Statistical Report on Associations Between Birthweight and Various Explanatory Variables

Sahil

27/03/24

Introduction

This report undertakes a comprehensive statistical analysis with the aim of identifying the key factors that significantly influence the birth weight of children. Drawing from a cohort of 327 children, this study meticulously examines various potential explanatory variables such as the mother's age, gestation period, sex of the child, maternal smoking habits during pregnancy, pre-pregnancy weight of the mother, and rate of growth in the first trimester. These factors are thoroughly explored to understand their relationships with birth weight, a critical indicator of newborn health.

The primary goal of this analysis is to ascertain which variables significantly predict birth weight, thereby providing insights that could be useful for medical professionals to enhance prenatal care and interventions. By employing rigorous statistical models and diagnostic checks, this investigation seeks to offer reliable and actionable findings that support better health outcomes for newborns.

Furthermore, the data set titled "BirthTrain.txt" serves as the training set for developing our statistical models. At the same time, the "BirthTest.txt" file provides a test set used solely for evaluating the predictive performance of these models. This structured approach ensures the study adheres to robust scientific methods, focusing on model development and validation stages.

Explotary Data Analysis

In this part we carry out some preliminary and **exploratory analysis**, then **clean and preprocess the data** repeating the process

Data Preprocessing

After importing the Birth training data, we perform a preliminary check of the dataframe by looking at the variables and their first few values.

```
## 'data.frame':    327 obs. of  7 variables:
##  $ age   : int  25 34 29 23 26 30 27 26 40 19 ...
##  $ gest  : int  294 252 280 280 266 245 273 273 287 287 ...
##  $ sex   : Factor w/ 2 levels "Female","Male": 1 1 2 1 1 2 1 1 2 2 ...
##  $ smokes: Factor w/ 3 levels "Heavy","Light",..: 3 3 1 2 2 1 3 3 3 3 ...
##  $ weight: num  65.2 58.6 57.3 65.9 59.1 49.1 69.5 75 82 58.3 ...
##  $ rate  : num  -0.01618 0.00678 0.03293 -0.02179 0.06346 ...
##  $ bwt   : num  3.87 2.44 3.4 2.82 2.94 2.4 3.08 3.6 3.6 3.93 ...
```

We can see that there are in total 327 observations. From the first few values. It can be seen that there is no data of wrong form, and we find the count of 'not a number' is 0. I changed structure for sex and smoke into factors for functional use in data frame. Next, we look at the distribution of each variable both visually and numerically to see if there are any anomalous values.

Summary Statistics

Based on the summary statistics, we can note that the average age of mothers is around 27.68 years, the average gestation period is approximately 276 days, the average pre-pregnancy weight of mothers is approximately 58.29 kg, the average growth rate during the first trimester is 0.028550 (suggesting a slight positive growth across the dataset), and the average birthweight is around 3.460 kg.

Summary of Data:

```
##       age              gest          sex         smokes        weight
##  Min.   :15.00   Min.   :224   Female:153   Heavy: 13   Min.   : 40.90
##  1st Qu.:24.00   1st Qu.:273   Male  :174   Light: 14   1st Qu.: 52.30
##  Median :27.00   Median :280                No   :300   Median : 56.80
##  Mean   :27.68   Mean   :276                            Mean   : 58.29
##  3rd Qu.:31.00   3rd Qu.:280                            3rd Qu.: 63.00
##  Max.   :42.00   Max.   :301                            Max.   :100.00
##       rate                bwt
##  Min.   :-0.040020   Min.   :1.640
##  1st Qu.: 0.004185   1st Qu.:3.170
##  Median : 0.026110   Median :3.440
##  Mean   : 0.028550   Mean   :3.461
##  3rd Qu.: 0.048285   3rd Qu.:3.760
##  Max.   : 0.124320   Max.   :5.680
```

When reviewing the plot, it was initially noted that there was a discrepancy in the units of measurement for child and mother weight. While the child's weight at birth was in grams, the mother's pre-pregnancy weight was given in kilograms. To rectify this, we converted the child's weight from grams to kilograms before inputting the plot into the boxplot.

Upon comprehensive analysis, it was observed that most of the data points fell within a similar range, with the exception of the Gestation period, which is measured in days. To provide a more comprehensive view, it would be strategic to analyze the distributions for some of the data in groups, such as the birth weight of male and female children in the same plot. However, some data points, such as the mother's age and the Gestation period, should be analyzed separately as they are measured in different units.

Note: In this report the colours blue and pink will be used for data identifying male and female babies respectively for better visuals and to understand the distribution of both sexes.

Regarding other data points, I did not remove them for the following three reasons:

- The physical conditions of different people can vary tremendously between different people; we cannot exclude them from the model just because they have exceptional physical conditions
- At this stage, we cannot decide whether they are outliers or high leverage point
- despite being special, these measurements seem to be correct, so special care needs to be taken

Therefore, we do not remove these deviating points at this step.

Observations from the graphs:

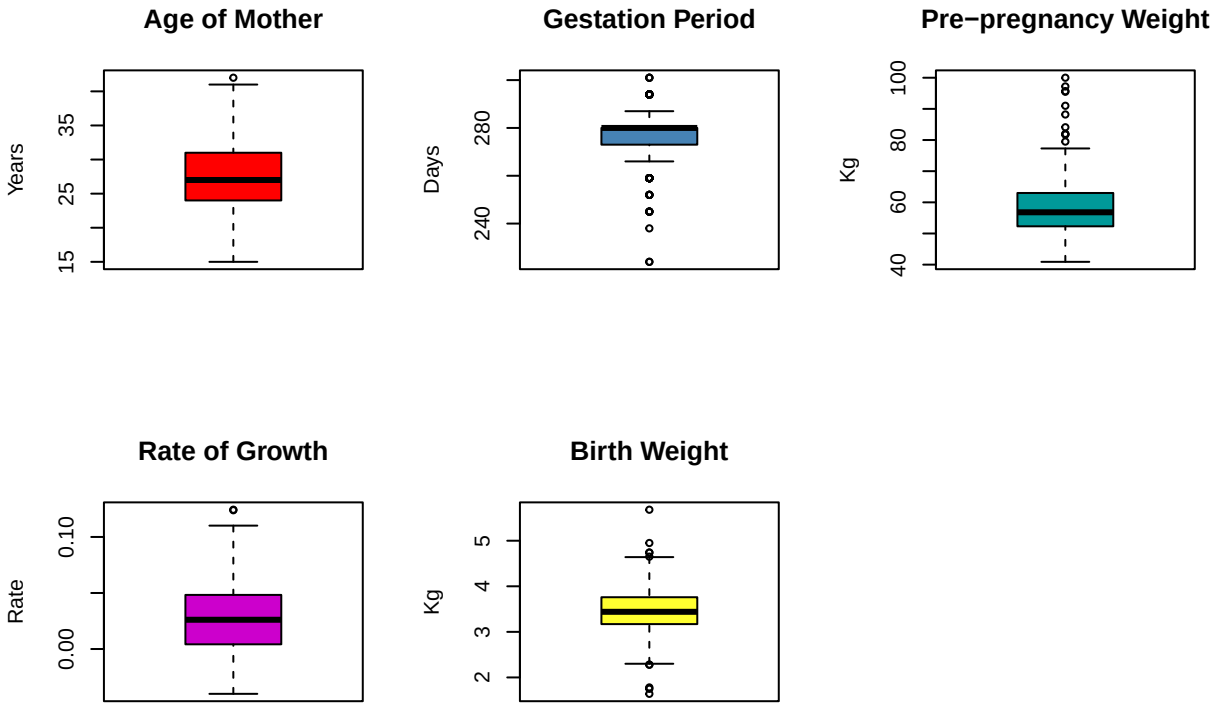
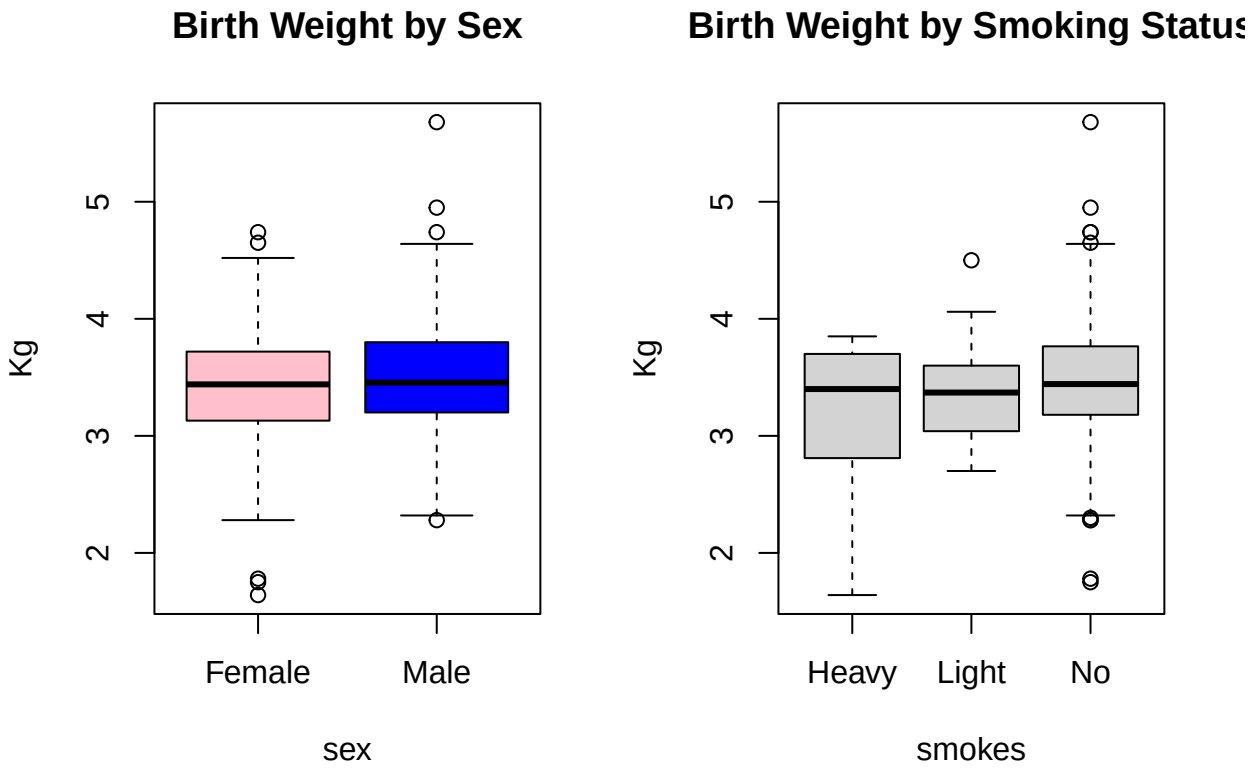


Figure 1: Group/Indiviual Data

- **Age of Mother**: The median age of mothers is around 30. There is a broader spread of ages in younger mothers (under 30) compared to older mothers (over 30).
- **Gestational Period**: The median gestational period is around 260 days. There is a smaller range of gestational periods compared to the other variables.
- **Pre-pregnancy Weight**: The median pre-pregnancy weight is around 60 kg. The distribution of pre-pregnancy weight is skewed to the right, with more mothers having higher weights.
- **Birth Weight**: The median birth weight is around 3000 grams (or 3 kg). Birth weight distribution is also skewed to the right, with more babies having higher birth weights.

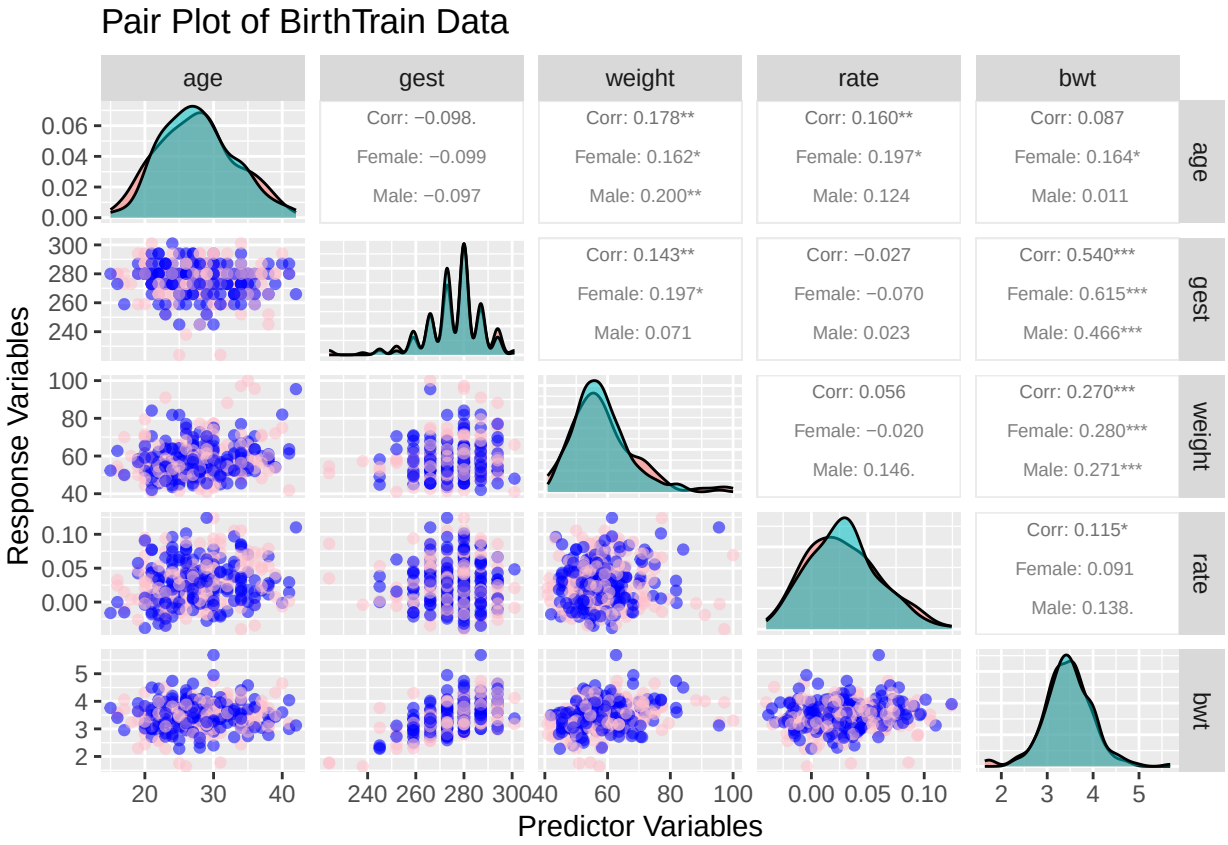


- **Birth weight by sex:** The median birth weight appears slightly higher for males than females. The box plots also show a more comprehensive range of male birth weights.
- **Birth weight by smoking status:** The median birth weight appears to be lower for mothers who smoke compared to those who do not smoke. The interquartile range (IQR) is similar for both groups. A few outliers in the "Heavy Smokes" group have a lower birth weight than the rest of the data.

Meanwhile, it is easy to see that there are observations with the rate of growth being negative, which may be an error, steaming from an incorrect recording or measurement. However, realistically, there can be negative growth of children due to health complications. In order to make the model more generalization, we choose to keep these observations.

Explotary Analysis

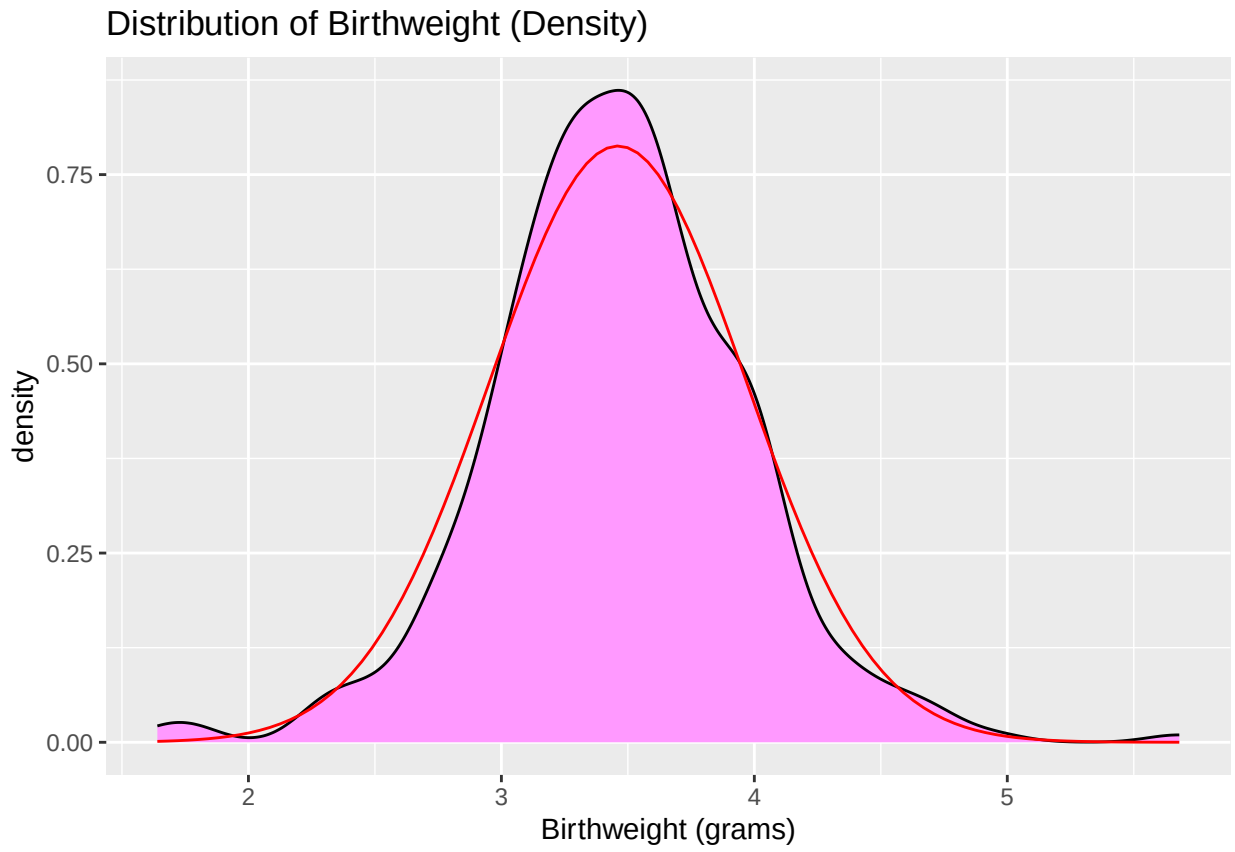
I explore further using different plots after the data has been vetted I explored it further using different plots.By directly pairing the variables with each other and plot their scatter matrix.I chose to have the data points sex and smokes to neglected as they are factors in the data structure.



So based on the visual output of the matrix plot with the scatter plot, histogram and regresson coefficent we can tell there exists a positive association ship between the Birthweight of child and Gestation period. Which shows co-linearity between the two variables. This is clearly given as you an interpret this information as the weight of the baby when it is born is impacted based on how long it spends in the Gestation state absorbing more nutrition and growing.This can also be said about the weight of the mother, while it is not as strong in correlation but it can still definetly be included. While in the case of the age of the women and growth rate these may not have a linear effect or their effects are affected by other variables.

Distribution of Birthweight

Based on the density plot, the birthweight distribution deviates from normality. The curve is skewed to the right, indicating a higher proportion of babies with heavier birth weights. This is evident when compared to the overlaid normal distribution (red line), which shows a greater concentration of data points in the tails than predicted by the theoretical distribution. This suggests that factors beyond a simple normal distribution are influencing birth weight. It would be interesting to explore these factors further through correlation analysis or other statistical methods to understand better the underlying mechanisms contributing to birthweight variations.



Model Selection

```
##
## Call:
## lm(formula = bwt ~ (age + gest + weight + rate + smokesHeavy +
##     smokesLight + smokesNo + sex), data = BirthTrain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.14101 -0.26264 -0.00988  0.25556  1.80105
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.843083   0.581790  -6.606 1.66e-10 ***
## age          0.008262   0.004326   1.910  0.05704 .
## gest         0.023434   0.002045  11.458  < 2e-16 ***
## weight       0.008979   0.002410   3.726  0.00023 ***
## rate         1.711712   0.719352   2.380  0.01792 *
## smokesHeavy -0.167604   0.116365  -1.440  0.15075
## smokesLight -0.055795   0.112007  -0.498  0.61873
## smokesNo           NA         NA      NA       NA
## sexMale      0.083385   0.045429   1.835  0.06736 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.4085 on 319 degrees of freedom
## Multiple R-squared:  0.3631, Adjusted R-squared:  0.3491
## F-statistic: 25.98 on 7 and 319 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
##
## Model 1: bwt ~ age + gest + weight + rate + sex
## Model 2: bwt ~ (age + gest + weight + rate + smokesHeavy + smokesLight +
##     smokesNo + sex)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    321 53.620
## 2    319 53.243  2   0.37713 1.1298 0.3244
```

By looking at the p-value for all the variables, we can conclude that the Gestation period and pre-pregnancy weight are the strongest predictors with very low p-values with a significance level of 0, giving us tremendous statistical importance to our model. It is clear from this that the growth rate is essential, as the p-value is at a 1 significance level, while other factors have no use to us in the model. Now we dig into the impact the factor of smoking has on the child's weight when it is born if the mother is smoking. After the F-test, we get the p-value of 0.3244, which exceeds standard thresholds for statistical significance such as 0.05 or 0.01, suggesting that adding the smoking variables $smokesHeavy$ and $smokesLight$ to the model does not significantly improve its ability to predict birth weight, rather having sex of the child is a better factor to include. Adding more predictors to a model can increase the risk of overfitting, especially if those predictors do not significantly enhance model performance.

```
##
## Call:
## lm(formula = bwt ~ gest + weight + rate + sex, data = BirthTrain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12972 -0.27006 -0.00662  0.22602  1.82018
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.604385   0.560448  -6.431 4.57e-10 ***
## gest         0.023110   0.002036  11.352  < 2e-16 ***
## weight       0.010052   0.002372   4.239 2.94e-05 ***
## rate         1.867587   0.713321   2.618  0.00926 **
## sexMale      0.089895   0.045527   1.975  0.04917 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4104 on 322 degrees of freedom
## Multiple R-squared:  0.3513, Adjusted R-squared:  0.3432
## F-statistic: 43.59 on 4 and 322 DF,  p-value: < 2.2e-16
```

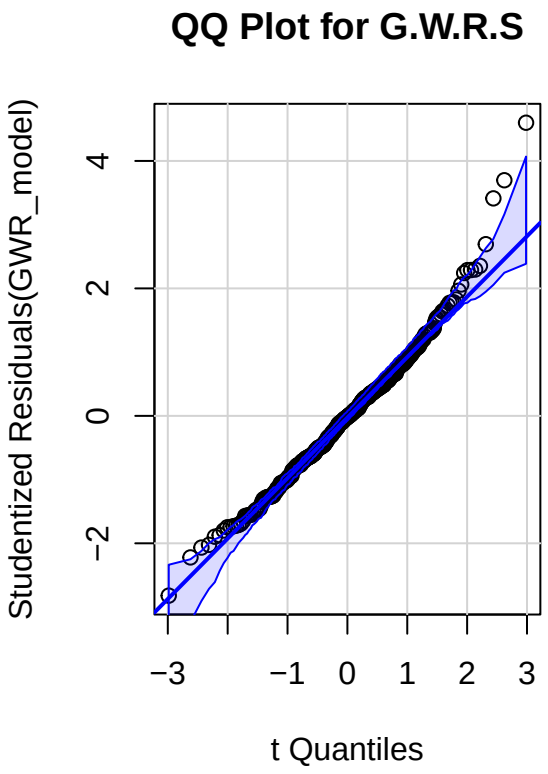
The linear regression output indicates a significant relationship between gestation period, maternal pre-pregnancy weight, and growth rate with birth weight, as all predictors show p-values well below the conventional significance level of 0.001 (with the exception of sex at 0.01). The model explains approximately 34.32% of the variance in birth weight (Multiple R-squared: 0.3513), which is moderately effective but suggests other factors might also be necessary.

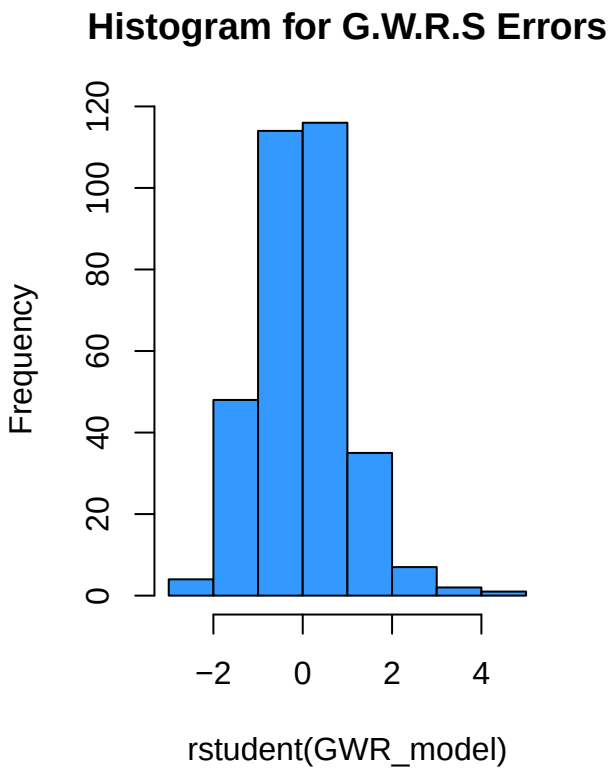
So now we will identify the model as the **G.W.R.S** model. The model is just nested in the full model as the model stands for the **gestation** period, maternal pre-pregnancy **weight**, growth rate, and birth **weight** of the baby. The abbreviation makes it easy to refer to later in the report and appreciates the parameters.

Model Checking and Validation

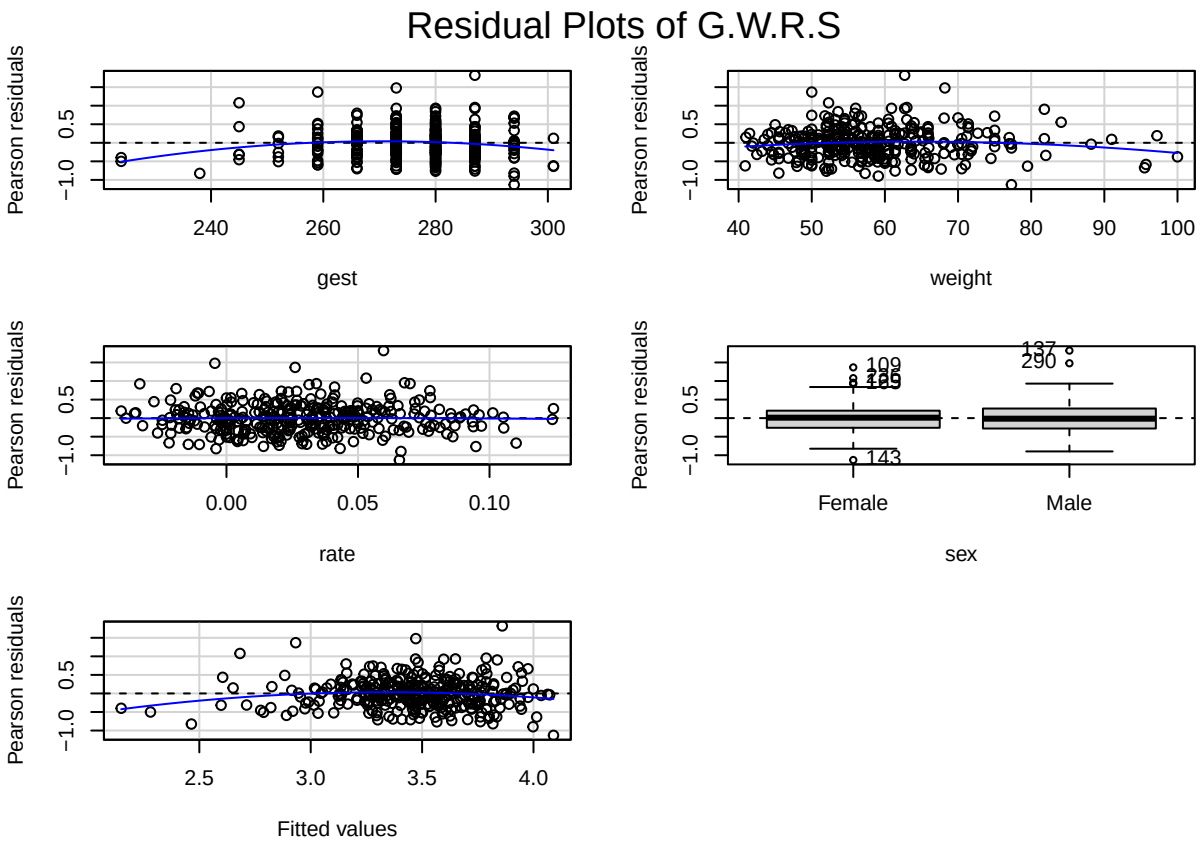
In this part we will check whether our selected models violate the **model assumptions on erros**:

- Zero mean: $E(\epsilon_i) = 0$ for all i
- Normality: $\epsilon_i \sim N(0, \sigma^2)$ for all i
- Uncorrelated: $Cov(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$
- Homoscedasticity (equal variance): $Var(\epsilon_i) = \sigma^2$ for all i

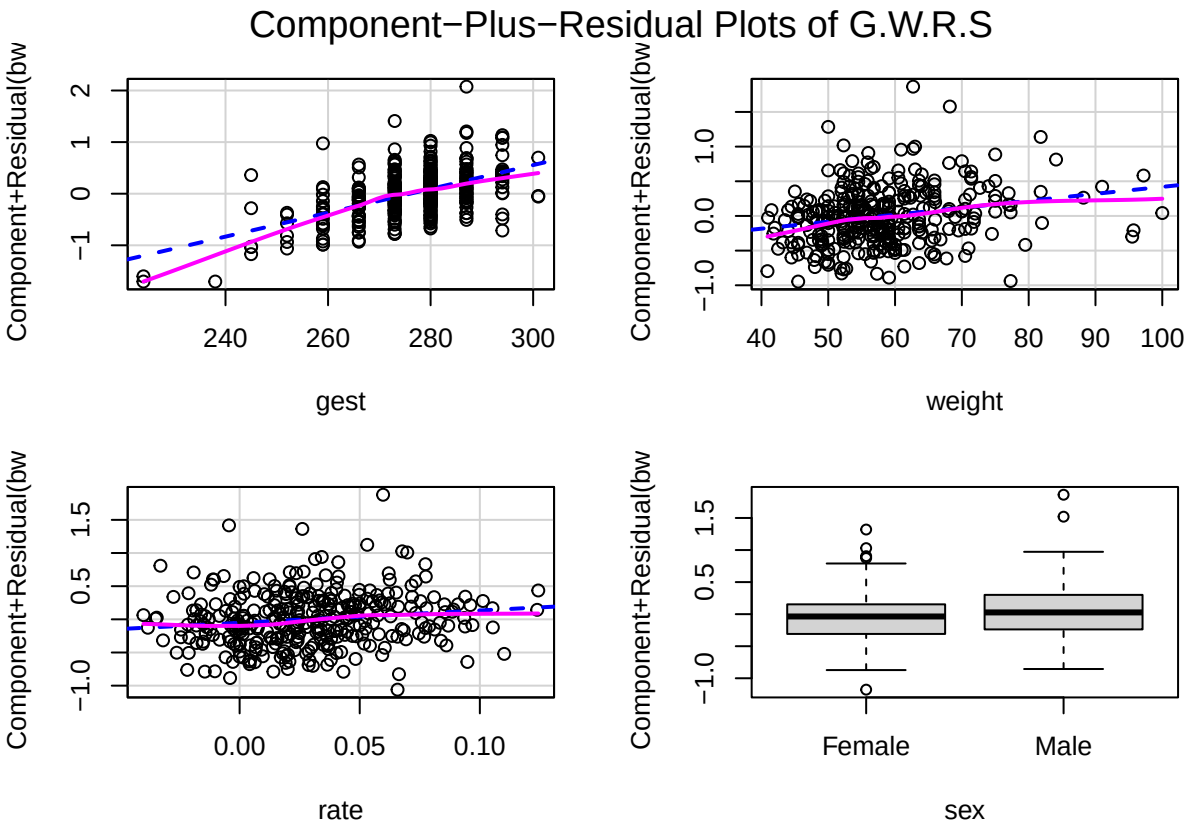




The Q-Q plot for the G.W.R.S model shows some deviation from the theoretical line, particularly in the upper tail, indicating possible issues with outliers or the distribution of residuals not being perfectly normal. The histogram of the residuals exhibits a slightly right-skewed distribution, as evident from the tail extending more towards positive values, which suggests that some model predictions are systematically underestimating the actual values. Both diagnostic plots highlight potential issues with the model fit, suggesting that the model might benefit from further investigation into outlier management, data transformation, or reconsidering the model assumptions to improve normality and reduce skewness.



The residual plots reveal inconsistent variance for the gestation period (gest) with increasing variance at higher values, indicating a need for variance-stabilizing transformations. The plots for weight and rate show scattered residuals with no apparent problematic trends, suggesting adequate model fit for these predictors.



The residual plots for the G.W.R.S model reveal various patterns across different predictors. For gestation (gest), the residuals show a slightly increasing trend as gestation increases, suggesting potential non-linearity or an inadequate model specification for this variable. The plots for weight and rate exhibit relatively random scatter, indicating no apparent issues with these predictors. However, a slight inconsistency in variance could be present, as seen in the spread of residuals across weight values. The sex plot highlights variances in residuals between females and males, with potential outliers identified in both categories, which may influence model performance. The residuals vs fitted values plot shows a lack of clear pattern, which generally indicates no significant issues with model fit across the range of predicted values. However, outliers or high leverage points suggest that the model might benefit from further robustness checks or modifications.

Predictive Ability

In this section, we perform prediction using our resulting model.

```
## MSE for G.W.R.S model:  136180.2
```

```
## MSE for Full Model:  134609.6
```

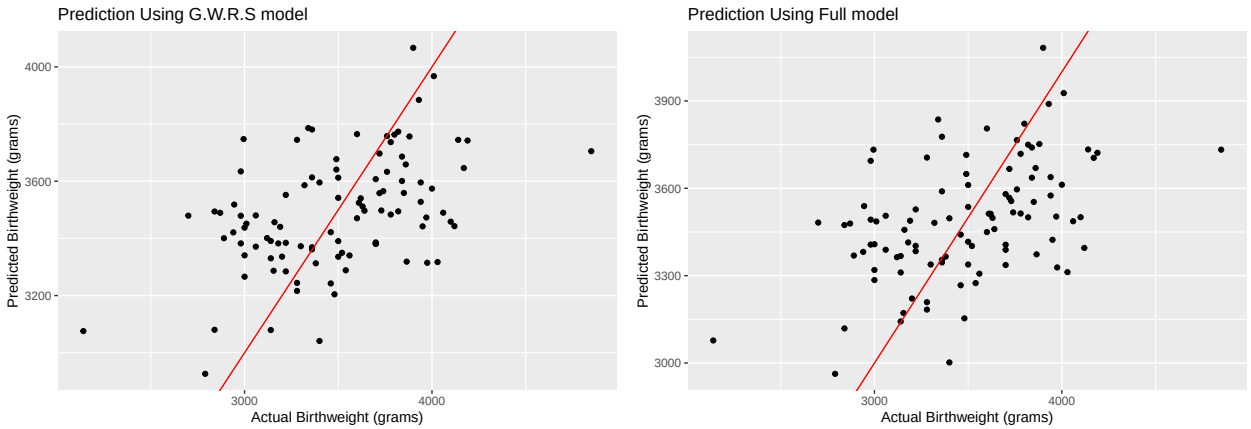
G.W.R.S Model: This model results in a Mean Squared Error (MSE) of 136180.2, which is a measure of the average squared difference between observed and predicted birth weights using variables such as gestation, weight, sex and rate.

Full Model: This includes additional predictors like age and various smoking statuses (smokes) alongside gestation, weight, and rate. The MSE for this model is 134609.6, which is lower than the G.W.R.S model's.

The lower MSE for the Full Model suggests that including more explanatory variables (age and smoking status) improves the model's predictive accuracy on the test dataset. This outcome implies that these additional factors are relevant in predicting birth weight, and their inclusion reduces prediction errors, enhancing the model's performance. However if we recall our previous findings the p-value for these additional factors which are in the full model imply they have little impact on the models ability to predict the weight of the baby when born. It is clear, to avoid over fitting it is best to go with G.W.R.S model although a higher MSE, its best to consider this as a trade off.

Load Test Data and Predict

This is the Prediction using the G.W.R.S model was built and it can be seen that the points obtained by matching the predicted values with the actual values are evenly distributed around the baseline y = x, indicating that our models perform relatively well. Simular can be said about the full model.



The R^2 values for the G.W.R.S Model and the Full Model are relatively similar, indicating that both models explain approximately 23% of the variance in the dependent variable. The adjusted R^2, which accounts for the number of predictors in the model, is lower for both models, with the Full Model experiencing a slightly more significant reduction, suggesting that the additional variables in the Full Model may not be contributing significantly to the explanation of the variance. This slight decrease in the $adjusted\_R^2$ in the Full Model compared to the G.W.R.S Model implies that the extra predictors may not justify their inclusion, considering their impact on the model's complexity and overall explanatory power. The Full Model, demonstrated a superior ability to predict birth weights accurately across all ranges.

```
## G.W.R.S Model - R^2: 0.2230859 Adjusted R^2: 0.1903737
```

```
## Full Model - R^2: 0.232046 Adjusted R^2: 0.1736147
```

Conclusions

Our extensive investigation into the variables influencing birth weight has yielded some compelling associations. These findings are critical for expectant mothers and healthcare providers to ensure the best possible outcome for babies. Our research has shown that closely monitoring the length of pregnancy is crucial as more extended gestation periods are positively associated with higher birth weights. Additionally, the importance of maternal health and nutrition cannot be overstated, as our study has highlighted a positive correlation between pre-pregnancy weight and childbirth weight. Early monitoring of pregnancies is also essential, as we found faster growth rates during the first trimester to be positively associated with higher birth weights. While maternal age, child sex, and smoking may be significant factors, our investigation found

no substantial correlation with birth weights. This suggests that focusing on the crucial factors of gestation length, maternal health and nutrition, and early monitoring of pregnancies can ensure the best possible outcome for mother and child.

Appendix

```
library(GGally)
library(ggplot2)
library(car)
library(knitr)


# Load training data
BirthTrain <- read.table("BirthTrain.txt", header = TRUE,sep = " ")
# Convert 'smokes' and 'sex' to factors
BirthTrain$smokes <- as.factor(BirthTrain$smokes)
BirthTrain$sex <- as.factor(BirthTrain$sex)
BirthTrain$bwt <- 0.001*(BirthTrain$bwt)


# Check the structure of the data
str(BirthTrain, useBytes = TRUE)
summary(BirthTrain)
par(mfrow=c(2,3))  # Set up the plotting area to have 2 rows and 3 columns

# Numerical variables
boxplot(BirthTrain$age, main="Age of Mother", ylab="Years",
        col = c("red"))
boxplot(BirthTrain$gest, main="Gestation Period", ylab="Days",
        col = c("steelblue"))
boxplot(BirthTrain$weight, main="Pre-pregnancy Weight", ylab="Kg",
        col = c("#009999"))
boxplot(BirthTrain$rate, main="Rate of Growth", ylab="Rate",
        col = c("#CC00CC"))
boxplot(BirthTrain$bwt, main="Birth Weight", ylab="Kg",
        col =c("#FFFF30"))

# Reset plot area for factor variables
par(mfrow=c(1,2))

# Factor variables
boxplot(bwt ~ sex, data=BirthTrain, main="Birth Weight by Sex", ylab="Kg",
        col =c("pink","blue"))
boxplot(bwt ~ smokes, data=BirthTrain, main="Birth Weight by Smoking Status",
        ylab="Kg",c("#CC6600"))
# Define color palette for sexes
color_palette <- c("Male" = "blue", "Female" = "pink")

# Select columns for the pair plot, excluding 'sex' and 'smokes'
selected_columns <- setdiff(names(BirthTrain), c("sex", "smokes"))
```

```r
# Generate the pair plot with the selected columns
p <- {ggpairs(BirthTrain, columns = selected_columns,
              aes(color = sex, alpha = 0.6),
              upper = list(continuous = wrap("cor", size = 2.5)),
              lower = list(continuous = "points"),
              title = "Pair Plot of BirthTrain Data",
              axisLabels = "show") +
              scale_color_manual(values = color_palette, name = "Gender")+
              labs(x = "Predictor Variables", y = "Response Variables")
}
print(p)
ggplot(BirthTrain, aes(x = bwt)) +
geom_density(aes(y = after_stat(density)), fill = "#FF99FF", color = "black") +
stat_function(fun = dnorm,
args = list(mean = mean(BirthTrain$bwt), sd = sd(BirthTrain$bwt)),color = "red") +
labs(title = "Distribution of Birthweight (Density)", x = "Birthweight (grams)")
# Create dummy variables for each category
BirthTrain$smokesHeavy <- ifelse(BirthTrain$smokes == "Heavy", 1, 0)
BirthTrain$smokesLight <- ifelse(BirthTrain$smokes == "Light", 1, 0)
BirthTrain$smokesNo <- ifelse(BirthTrain$smokes == "No", 1, 0)


# Remove the original 'smokes' variable
BirthTrain$smokes <- NULL
#head(BirthTrain[, c("smokesHeavy", "smokesLight", "smokesNo")])

# Fit a linear model including all smoking categories
model_all_smokes <- lm(bwt ~ (age + gest + weight + rate + smokesHeavy
                              + smokesLight + smokesNo + sex), data = BirthTrain)

# View the summary of the model
summary(model_all_smokes)

# Assume there's another model for comparison that perhaps does not include smoking variables
model_without_smokes <- lm(bwt ~ age + gest + weight + rate + sex,
                           data = BirthTrain)
# Use ANOVA to compare models
anova(model_without_smokes, model_all_smokes)

GWR_model <- lm(bwt ~ gest + weight + rate +sex, data = BirthTrain)

summary(GWR_model)
par(mfrow = c(1, 2))
qqPlot(GWR_model,main = 'QQ Plot for G.W.R.S',id=list(method="n", n=2, cex=1,
                                                      col=carPalette()[1], location="lr"))
hist(rstudent(GWR_model),main = 'Histogram for G.W.R.S Errors',col = "#3399ff")
#Skew to the right
a <-residualPlots(GWR_model,main='Residual Plots of G.W.R.S',tests=FALSE)
crPlots(GWR_model,main='Component-Plus-Residual Plots of G.W.R.S')
test_data <- read.table("BirthTest.txt", header = TRUE)

GWRS_model <- lm(bwt ~ gest + weight + rate +sex, data = test_data)
```

```r
predictions_best <- predict(GWRS_model , newdata = test_data)
full_model <- lm(bwt ~ age + gest + sex + smokes + weight + rate, data = test_data)
predictions_full <- predict(full_model, newdata = test_data)
mse_GWRS <- mean((test_data$bwt - predictions_best)^2)
cat("MSE for G.W.R.S model: ", mse_GWRS, "\n")
mse_full <- mean((test_data$bwt - predictions_full)^2)
cat("MSE for Full Model: ", mse_full, "\n")
# Load test data
BirthTest <- read.table("BirthTest.txt", header = TRUE)
BirthTest$smokes <- as.factor(BirthTest$smokes)
BirthTest$sex <- as.factor(BirthTest$sex)

# Predict birthweight
predictions <- predict(GWRS_model, newdata = BirthTest)
pred_full <- predict(full_model, newdata = BirthTest)
par(mar = c(4, 4, .1, .1))
# Compare actual and predicted birthweight
h<-ggplot(BirthTest, aes(x = bwt, y = predictions)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, color = "red") +
  labs(title = "Prediction Using G.W.R.S model", x = "Actual Birthweight (grams)",
       y = "Predicted Birthweight (grams)")

f<-ggplot(BirthTest, aes(x = bwt, y = pred_full)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, color = "red") +
  labs(title = "Prediction Using Full model", x = "Actual Birthweight (grams)",
       y = "Predicted Birthweight (grams)")
print(h)
print(f)
# Assuming your models are named GWR_model and full_model and were created using lm()

# Fit the models (Example)
GWR_model <- lm(bwt ~ gest + weight + rate +sex, data = test_data)
full_model <- lm(bwt ~ age + gest + sex + smokes + weight + rate, data = test_data)

# Get summary
summary_GWR <- summary(GWR_model)
summary_full <- summary(full_model)

# Extract R-squared and Adjusted R-squared
R2_GWR <- summary_GWR$r.squared
adjR2_GWR <- summary_GWR$adj.r.squared

R2_full <- summary_full$r.squared
adjR2_full <- summary_full$adj.r.squared

# Print the values
cat("G.W.R.S Model - R^2:", R2_GWR, "Adjusted R^2:", adjR2_GWR, "\n")
cat("Full Model - R^2:", R2_full, "Adjusted R^2:", adjR2_full, "\n")
```