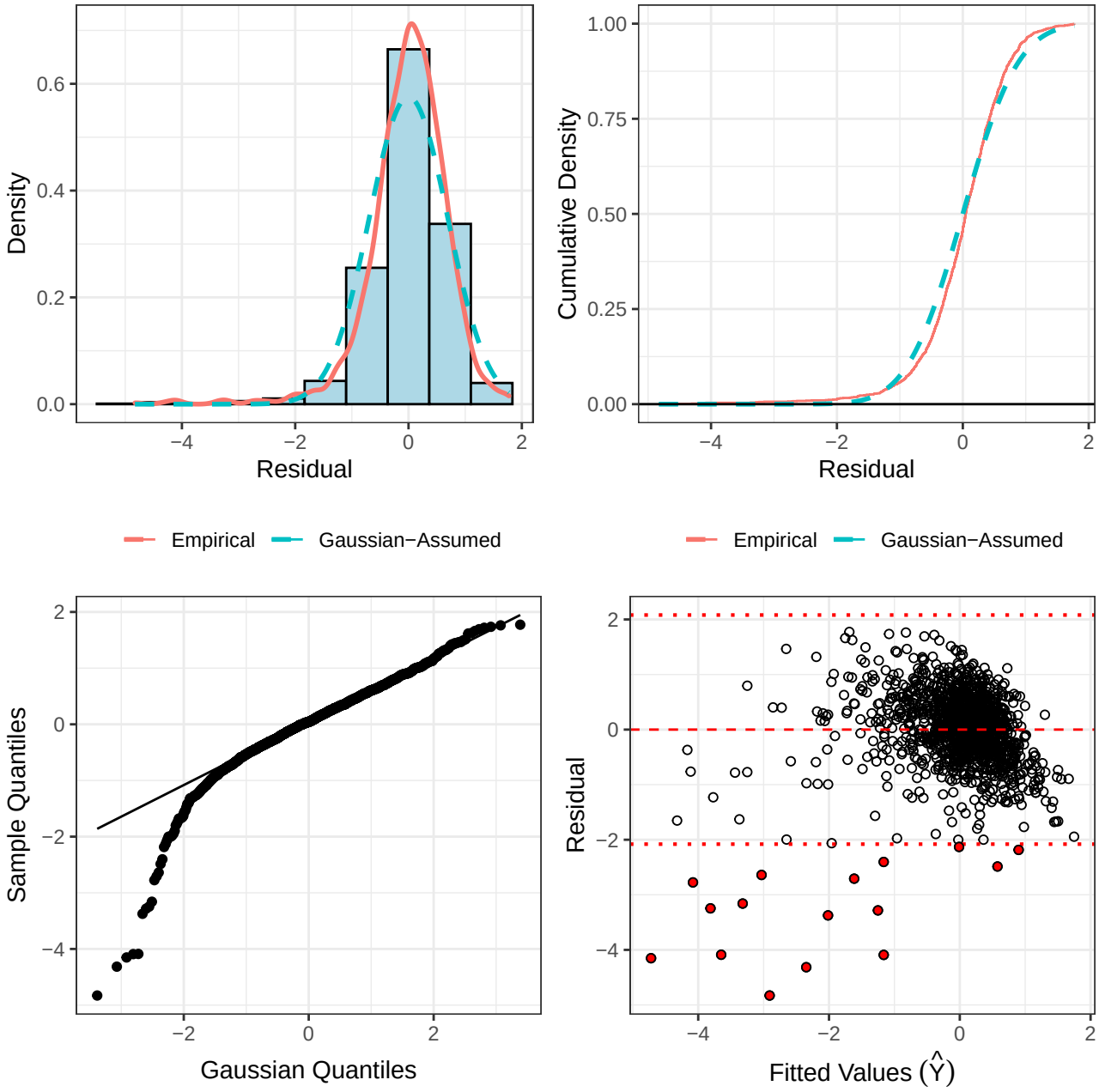





```
lm(WeightLogSC ~ Plural + Sex + MomAgeSC + WeeksSC + RaceMom +
     Marital + GainedSC + Smoke + Premie, births) -> model.2.accur

modelSummary(model.2.accur)
```

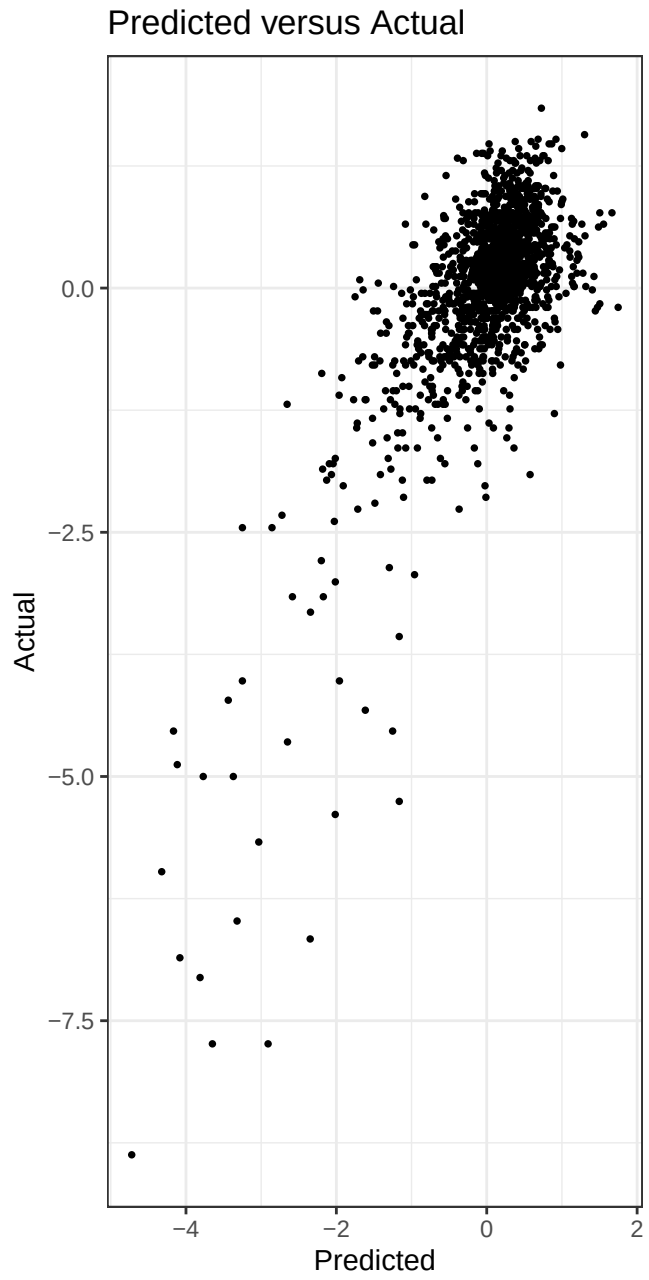
```
##                          Estimate Std. Error     t value     Pr(>|t|)
## (Intercept)             0.10329787 0.15222022  0.67860807 0.4974989472
## Plural2                -1.01191972 0.11426320 -8.85604241 0.0000000000
## Plural3                -1.55437979 0.35423513 -4.38798876 0.0000123026
## SexMale                 0.10769027 0.03704422  2.90707360 0.0037060130
## MomAgeSC                0.10189593 0.02113775  4.82056744 0.0000015884
## WeeksSC                 0.55647332 0.02739770 20.31095005 0.0000000000
## RaceMomBlack           -0.15127673 0.15330096 -0.98679577 0.3239141352
## RaceMomChinese          0.14223435 0.51357488  0.27694959 0.7818599036
```

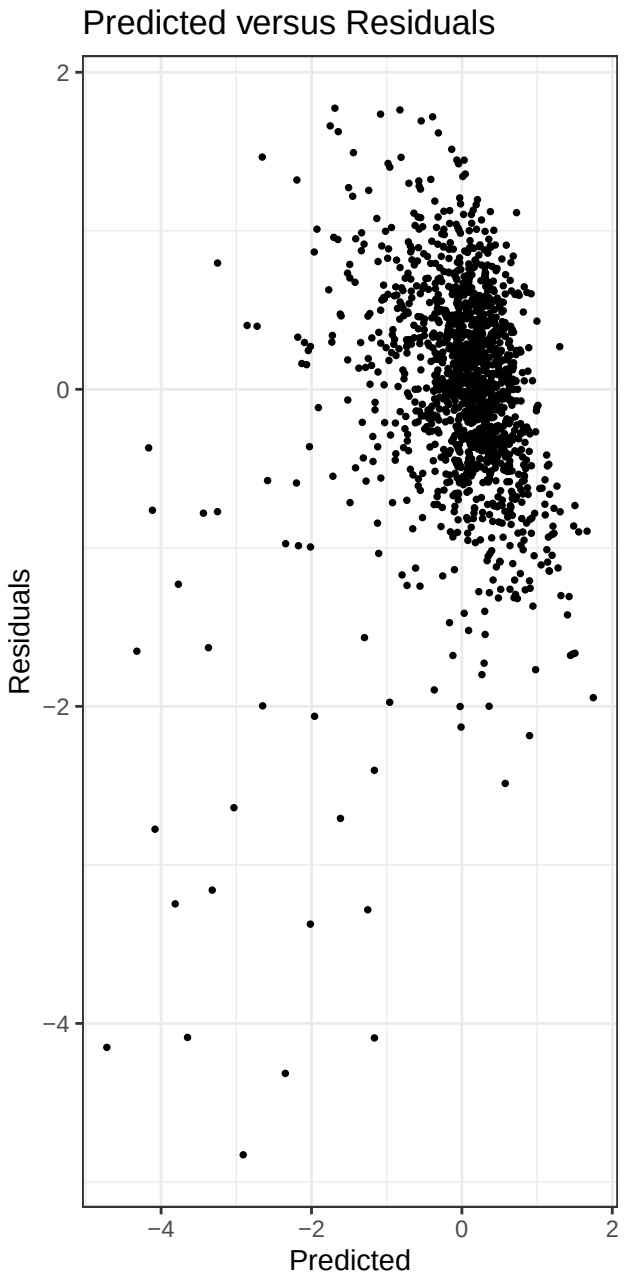
```
## RaceMomFilipino         -1.20670223 0.71049820 -1.69838886 0.0896577113
## RaceMomJapanese          0.01131033 0.15931074  0.07099541 0.9434115974
## RaceMomOther Asian / PI -0.15487126 0.21030517 -0.73641204 0.4616039703
## RaceMomWhite            -0.01995089 0.15111880 -0.13202126 0.8949865415
## MaritalUnmarried        -0.07303122 0.04794403 -1.52326000 0.1279206442
## GainedSC                 0.15321292 0.01884183  8.13153173 0.0000000000
## SmokeYes                -0.24094569 0.05460323 -4.41266412 0.0000109974
## PremieYes               -0.15266028 0.07961383 -1.91750960 0.0553775351
## [1] "R-squared: 0.524330514867102"
## [1] "Adjusted R-Squared: 0.519208445752246"
## [1] "Sigma: 0.693391342783968"
## [1] "AIC: 2984.63630846545"
## [1] "BIC: 3073.89711216773"
## [1] "Quantile Departure: 0.157310064009036"
```



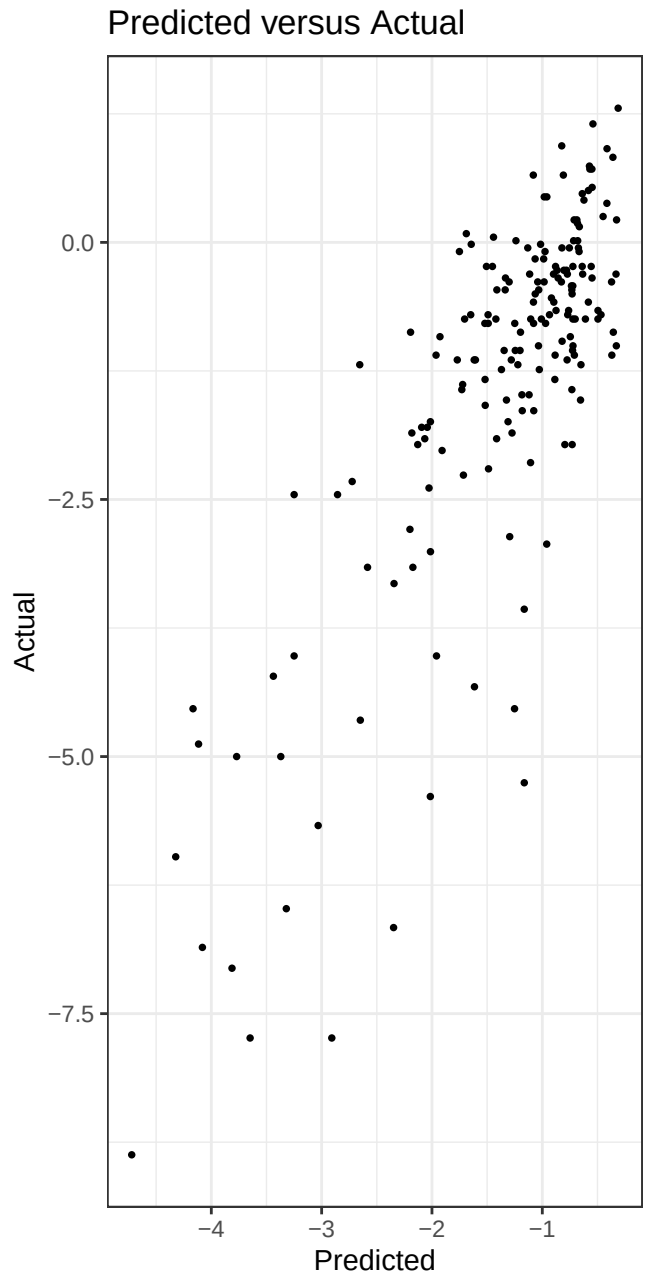

```
predictForSubsets(model.2.accur, "WeightLogSC",
                  births,
                  births %>% filter(Premie=="Yes"),
                  births %>% filter(Premie=="No"))
```

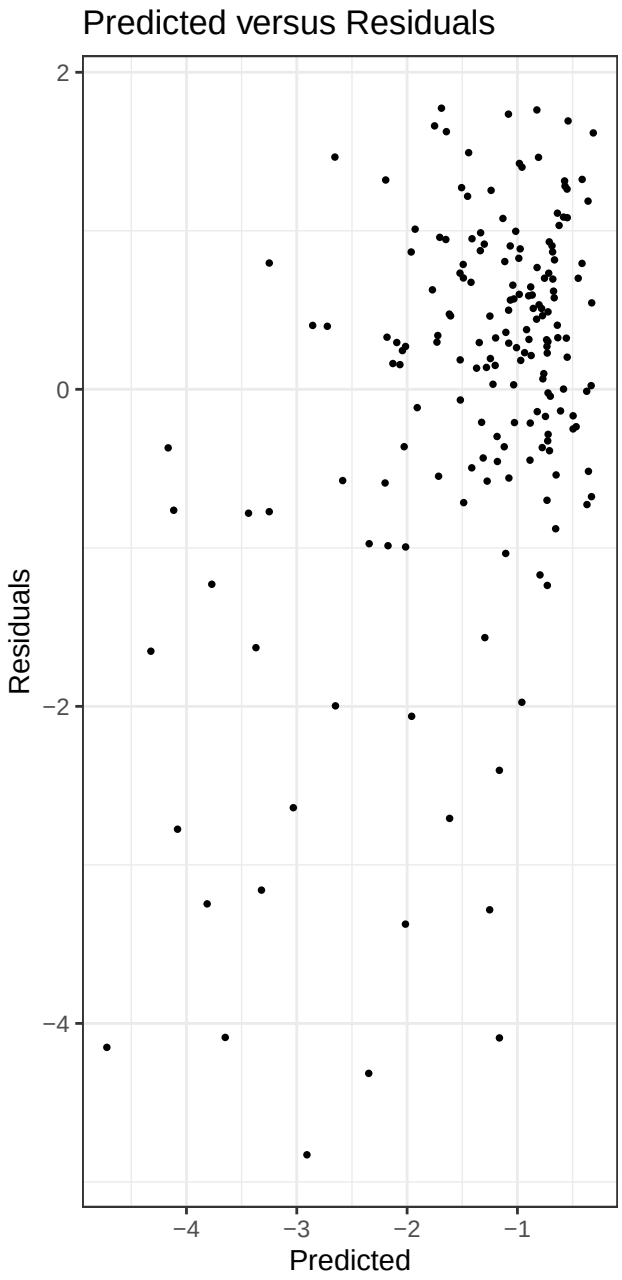
```
## [1] "Subset 1 R-Squared: 0.524330514867102"
## [1] "Subset 1 Mean Abs. Error: 0.493034075135682"
```



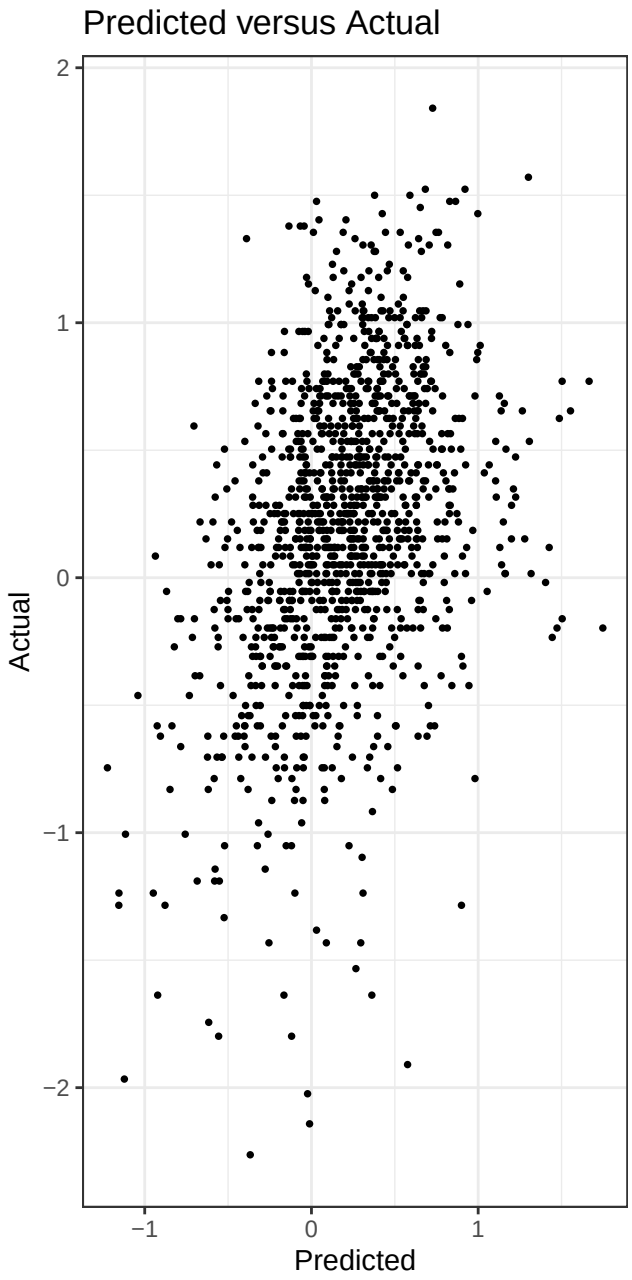


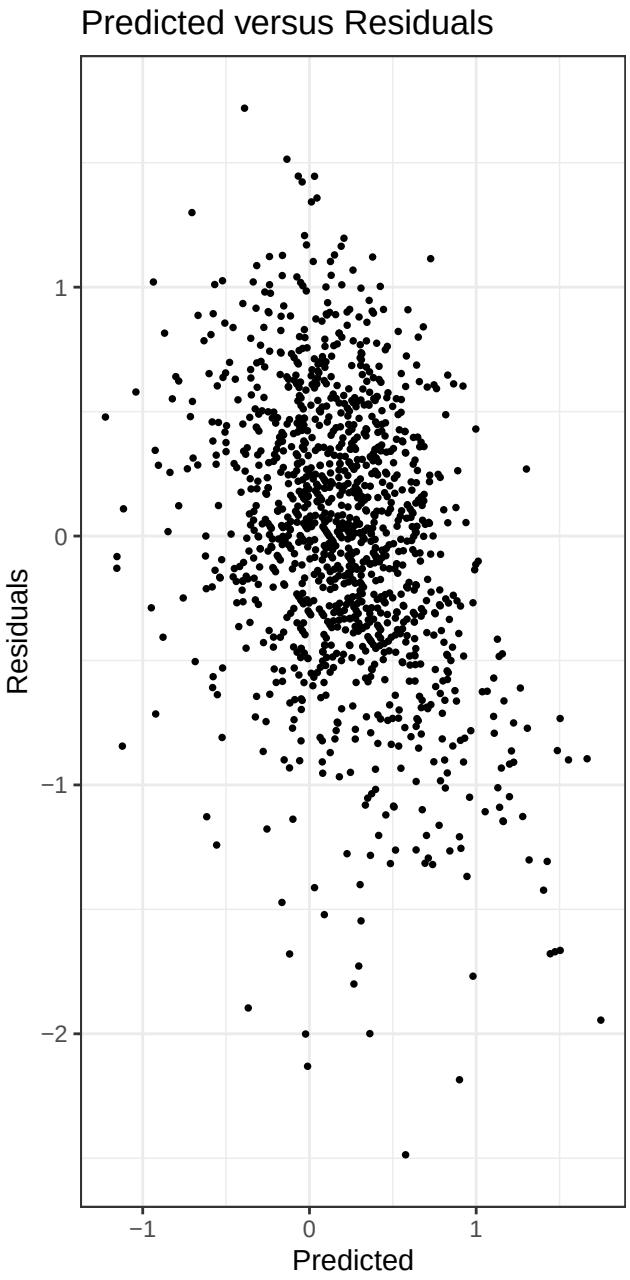

```
## [1] "Subset 2 R-Squared: 0.655114101459069"
## [1] "Subset 2 Mean Abs. Error: 0.884841254580077"
```






```
## [1] "Subset 3 R-Squared: 0.153089471782906"
## [1] "Subset 3 Mean Abs. Error: 0.435283994126369"
```





Subset	R-Squared	Mean Absolute Error
Births	0.4570	0.5714
Preemies	0.5510	0.7090
Not Preemies	0.1755	0.5511

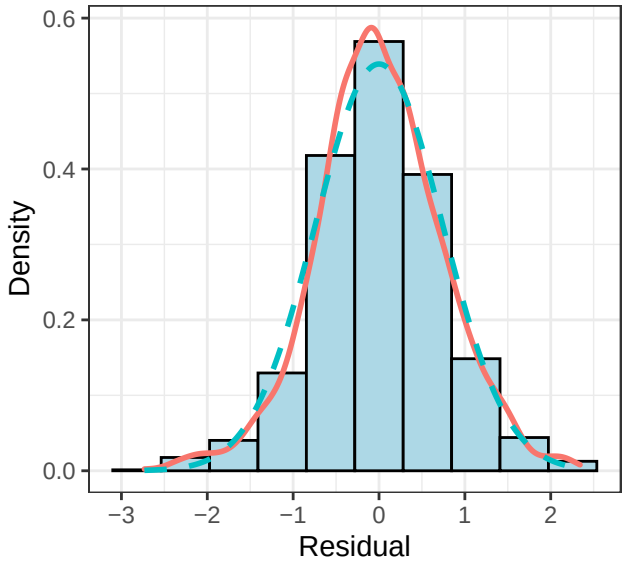
Table 3: Summary of all adjusted first order regression model R-squared, Adj R-squared, RSE, AIC, and BIC.

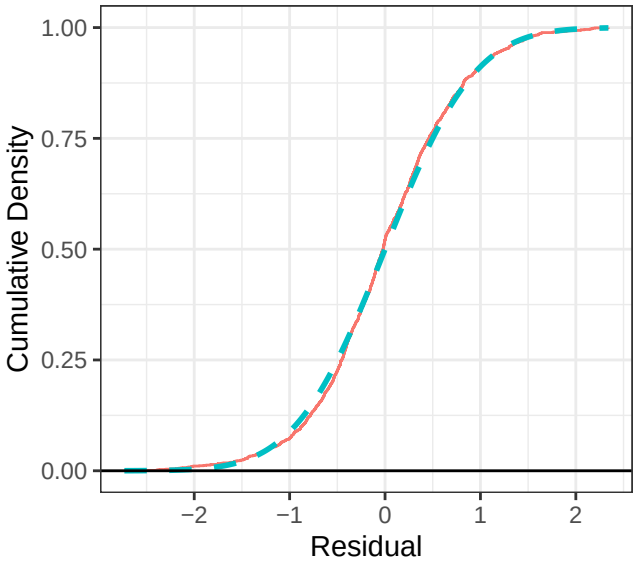
Part 2: AIC/BIC Iteration 1

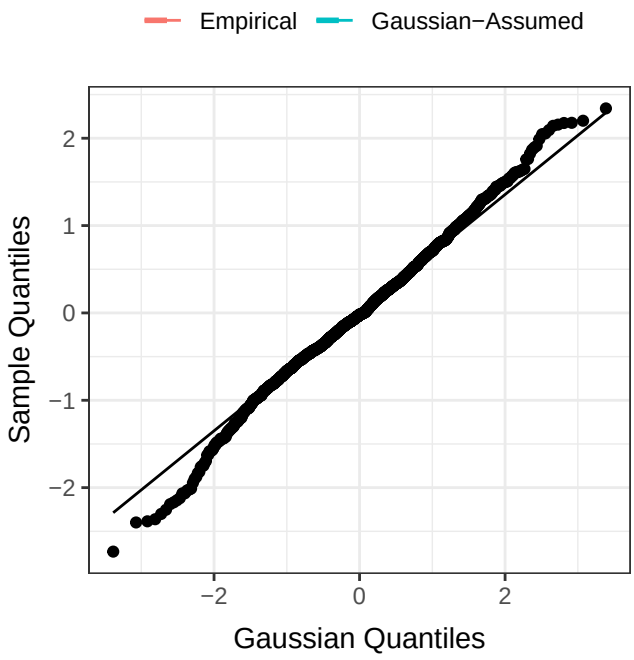
```
## Assumptions Model

x <- model.matrix(model.2.assu)[,-1]

y <- births$WeightGmSC
```

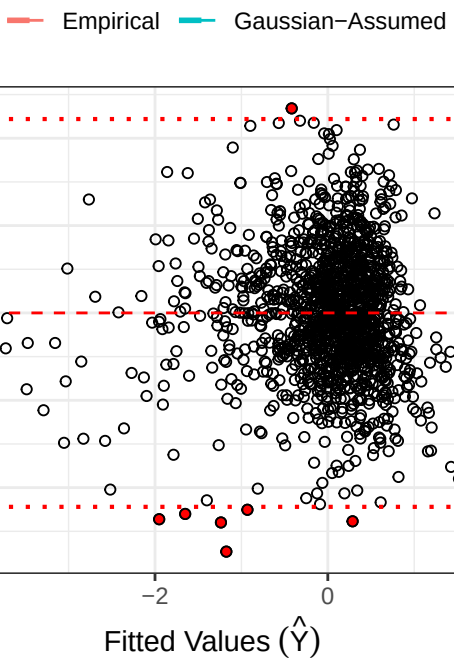
```r
xy <- as.data.frame(cbind(x,y))
best.subsets.aic <- bestglm(xy, IC="AIC", TopModels = 5)
best.model.aic <- best.subsets.aic$BestModel
modelSummary(best.model.aic)
```

```
##                    Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)       0.1155533 0.03565298  3.241056 0.0012189440
## Plural2          -1.1195941 0.12177905 -9.193651 0.0000000000
## Plural3          -1.4808912 0.37795193 -3.918200 0.0000935265
## SexMale           0.1502685 0.03951570  3.802753 0.0001492650
## MomAgeSC          0.1054359 0.02244925  4.696635 0.0000029051
## WeeksSC           0.4196465 0.02914455 14.398801 0.0000000000
## RaceMomBlack     -0.1665708 0.05073699 -3.283025 0.0010523894
## RaceMomFilipino  -1.4157231 0.74133018 -1.909706 0.0563756484
## MaritalUnmarried -0.0790069 0.05003876 -1.578914 0.1145821349
## GainedSC          0.1737710 0.01995059  8.710069 0.0000000000
## SmokeYes         -0.3216242 0.05699092 -5.643428 0.0000000202
## PremieYes        -0.3511604 0.08474998 -4.143486 0.0000362603
## [1] "R-squared: 0.456717769215429"
## [1] "Adjusted R-Squared: 0.452439956374606"
## [1] "Sigma: 0.739973001957095"
## [1] "AIC: 3163.90042969188"
## [1] "BIC: 3232.15869134656"
## [1] "Quantile Departure: 0.0474199906966593"
```





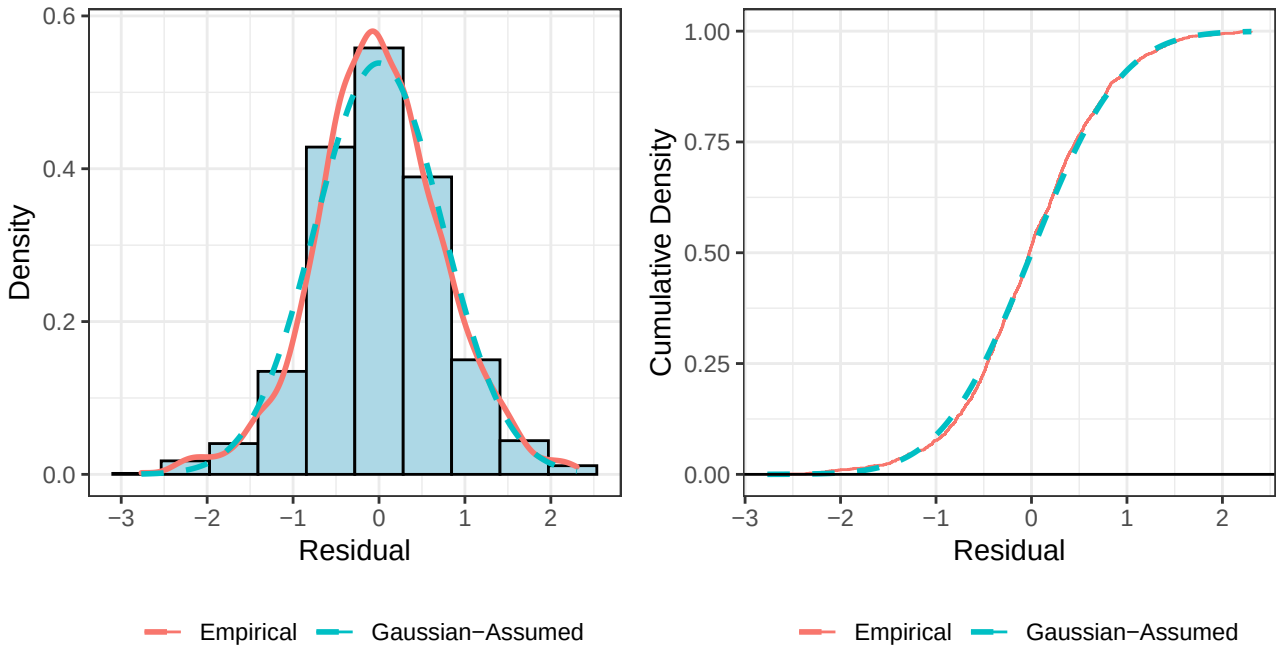


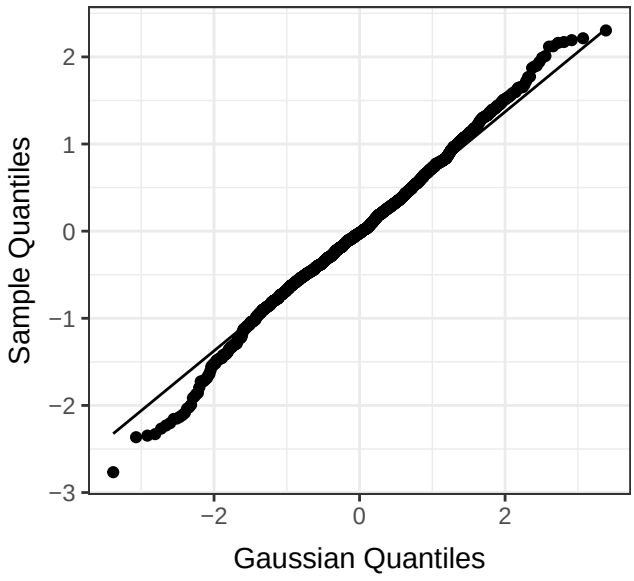


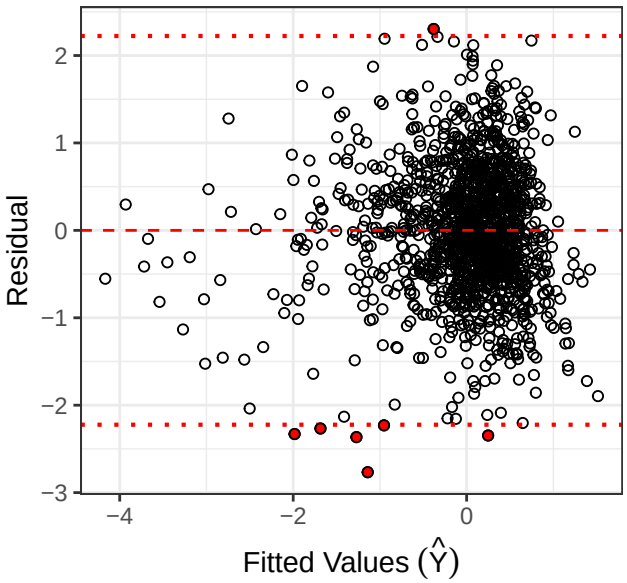

```
best.subsets.bic <- bestglm(xy, IC="BIC", TopModels = 5)
best.model.bic <- best.subsets.bic$BestModel
modelSummary(best.model.bic)
```

```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   0.09475252 0.03368651  2.812773 0.0049802705
## Plural2      -1.12009473 0.12195379 -9.184583 0.0000000000
## Plural3      -1.47212718 0.37849237 -3.889450 0.0001051871
## SexMale       0.15216134 0.03956124  3.846223 0.0001253459
## MomAgeSC      0.11860642 0.02041483  5.809816 0.0000000077
## WeeksSC       0.41869752 0.02918503 14.346313 0.0000000000
## RaceMomBlack -0.19170667 0.04813574 -3.982627 0.0000716623
## GainedSC      0.17381907 0.01997972  8.699777 0.0000000000
## SmokeYes     -0.33430893 0.05643831 -5.923439 0.0000000040
```

```
## PremieYes    -0.35772800 0.08478370 -4.219302 0.0000260829
## [1] "R-squared: 0.45431789056999"
## [1] "Adjusted R-Squared: 0.450807426678017"
## [1] "Sigma: 0.741075281818239"
## [1] "AIC: 3166.11079739794"
## [1] "BIC: 3223.86778802883"
## [1] "Quantile Departure: 0.0461987895225344"
```








```
regsubsets.out <- regsubsets(WeightGmSC ~ Plural + Sex + MomAgeSC + WeeksSC + RaceMom +
                               Marital + GainedSC + Smoke + Premie,
                             data=births, nbest = 1, nvmax=15)

as.data.frame(summary(regsubsets.out)$outmat)
```

```
##           Plural2 Plural3 SexMale MomAgeSC WeeksSC RaceMomBlack RaceMomChinese
## 1  ( 1 )                                     *
## 2  ( 1 )                                     *
## 3  ( 1 )        *                            *
## 4  ( 1 )        *                    *       *
## 5  ( 1 )        *                    *       *
## 6  ( 1 )        *                    *       *
## 7  ( 1 )        *                    *       *           *
## 8  ( 1 )        *       *            *       *           *
## 9  ( 1 )        *       *       *    *       *           *
## 10  ( 1 )       *       *       *    *       *           *
## 11  ( 1 )       *       *       *    *       *           *
## 12  ( 1 )       *       *       *    *       *           *
## 13  ( 1 )       *       *       *    *       *           *              *
## 14  ( 1 )       *       *       *    *       *           *
## 15  ( 1 )       *       *       *    *       *           *              *
##           RaceMomFilipino RaceMomJapanese RaceMomOther Asian / PI RaceMomWhite
## 1  ( 1 )
## 2  ( 1 )
## 3  ( 1 )
## 4  ( 1 )
## 5  ( 1 )
## 6  ( 1 )
## 7  ( 1 )
## 8  ( 1 )
## 9  ( 1 )
## 10  ( 1 )              *
## 11  ( 1 )              *
## 12  ( 1 )              *                                  *
## 13  ( 1 )              *                                  *
## 14  ( 1 )              *              *                   *            *
## 15  ( 1 )              *              *                   *            *
##           MaritalUnmarried GainedSC SmokeYes PremieYes
## 1  ( 1 )
## 2  ( 1 )                        *
## 3  ( 1 )                        *
## 4  ( 1 )                        *
## 5  ( 1 )                        *        *
## 6  ( 1 )                        *        *         *
## 7  ( 1 )                        *        *         *
## 8  ( 1 )                        *        *         *
## 9  ( 1 )                        *        *         *
## 10  ( 1 )                       *        *         *
## 11  ( 1 )               *       *        *         *
## 12  ( 1 )               *       *        *         *
## 13  ( 1 )               *       *        *         *
## 14  ( 1 )               *       *        *         *
## 15  ( 1 )               *       *        *         *
```

```
fit.stats <- data.frame(num.variables=1:15,
                        adjr2 = summary(regsubsets.out)$adjr2,
                        bic=summary(regsubsets.out)$bic)
fit.stats
```

```
##    num.variables     adjr2       bic
## 1              1 0.3419165 -576.0584
## 2              2 0.3659468 -622.2229
```

```
## 3              3 0.3936644 -678.9560
## 4              4 0.4139286 -720.6032
## 5              5 0.4262316 -744.2497
## 6              6 0.4342432 -757.8164
## 7              7 0.4397681 -765.3984
## 8              8 0.4453965 -773.3810
## 9              9 0.4508074 -780.9513
## 10            10 0.4518552 -777.3989
## 11            11 0.4524400 -772.6604
## 12            12 0.4523523 -766.1931
## 13            13 0.4519692 -758.9668
## 14            14 0.4515905 -751.7534
## 15            15 0.4512188 -744.5593

## Of the models, BIC suggests removing Marital and all races
## except Black. AIC / radj2 would retain Filipino and Marital.
## Since there is only a single Philipino instance, we remove Filipino,
## as well as marital since the significance is so low.
```

5 Part 5: Final Model(s) and Conclusions.

```
assessModel <- function(model) {
  print(modelSummary(model))
  print(vif(model))
  print(summary(model$residual))
  print(confint(model))
  lev <- model$model %>% mutate(h.values = hatvalues(model))
  print(summary(lev$h.values))
  p <- 2
  n <- nrow(model$model)
  high.lev <- lev %>% filter(h.values > 2*p/n)
  print(paste("High Lev.:", nrow(high.lev)))
  v.high.lev <- lev %>% filter(h.values > 3*p/n)
  print(paste("Very High Lev.:", nrow(v.high.lev)))
  new.resid <- model$model %>% mutate(stdres = rstandard(model),
                                      stures = rstudent(model))
  print("Standard Residual Quant.:")
  print(summary(new.resid$stdres))
  print("Studentized Residual Quant.:")
  print(summary(new.resid$stures))
  s.outliers.stdres <- new.resid %>% filter(abs(stdres)>3)
  print(paste("Strong Standard Residual Outliers:", nrow(s.outliers.stdres)))
  print(s.outliers.stdres)
  s.outliers.stures <- new.resid %>% filter(abs(stures)>3)
  print(paste("String Studentized Residual Outliers:", nrow(s.outliers.stures)))
  print(s.outliers.stures)
  cooks.values <- model$model %>% mutate(cooks = cooks.distance(model))
  print("Cook's Values:")
  print(summary(cooks.values$cooks))
  cooks.strong <- cooks.values %>% filter(cooks>1)
  print(paste("Strong C. Values:", nrow(cooks.strong)))
}

assessModel(mod.accur.final)
```

```
## Error in summary(model):  object 'mod.accur.final' not found
```

```
assessModel(mod.accur.final.part)
## Error in summary(model):  object 'mod.accur.final.part' not found
assessModel(min.vif.model)
## Error in summary(model):  object 'min.vif.model' not found
```

Highest Accuracy Model:

$$\hat{y} = 0.119 + -1.165 \cdot \text{I}(\text{Plural} = 2) + -2.006 \cdot \text{I}(\text{Plural} = 3) + 0.104 \cdot \text{I}(\text{Sex} = \text{Male}) \quad +0.110 \cdot \text{MomAgeSC} + 0.398 \cdot \text{WeeksSC} -$$

Highest accuracy with the slowest departure from Minimizing residuals:

$$\hat{y} = 0.1122 + -1.2928 \cdot \text{I}(\text{Plural} = 2) + -0.7222 \cdot \text{I}(\text{Plural} = 3) + 0.1476 \cdot \text{I}(\text{Sex} = \text{Male}) \quad +0.1653 \cdot \text{MomAgeSC} - 0.0259 \cdot \text{Mo}$$

max accuracy with limited variables

$$\hat{y} = 0.1726 + -1.1681 \cdot \text{I}(\text{Plural} = 2) + -0.9379 \cdot \text{I}(\text{Plural} = 3) + -0.1535 \cdot \text{I}(\text{Black} = \text{TRUE}) \quad +0.9901 \cdot \text{I}(\text{Premie} = \text{Yes}) + 0.$$