The Impact that Smoking Makes on Birth Weights

Ansh Mujral Keenan Serrao

2024-10-10

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Contribution

Student 1 was responsible for the Introduction, Question 1, Question 3, and the formatting of the document. Student 2 was responsible for Questions 2, Questions 4, Advanced Analysis, and Conclusion. We both worked together on the code and double checked each others work.

Use of GPT

The use of gpt was limited. It was only used for the help of the scatter plot creation in the advanced analysis visualization.

1. Introduction

Birth weight is an important factor in newborn health, influencing health outcomes and possible treatments that may need to be undertaken. Studies have shown that low birth weight increases the risk of early deaths in newborns as well as negative impacts as they grow up. Smoking has been a large factor that is thought to influence birth weights.

In this analysis we are looking into the effect that maternal smoking has on the birth weights of babies. Our data comes from the Child Health and Development dataset based on pregnancies from Kaiser Health Plan in Oakland, Ca from 1960 to 1967. The main goal of this analysis is to compare the baby weights of smoking and non smoking mothers. We will be evaluating this through numerical, graphical, and incidence analysis. The report will address the difference in birth weight distributions, the percentage of low-birth babies, and multivariate regression on independent variable.

The data in this analysis consists of 1236 babies, all of which are boys, single births, and lived at least 28 days. This ensures consistency between individual data points and allows us to minimize confounding variables or external factors that could influence the results that we provide.

For each baby present, we have various pieces of information available, and can be summarized in the following list: Although our data has whole numbers which are discrete, weights, time-series (gestation), height are all continuous data in the real world.

- bwt: birth weight (in ounces), **Numerical Continuous**, normal distribution
- gestation: length of gestation (in days), **Numerical Continuous**, clustered around 280 days or full term
- parity: binary indicator for a first pregnancy (0 = first pregnancy), **Categorical Ordinal**,
- age: mother's age (in years), **Numerical Continuous**,
- height: mother's height (in inches), **Numerical Continuous**, near normal distribution
- weight: mother's weight (in pounds), **Numerical Continuous**, near normal distribution
- smoke: binary indicator for whether the mother smokes (0 = no), **Categorical Nominal**

2. Analysis

2.1 Question 1

2.1.1 Method The goal of this question is to understand the data that we are presented with, as well as clean our data by removing things such as null values and outliers. We can do this by first loading our data into a dataframe, and running appropriate functions to generate histograms. We can then filter our data for null values using the na.omit() function, as well as a custom function for removing outliers.

```r
babies <- read.csv("~/Desktop/babies.txt", sep="")
str(babies)
```

2.1.2 Analysis

```
## ’data.frame':    1236 obs. of  7 variables:
##  $ bwt      : int  120 113 128 123 108 136 138 132 120 143 ...
##  $ gestation: int  284 282 279 999 282 286 244 245 289 299 ...
##  $ parity   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ age      : int  27 33 28 36 23 25 33 23 25 30 ...
##  $ height   : int  62 64 64 69 67 62 62 65 62 66 ...
##  $ weight   : int  100 135 115 190 125 93 178 140 125 136 ...
##  $ smoke    : int  0 0 1 0 1 0 0 0 0 1 ...
```

```r
summary(babies)
```

```
##       bwt          gestation          parity            age
##  Min.   : 55.0   Min.   :148.0   Min.   :0.0000   Min.   :15.00
##  1st Qu.:108.8   1st Qu.:272.0   1st Qu.:0.0000   1st Qu.:23.00
##  Median :120.0   Median :280.0   Median :0.0000   Median :26.00
##  Mean   :119.6   Mean   :286.9   Mean   :0.2549   Mean   :27.37
##  3rd Qu.:131.0   3rd Qu.:288.0   3rd Qu.:1.0000   3rd Qu.:31.00
##  Max.   :176.0   Max.   :999.0   Max.   :1.0000   Max.   :99.00
##      height          weight         smoke
##  Min.   :53.00   Min.   : 87   Min.   :0.0000
##  1st Qu.:62.00   1st Qu.:115   1st Qu.:0.0000
##  Median :64.00   Median :126   Median :0.0000
##  Mean   :64.67   Mean   :154   Mean   :0.4644
##  3rd Qu.:66.00   3rd Qu.:140   3rd Qu.:1.0000
##  Max.   :99.00   Max.   :999   Max.   :9.0000
```

```r
bwt_description <- ggplot(babies, aes(x=bwt)) + geom_histogram()
gestation_description <- ggplot(babies, aes(x=gestation)) + geom_histogram()
parity_description <- ggplot(babies, aes(x=parity)) + geom_histogram()
age_description <- ggplot(babies, aes(x=age)) + geom_histogram()
height_description <- ggplot(babies, aes(x=height)) + geom_histogram()
weight_description <- ggplot(babies, aes(x=weight)) + geom_histogram()
smoke_description <- ggplot(babies, aes(x=smoke)) + geom_histogram()
```

```r
cleaned_df <- na.omit(babies)

remove_outliers <- function(x) {
  Q1 <- quantile(x, 0.25, na.rm = TRUE)
  Q3 <- quantile(x, 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  x[x < (Q1 - 1.5 * IQR) | x > (Q3 + 1.5 * IQR)] <- NA
  return(x)
}

cleaned_df <- cleaned_df %>%
  mutate(across(where(is.numeric), remove_outliers)) %>%
  na.omit()
```

2.1.3 Conclusion Based on our histograms, we can see that the body weight column is **approximately normally distributed**, the gestation column looks symmetrical with some extreme outliers, the age column has a slight right skew with some outliers, the height and weight columns look **relatively symmetrical** but both have outliers. While a lot of the columns appear to be normally distributed, we cannot say that this is a simple random sample as it is specific to a certain location and time period. While this population may be able to be generalized to the rest of Oakland in the 1960s, it may not be fully representative of other time periods or locations in the US.

2.2 Question 2

2.2.1 Method The goal of this question is to summarize numerically the birth weights of babies that are born to mothers that smoked versus mothers that did not smoke. To do this, we split the data into two groups, one group being smokers and the other being non smokers.

```r
nonsmoker_df <- cleaned_df %>% filter(smoke == 0)
smoker_df <- cleaned_df %>% filter(smoke == 1)
```

Then for each group we calculated the summary statistics for birth weights. We calculated the the minimum value, maximum value, mean, median, quartile 1, quartile 2, quartile 3, and the standard deviation. The way we calculated it is shown more below in 2.2.2 Analysis.

```r
nonsmoker_min_bwt <- min(nonsmoker_df$bwt)
nonsmoker_max_bwt <- max(nonsmoker_df$bwt)
smoker_min_bwt <- min(smoker_df$bwt)
smoker_max_bwt <- max(smoker_df$bwt)

# mean
smoker_mean_bwt <- mean(smoker_df$bwt)
nonsmoker_mean_bwt <- mean(nonsmoker_df$bwt)

# median
smoker_median_bwt <- median(smoker_df$bwt)
nonsmoker_median_bwt <- median(nonsmoker_df$bwt)

# quartiles
smoker_q1_bwt <- quantile(smoker_df$bwt, 0.25)
smoker_q2_bwt <- quantile(smoker_df$bwt, 0.50)
smoker_q3_bwt <- quantile(smoker_df$bwt, 0.75)
nonsmoker_q1_bwt <- quantile(nonsmoker_df$bwt, 0.25)
nonsmoker_q2_bwt <- quantile(nonsmoker_df$bwt, 0.50)
nonsmoker_q3_bwt <- quantile(nonsmoker_df$bwt, 0.75)

# sd
smoker_std_bwt <- sd(smoker_df$bwt)
nonsmoker_std_bwt <- sd(nonsmoker_df$bwt)
```

2.2.2 Analysis Here is a table of values to show the numerical comparison between the two.

Table 1: Smoker vs Non-Smoker Summary

Col1	Smoker	Non Smoker
Min	78	78
Max	163	164
Mean	114.23	123.42
Median	115	124
Q1	103	113
Q2	115	124
Q3	126	133
SD	16.77	15.12

2.2.3 Conclusion From the results it shows that the mean birth weight of babies born to non-smokers **(123.42 oz)** is higher compared to the mean birth weight of babies born to smokers **(115.23 oz)**. It also shows that the median birth weight of babies born to non smokers **(124 oz)** is higher than median birth weight of babies born to smokers **(115 oz)**. We can also see that both distributions have no skewness from the mean and median values. **Both values are approximately equal meaning that the data is symmetrical**. From the standard deviations we can see that the distribution is slightly more variable for the babies born to smokers (16.77) compared to non smokers(15.16). **From our results and findings it is easy to see that lower birth weights are associated with mothers that smoke during pregnancy.**

2.3 Question 3

2.3.1 Method The main goal for this question is to summarize the 2 distributions created above in a graphical manner. To do so, we will plot the bwt column of each of the dataframes from the previous question using histograms and compare the distributions.

```
smoker_bwt_description <- ggplot(smoker_df, aes(x=bwt)) + geom_histogram()
nonsmoker_bwt_description <- ggplot(nonsmoker_df, aes(x=bwt)) + geom_histogram()
```

2.3.2 Analysis

2.3.3 Conclusion Looking at the 2 histograms, we can see that the distribution of body weight for smokers is shifted to the left compared to the distribution of body weight for non-smokers. **The shift in distribution can indicate a difference in the underlying patterns influenced by whether or not a pregnant woman smokes**. This can lead us to believe that **smoking does have some impact on the body weight of a baby** as the distributions are of similar shapes, but are centered around different values.

2.4 Question 4

2.4.1 Method For question 4, the goal is to find out the percentage of low-birth-weight babies which are babies weighing under 100 ounces. To find this out, we filter the data to count the number of babies that weight under 100 pounds in both groups and then we calculate the percentage using the formula (value/total * 100)

2.4.2 Analysis When the low birth weight is defined at 100

```
low_bwt_smoker <- mean(smoker_df$bwt < 100) * 100
low_bwt_nonsmoker <- mean(nonsmoker_df$bwt < 100) * 100
```

When low birth weight is defined at a different threshold both greater and lower than 100

```
low_bwt_smoker_90 <- mean(smoker_df$bwt < 90) * 100
low_bwt_nonsmoker_90 <- mean(nonsmoker_df$bwt < 90) * 100

# Raising the threshold to 110 ounces
low_bwt_smoker_110 <- mean(smoker_df$bwt < 110) * 100
low_bwt_nonsmoker_110 <- mean(nonsmoker_df$bwt < 110) * 100

# Output the results
list(
  low_bwt_smoker_90 = low_bwt_smoker_90,
  low_bwt_nonsmoker_90 = low_bwt_nonsmoker_90,
  low_bwt_smoker_110 = low_bwt_smoker_110,
  low_bwt_nonsmoker_110 = low_bwt_nonsmoker_110
)
```

```
## $low_bwt_smoker_90
## [1] 5.555556
```

```
##
## $low_bwt_nonsmoker_90
## [1] 1.72144
##
## $low_bwt_smoker_110
## [1] 39.13043
##
## $low_bwt_nonsmoker_110
## [1] 15.64945
```

2.4.3 Conclusion The incidence of low-birth-weight babies is higher for smokers at 18.26% compared to that of non-smokers at 5.32%. We find this to be a **huge difference and conclude that mothers that smoke during pregnancy is a very big factor for low birth weight**.

In an initial analysis, low birth weight was defined as less than 100 ounces. **This corresponded to 18.36% and 5.32% of babies born having a low body mass at the threshold for smokers and non-smoker, respectively**. But the threshold that you choose makes a big difference in whether these groups differ or not. If we set the threshold value lower (e.g. 90 ounces), fewer babies will qualify as low birth weight, and both groups would demonstrate a decline in incidence rates for being classified such. Conversely, increasing the threshold (e.g., to 110 ounces) will lead more babies in both scenarios being classified as low birth-weight.

The more we adjust the threshold, the incidence between smokers and nonsmokers looks different:

Rates were reduced in both groups with the threshold at 90 ounces, but the relative difference between smokers and non-smokers remain significant because smoking results in lower birth weights. That said, **with fewer babies in the "low birth weight" camp overall this absolute difference between groups may appear smaller**.

If we raise it to 110 ounces, the incidence rate will increase along with the number of babies in these two groups. But more babies in the smoker group will probably be labeled as low-birth-weight infants, though this will only serve to widen the gap between smokers and nonsmokers. These **results enhance the influence of smoking on child's early weight outcomes, allowing inferring that smoking is bad in the neonatal process**.

Adjusting the cutoff for low birth weight impacts both the overall occurrence rates and the contrast between individuals who smoke and those who do not. Decreasing the cutoff point leads to a decrease in the number of infants identified as having low birth weight, which may lessen the observed disparity between the two groups. Increasing the threshold also raises the occurrence in both groups while emphasizing the difference between smokers and non-smokers even more. **This study shows that no matter the cutoff point, smoking is still a major factor in lower birth weights; however, the threshold chosen can impact how we perceive the seriousness of this risk.**

2.5 Question 5

2.5.1 Method Here, we will rely on the variability and outcomes of our previous questions, namely summary statistics, histograms, and incidence rates. In this analysis, we will gauge the benefits and drawbacks of each method by examining how well they reflect smoking exposure in relation to birth weight.

```


nonsmoker_min_bwt
```

2.5.2 Analysis

```
## [1] 78
```

```
nonsmoker_max_bwt
```

```
## [1] 164
```

```
nonsmoker_mean_bwt
```

```
## [1] 123.4241
```

```
nonsmoker_median_bwt
```

```
## [1] 124
```

```
nonsmoker_q1_bwt
```

```
## 25%
## 113
```

```
nonsmoker_q2_bwt
```

```
## 50%
## 124
```

```
nonsmoker_q3_bwt
```

```
## 75%
## 133
```

```
nonsmoker_std_bwt
```

```
## [1] 15.15593
```

```
smoker_min_bwt
```

```
## [1] 78
```

```
smoker_max_bwt
```

```
## [1] 163
```

```
smoker_mean_bwt
```

```
## [1] 115.2343
```

```
smoker_median_bwt
```

```
## [1] 115
```

```
smoker_q1_bwt
```

```
## 25%
## 103
```

```
smoker_q2_bwt
```

```
## 50%
## 115
```

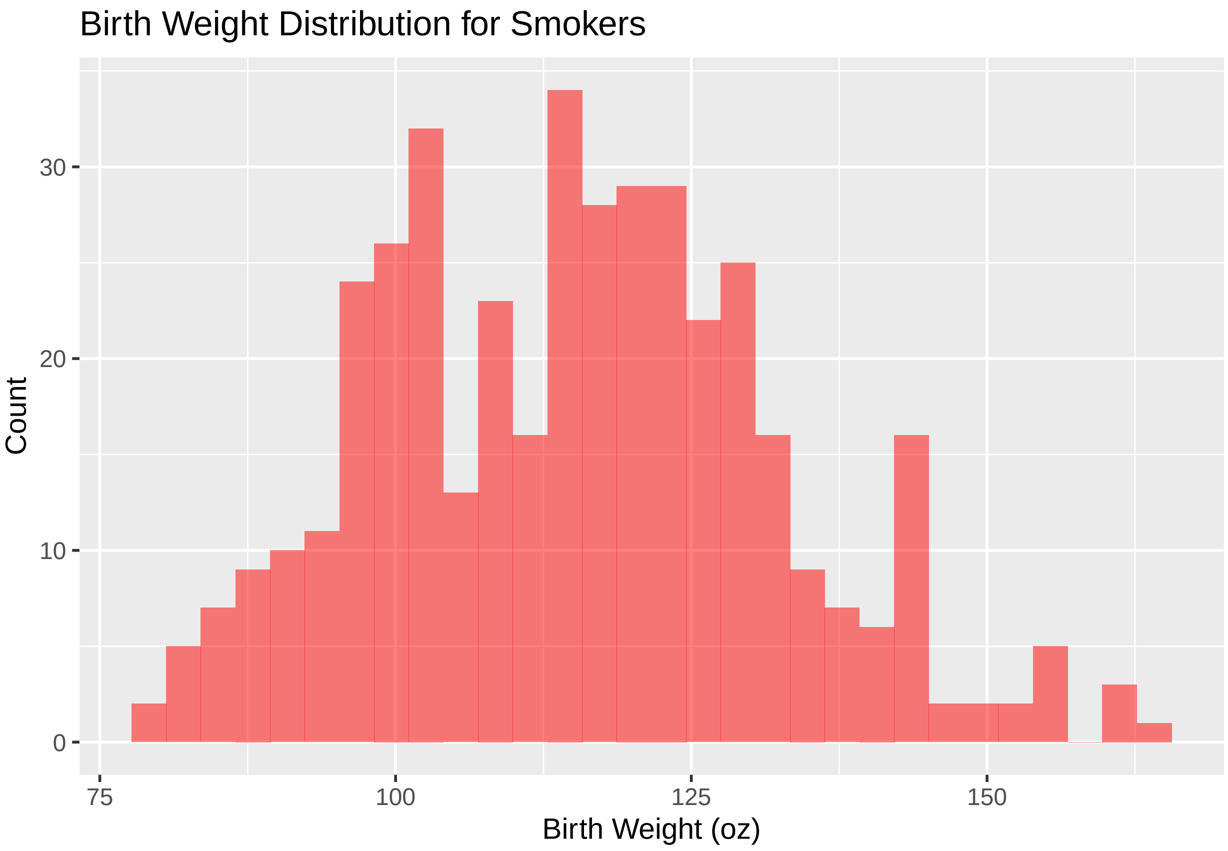
```
smoker_q3_bwt
```

```
## 75%
## 126
```

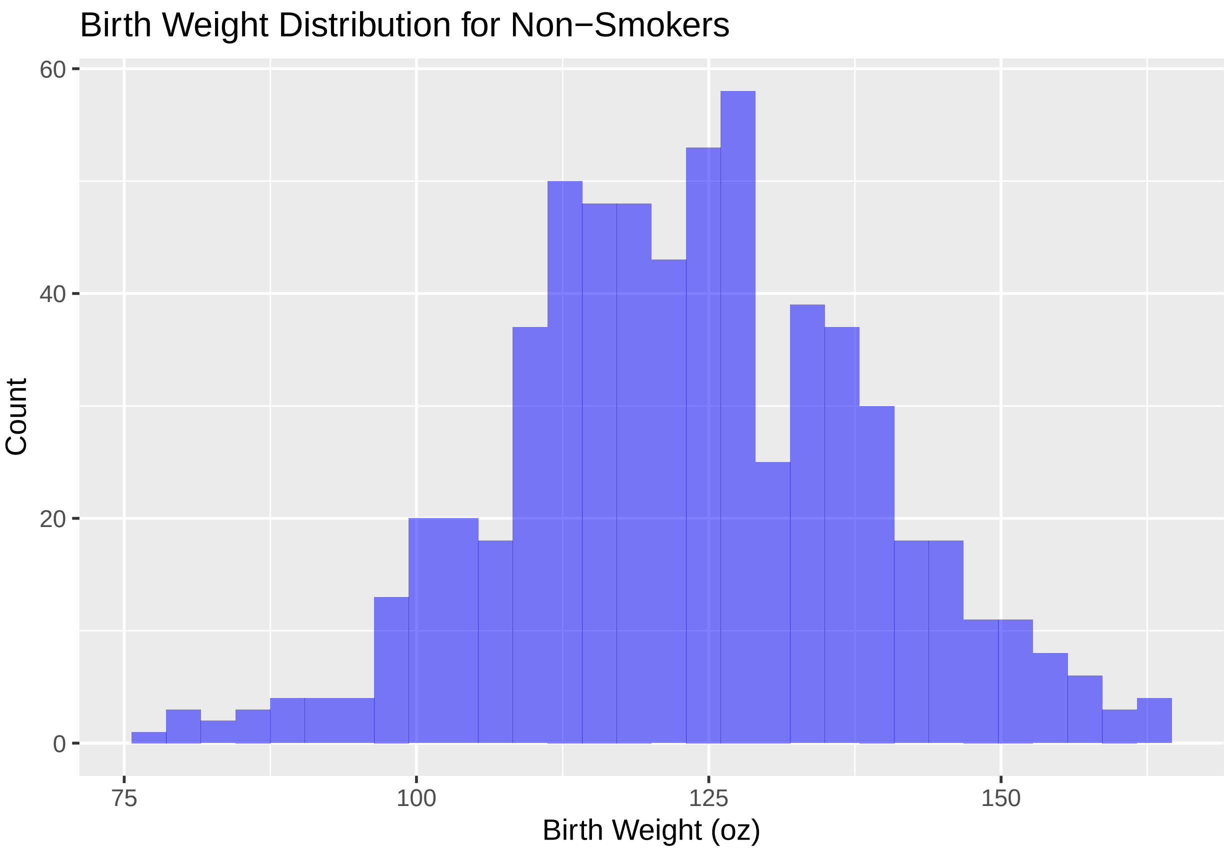
```
smoker_std_bwt
```

```
## [1] 16.76733
```

```
# Graphical Comparison
# Histograms showing birth weight distributions for smokers and non-smokers
ggplot(smoker_df, aes(x = bwt)) + geom_histogram(bins = 30, fill = "red", alpha = 0.5) +
  labs(title = "Birth Weight Distribution for Smokers", x = "Birth Weight (oz)", y = "Count")
```



```r
ggplot(nonsmoker_df, aes(x = bwt)) + geom_histogram(bins = 30, fill = "blue", alpha = 0.5) +
  labs(title = "Birth Weight Distribution for Non-Smokers", x = "Birth Weight (oz)", y = "Count")
```



```r
# Incidence-Based Comparison
# Proportion of low-birth-weight babies (under 100 oz) for smokers and non-smokers
low_bwt_smoker
```

```
## [1] 18.35749
```

```r
low_bwt_nonsmoker
```

```
## [1] 5.320814
```

2.5.3 Conclusion

Based on the analysis:

Numerical Comparisons: The smoker's summary statistics tend to have lower birth weights than non-smokers. **This is evidenced by the mean and median, which indicate that smokers' babies weigh on average 7-8 ounces less**. However, the variation is slightly higher among smokers, indicating more spread in birth weights. This method is more accurate, but it does not show the shape or outliers.

Graphical Comparisons: Smokers' histograms offer a visual view of the distribution of birth weights. We can see that **non-smoker babies tend to have higher birth weights, while smokers' babies show a shift toward lower weights**. This method is useful for identifying overall trends but can be subjective and lacks precision.

Incidence Comparisons: The percentage of low-birth-weight babies is higher for smokers, with 18.36% compared to 5.32% for non-smokers, for babies weighing under 100 ounces. This approach is easy to interpret and useful for making practical recommendations, but **it simplifies the data by focusing on a cutoff**.

Numerical comparisons are precise but may limit distribution insights. Graphical comparisons provide a good visual overview of trends but need to be backed by precise data. Incidence comparisons offer practical insights but may oversimplify the data. **Combining all three approaches gives a comprehensive understanding of how smoking during pregnancy affects birth weight**.

3. Advanced Analysis

3.1 Method Our current question only tackle the idea of how smoking affects birth weights. Our data also includes age and gestation period as well. With this information we can perform multivariate regression analysis to see how gestation, age, and smoking status all affect baby weights. To do this we will implement a multiple linear regression model where the independent variables are smoking status, mothers age, and gestation and birth wight is the dependent variable.

```
model <- lm(bwt ~ smoke + age + gestation, data = cleaned_df)
summary(model)
```

3.1 Analysis

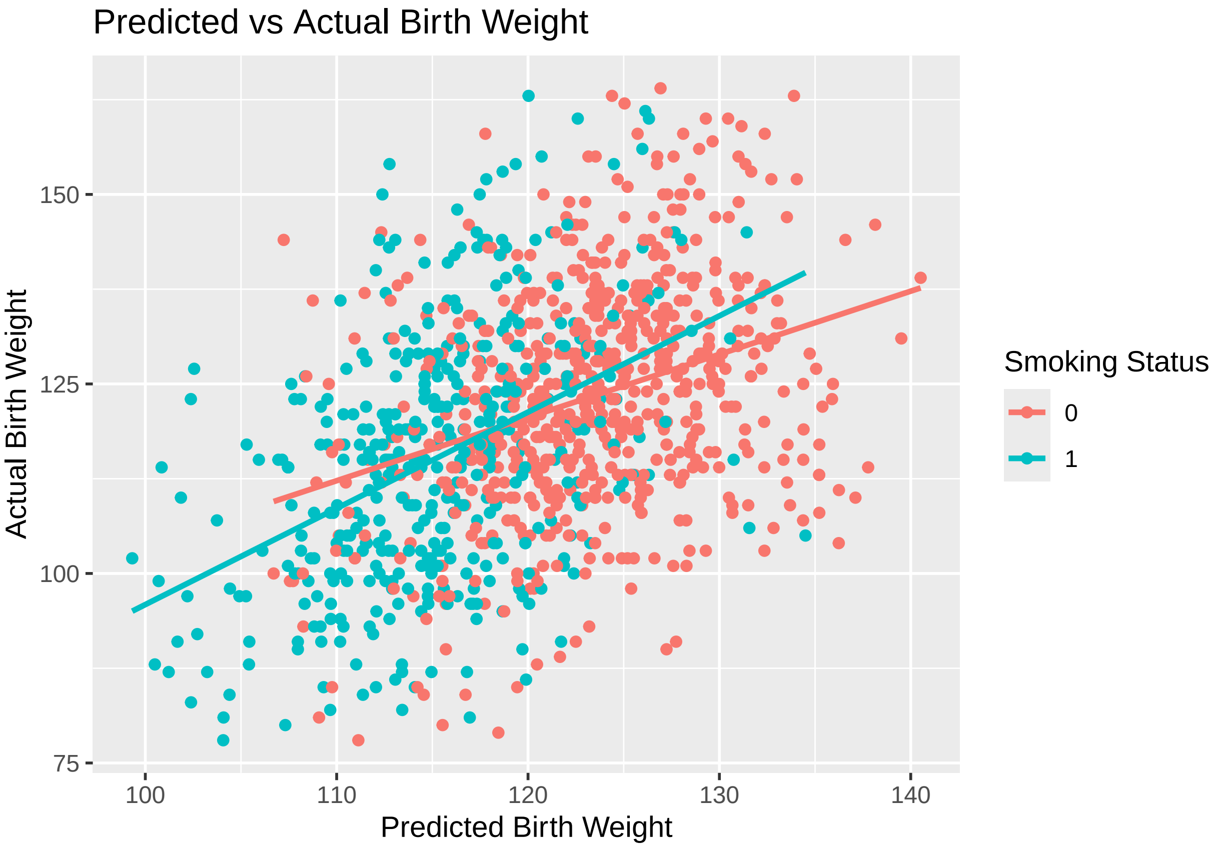
```
##
## Call:
## lm(formula = bwt ~ smoke + age + gestation, data = cleaned_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.451 -10.238  -0.227   9.400  42.971
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -24.2190    11.5859  -2.090   0.0368 *
## smoke        -7.2175     0.9328  -7.737 2.38e-14 ***
## age           0.1725     0.0796   2.167   0.0304 *
## gestation     0.5092     0.0401  12.699  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 14.72 on 1049 degrees of freedom
## Multiple R-squared:  0.1871, Adjusted R-squared:  0.1847
## F-statistic: 80.46 on 3 and 1049 DF,  p-value: < 2.2e-16
```

In this analysis you can see the coefficients for each predictor and how significant they are in terms of contributing to birth weight.

```
cleaned_df$predicted_bwt <- predict(model)

# Plot actual vs predicted birth weight, color by smoking status
ggplot(cleaned_df, aes(x = predicted_bwt, y = bwt, color = as.factor(smoke))) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Predicted vs Actual Birth Weight",
       x = "Predicted Birth Weight",
       y = "Actual Birth Weight",
       color = "Smoking Status")
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



3.1 Conclusion From the regression analysis you can see that the p value is under 0.05 and it indicates that smoke, age, and gestation are highly significant predictor of birth weight. Each coefficient is statistically significant at the 5% significance level showing how impactful a mothers age, their smoking status, and gestation period matters towards birth weight.

Also from the graph of predicted birth weight vs. actual birth weight, you can see the trend that non smokers predicted and actual birth weights are much high than that of smokers. The y-intercept where smoking status is 1 is significantly lower and the variability is much higher seen by the spread of the blue dots on the scatter plot.

Conclusion

This analysis aimed to study the influence of maternal smoking on babies' birth weights, utilizing both simple comparisons and an advanced multivariable regression model to account for the impact of smoking in combination with the mother's age and gestation period.

The most significant findings can be summarized as follows: Simple numerical and visualized comparisons: Babies born to mothers who smoked were significantly disadvantaged in terms of birth weight. **Specifically, their mean birth weight was about 115.23 ounces, while newborns of non-smoking mothers averaged 123.42 ounces. Moreover, smokers presented with higher variation.** Additionally it was shown through the visualizations that the center and distribution of the babies born to mothers that smoked was shifted to the left.

Incidence-based analysis: The percentage of babies with low birth weight—below 100 ounces—was significantly higher among smokers (18.36%) compared to non-smokers (5.32%). Hence, smoking is a significant risk factor for low birth weight, a crucial health condition.

Multivariable regression findings: Smoking during pregnancy has an independent, statistically significant effect on reducing birth weight. While controlling for mother's age and gestation period, smoking was **associated with a 7.22-ounce reduction in birth weight.** Gestation period and occasionally the mother's age also played an important role, with longer gestation and younger mothers yielding heavier babies.

Implications: The harmful effect of smoking highlights the importance of public health recommendations advising pregnant women to avoid smoking. Prenatal care ensuring optimal gestational duration is emphasized, as extended pregnancy allows for the safer delivery of a heavier newborn. Moreover, the mother's age is a small but significant predictor of birth weight. Thus, older mothers might observe a negligible increase in newborn weight. Conclusively, this report affirms that smoking is a major predictor of reduced birth weight in newborns, even when considering other factors like the mother's age and gestation. Therefore, smoking during pregnancy should be targeted by public health interventions.

We conclude that smoking leads to lower baby weights compared to that of non smokers.