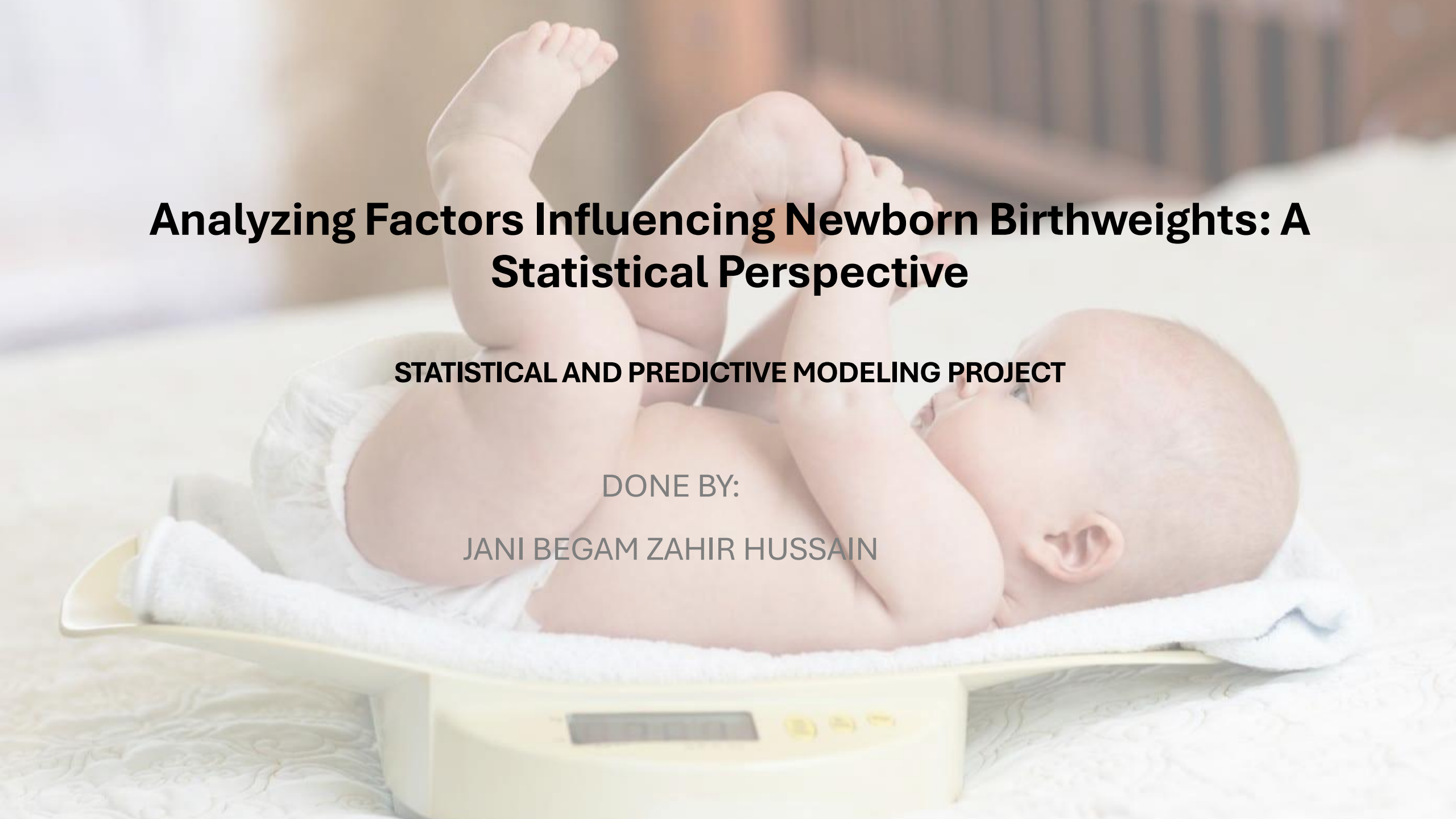


Analyzing Factors Influencing Newborn Birthweights: A Statistical Perspective

STATISTICAL AND PREDICTIVE MODELING PROJECT

DONE BY:

JANI BEGAM ZAHIR HUSSAIN

DESCRIPTION OF THE RESEARCH REQUIREMENTS

Introduction: Birth weight is a crucial parameter of a newborn's health. Various factors may affect it, including the mother's lifestyle, health, and demographics.

Mr. John Hughes created a dataset, Birthweight.csv, to investigate these connections, which includes data on 1174 infants and their moms.

Objective: Use a basic linear regression model to examine the effect of mother smoking on birth weight.

- Use a multivariate regression model to examine the combined impact of additional variables on birth weight, including gestation length, mother's region, age, height, and weight.

About the Dataset: The dataset includes 1174 records and 7 key features Like Gestation, Region, Age, Height, Weight, Smoke, BWT

Problem Statement and Solution: Mr. John Hughes wants to investigate the variables affecting birth weight using regression analysis. This involves examining the combined impact of all maternal covariates (multivariate regression) as well as the effect of the mother's smoking habit (basic linear regression).

Proposed Solution: A simple linear regression approach to investigate the connection between birth weight and smoking. A model of multivariate regression is used to assess the combined impact of pregnancy length, region, age, height, weight, and smoking on birth weight.

Steps in conducting the analysis:

Step-1 : Descriptive Statistics

Step-2 : Data Visualization

Step-3 : Conducting T- Test

Step-4 : Simple Linear Regression

Step-5 : Multiple Linear Regression

Step-6 : Conclusion

BASIC STATISTICS

The basic statistics for the data

Variable	Mean	SD	Min	Max
bwt	3584.71	337.86	3061.8	4054.1
gestation	276.67	19.12	244	299
age	28.44	3.88	24	34
height	161.89	5.13	157	170
weight	57.81	11.26	42.2	80.7
smoke	0.33	0.50	0	1

Observations

- **Birth Weight (bwt):**
  - The mean birth weight is 3584.71 grams, indicating the average weight of babies in the sample.
  - The SD is 337.86 grams, suggesting moderate variability in birth weights.
  - The minimum value (3061.8 grams) and maximum value (4054.1 grams) show the range of weights in the dataset, with the lightest and heaviest recorded weights.

- **Gestation:**
  - The mean gestational age is 276.67 days, representing the average duration of pregnancies.
  - The SD is 19.12 days, showing some variation in gestational lengths.
  - The minimum gestation (244 days) is notably shorter, while the maximum gestation (299 days) indicates full-term or extended pregnancies.

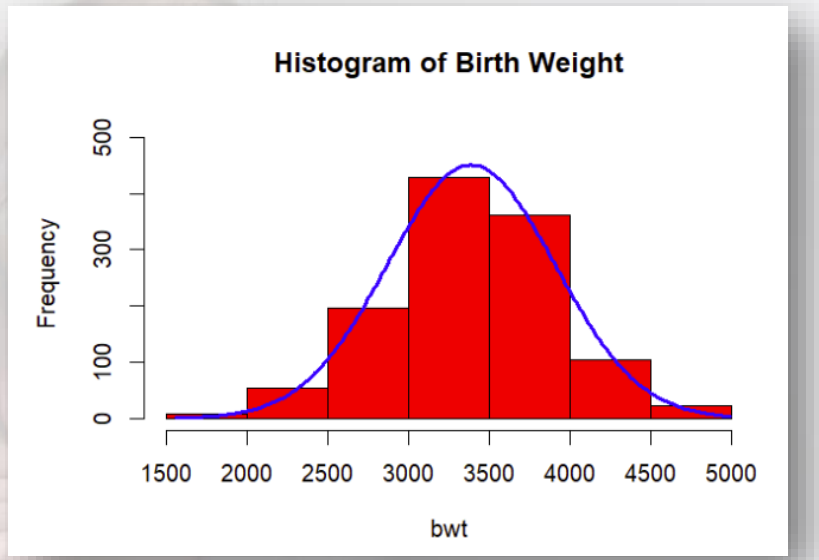
INTERPRETATION OF BASIC STATISTICS

- **Age:**
  - The mean age is 28.44 years, reflecting the average age of the individuals.
  - The SD is 3.88 years, indicating relatively low age variability.
  - The youngest individual is 24 years, and the oldest is 34 years, showing the age range of participants.
- **Height:**
  - The mean height is 161.89 cm, the average stature of individuals in the dataset.
  - The SD is 5.13 cm, indicating a slight variation in heights.
  - Heights range from a minimum of 157 cm to a maximum of 170 cm, representing shorter and taller individuals.
- **Weight:**
  - The mean weight is 57.81 kg, indicating the average body weight.
  - The SD is 11.26 kg, showing moderate variability in weight distribution.
  - The minimum weight is 42.2 kg, and the maximum weight is 80.7 kg, revealing significant differences between the lightest and heaviest individuals.
- **Smoking Status (smoke):**
  - The mean is 0.33, suggesting 33% of individuals are smokers (coded as 1).
  - The SD is 0.50, reflecting an even split between smokers and non-smokers.
  - The binary variable ranges from 0 (non-smoker) to 1 (smoker), defining the smoking status.

HISTOGRAM

Findings from the Histogram and Normal Curve:

- **Distribution Shape:**
  - The histogram shows that the birth weight data is approximately normally distributed. The bars are roughly symmetric around the center, with a peak around 3000-3500 grams.
- **Mean and Standard Deviation:**
  - The normal curve overlaid on the histogram confirms the normal distribution. The mean birth weight appears to be around 3000-3500 grams, and the standard deviation seems to be about 500 grams (based on the spread of the curve).
- **Outliers:**
  - There are a few bars with lower frequencies at the extremes (below 2000 grams and above 4500 grams), which could be considered potential outliers. However, without more context about the data collection process and criteria for outliers, it's difficult to definitively label them as such.
- **Interpretation:**
  - The histogram and normal curve suggest that birth weights in the dataset are distributed around a central value (the mean) with a certain degree of variability (the standard deviation). This is a common pattern in many natural phenomena, and it's useful for understanding the typical range of birth weights and identifying potential outliers or unusual cases.



HISTOGRAM OF THE DEPENDENT VARIABLE(BWT)

Explanation:

Data Loading:

- The code assumes you have a dataset named data with a variable bwt representing birth weight.

Histogram Creation:

- hist(x, breaks=10, col="red", xlab="bwt", main="Histogram of Birth Weight")
- This line creates a histogram of the birth weight data (x) with 10 bins (controlled by breaks=10).
- The bars are colored red (col="red"), the x-axis is labeled "bwt", and the title is "Histogram of Birth Weight".

Normal Curve Overlay:

- xfit<-seq(min(x),max(x),length=40)

  Creates a sequence of 40 equally spaced points from the minimum to the maximum birth weight values.
- yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))

  Calculates the density of the normal distribution at each point in xfit, using the mean and standard deviation of the birth weight data.
- yfit <- yfit*diff(h$mids[1:2])*length(x)

  Scales the density values to match the histogram's y-axis.
- lines(xfit, yfit, col="blue", lwd=2)

  Plots the scaled normal curve on top of the histogram as a blue line.

HYPOTHESES RELATED TO THE T-TEST

HYPOTHESES:

- **Null Hypothesis (H0):** The true mean birthweight is equal to 3400 grams (μ=3400).
- **Alternative Hypothesis (Ha):** The true mean birthweight is not equal to 3400 grams (μ≠3400).
- This is a two-tailed test because we are testing for a difference in either direction.

RESULTS AND EXPLANATION OF THE T-TEST:

- **t-value:** -0.9
  - The t-value measures how far the sample mean deviates from the hypothesized population mean (3400 grams) in terms of standard errors. A small t-value suggests little difference between the sample mean and the hypothesized mean.
- **Degrees of Freedom (df):** 1173
  - This is based on the sample size (n−1) and reflects the number of independent observations.
- **p-value:** 0.4
  - The p-value indicates the probability of observing a t-value as extreme as -0.9 (or more extreme) under the null hypothesis. A p-value of 0.4 is much greater than the common significance level (α=0.05), suggesting insufficient evidence to reject the null hypothesis.

CONT.

95% CONFIDENCE INTERVAL (CI): [3357, 3417]

- This range represents the plausible values for the true mean birthweight. Since the hypothesized mean (3400) lies within this interval, it is consistent with the observed data.

Sample Mean: 3387 grams

- This is the mean birthweight observed in the sample.

What Was Tested?

- Is the provided dataset's average birthweight (bwt) different from 3400 grams? This is done using a one-sample t-test.

Results of T-test

- Sample Mean: The average birthweight in your data is 3386.79 grams (from the mean of x in the result).
- p-Value: The p-value is 0.3838 (from the p-value in the result).
- Since the p-value is greater than 0.01 (1% significance level), there isn’t enough evidence to conclude that the average birthweight is significantly different from 3400 grams.

CONT.

Confidence Interval

- Confidence Interval: The 95% confidence interval for the true average birthweight is between 3357.03 grams and 3416.54 grams (from the 95 percent confidence interval).
- Since 3400 is inside this range, it's reasonable to believe the true average birthweight could be 3400 grams.

```
        One Sample t-test

data:  data$bwt
t = -0.9, df = 1173, p-value = 0.4
alternative hypothesis: true mean is not equal to 3400
95 percent confidence interval:
 3357 3417
sample estimates:
mean of x
     3387
```

Conclusion of the hypothesis

- t-Statistic: The test statistic t=−0.9 (from the t value) is small, meaning the difference between the sample mean (3387) and 3400 is minor.
- We fail to reject the null hypothesis, meaning we do not have enough evidence to say the average birthweight is different from 3400 grams.

Interpretation:

- While the observed mean birthweight (3387 grams) is slightly lower than 3400 grams, the difference is not statistically significant.
- The variability in the data (captured by the standard deviation) and the sample size contribute to the wide confidence interval, which includes 3400 grams. Thus, we cannot rule out the possibility that the true mean is equal to 3400 grams based on this analysis.

LINEAR REGRESSION

Introduction: **Linear Regression** is a statistical method used to model the relationship between a dependent variable (response) and one or more independent variables (predictors). The goal is to find a linear equation that best predicts the dependent variable from the predictors.

Linear regression for this dataset: Linear regression is applied to this dataset to explore the relationship between maternal smoking and newborn birthweight. The dependent variable is bwt (birthweight in grams), while the independent variable is smoke, a binary indicator of whether the mother smoked during pregnancy (0 = Non-smoker, 1 = Smoker). The objective is to determine if smoking significantly affects birthweight and to quantify this impact. The model estimates the average birthweight for non-smoking mothers and predicts the reduction in birthweight for smoking mothers, providing insights into the potential adverse effects of smoking on newborn health.

Hypothesis:

- **Null Hypothesis (H_0)**: Smoking has no effect on birthweight ($\beta=0$).
- **Alternative Hypothesis (H_a)**: Smoking significantly impacts birthweight ($\beta\neq0$).

Regression Equation: $\mathbf{bwt = 3489.49 - 262.69 \cdot smoke}$

- **Intercept (3489.49)**: Predicted birthweight for non-smoking mothers.
- **Slope (-262.69)**: Predicted decrease in birthweight (grams) for smoking mothers.

Prediction Example:

Using the regression model, the predicted birthweight for a smoker is calculated to be **3226.8 grams**, demonstrating the reduction in birthweight associated with smoking. This highlights the significant negative impact of maternal smoking on newborn health.

MODEL INTERPRETATION

1. Intercept (β_0=3489.49):

- When a mother does not smoke (smoke=0), the predicted birthweight is 3489.49 grams.
- This represents the baseline birthweight for non-smoking mothers in the dataset.

2. Slope (β_1=−262.69):

- For mothers who smoke (smoke=1), the predicted birthweight decreases by 262.69 grams compared to non-smoking mothers.
- This negative coefficient indicates that smoking has a negative effect on birthweight.

3. Statistical significance:

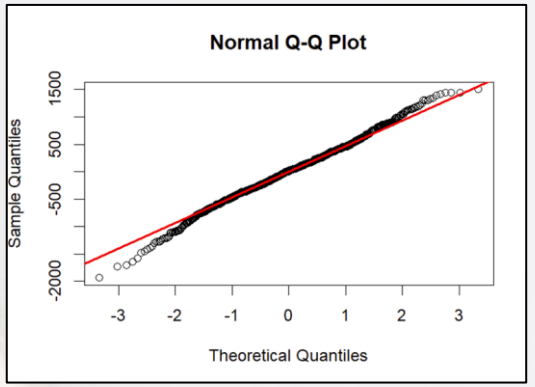
- The p-value for the slope ($p<0.001$) is highly significant, indicating that smoking has a statistically significant effect on birthweight.
- We can reject the null hypothesis (H_0:$\beta=0$) and conclude that there is a relationship between smoking and birthweight.

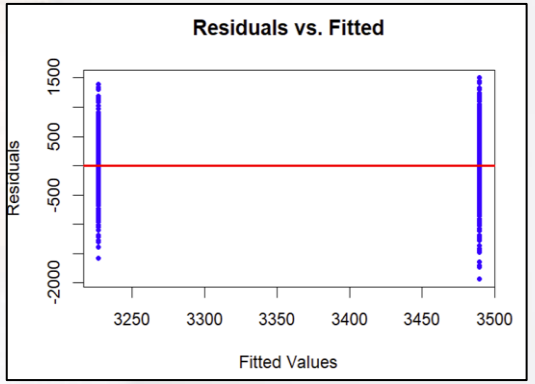
4. Goodness of fit (r-squared):

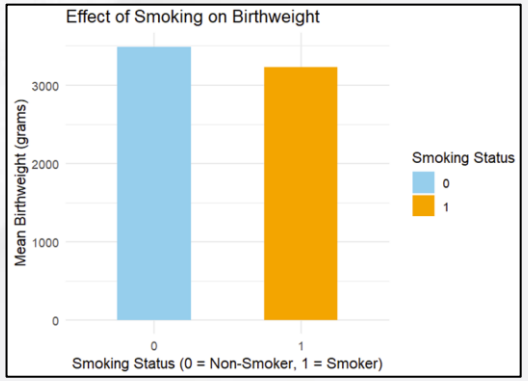
- The R^2 value is 0.061, meaning that only 6.1% of the variation in birthweight is explained by whether the mother smokes.
- While the effect of smoking is statistically significant, other factors not included in this model (e.g., Gestation period, maternal weight) likely contribute to the variation in birthweight.

METRIC	BACKWARD MODEL
Intercept (β_0)	3489.49 grams
Slope (β_1)	-262.69 grams
p-value (for β_1)	< 0.001
R^2	0.061 (6.1%)
Adjusted R^2	0.060
Residual Std. Error	503.8 grams

MODEL INTERPRETATION(CONT.)







5. Model assumptions:

Residuals vs. Fitted plot:

- The residuals are symmetrically distributed around zero, indicating that the model captures the linear relationship between smoke and bwt well.
- The consistent spread of residuals across fitted values supports the assumption of homoscedasticity.
- The bar chart compares the mean birthweight for babies born to non-smoking mothers (smoking status = 0) and smoking mothers (smoking status = 1).

Bar chart:

- Non-smoking mothers have babies with a higher mean birthweight (approximately 3489.5 grams).
- Smoking mothers have babies with a reduced mean birthweight (approximately 3226.8 grams), reflecting a decrease of about 262.69 grams as estimated by the regression model.
- This chart reinforces the regression results, highlighting the negative impact of smoking on birthweight.

Q-Q plot:

- The residuals align closely with the diagonal line, suggesting that they follow a normal distribution.
- This indicates that the model's assumptions for hypothesis testing and confidence intervals are valid.

6. Practical implication:

- Smoking during pregnancy is associated with a significant reduction in birthweight.
- This information can be used in public health campaigns to discourage smoking among expectant mothers.

MULTIPLE LINEAR REGRESSION

- Introduction **:** Multiple regression is a statistical technique used to model the relationship between a dependent variable (birthweight) and multiple independent variables (e.g., gestation, region, age, height, weight, smoke). It predicts outcomes and identifies the strength and direction of these relationships.
- Multiple linear regression for this dataset: Multiple regression is crucial for this dataset as it helps identify how factors like gestation and smoking influence birthweight while quantifying the precise impact of each variable, such as smoking reducing birthweight by approximately 235 grams. It also accounts for overlapping effects among predictors, like region and age, ensuring accurate insights. Additionally, the model enables birthweight predictions for new cases, supporting better prenatal care planning and interventions. The effectiveness of multiple regression lies in its ability to identify significant predictors through p-values, assess the model's accuracy using metrics like R-squared, and provide actionable insights by highlighting key factors, such as the importance of reducing smoking during pregnancy, to guide healthcare decisions.
- Hypothesis Statement:
  - **Null Hypothesis (H_0):** The coefficients of all predictors in the model are equal to zero. This means the independent variables (gestation, region, age, height, weight, smoke) do not collectively have a significant effect on birthweight.

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$$

  - **Alternative Hypothesis (H_1):** At least one predictor's coefficient is not zero, meaning that at least one independent variable significantly impacts birthweight.

$$H_1 : \text{At least one } \beta_i \neq 0$$

- Regression Equation

bwt (birthweight) i = β 0 + β 1 (gestation i) + β 2 (regionnorthwest i) + β 3 (regionsoutheast i)+ β 4 (regionsouthwest i) + β 5 (age i) + β 6 (height i) + β 7 (weight i) + β 8 (smoke i) + εi

Where: i = 1,2,...,n (the index for each observation in the dataset).

β 0 : Intercept term.

β 1 ,β 2 ,...,β 8 : Coefficients for each predictor variable.

εi : Residual error term for observation ii.

Bwt = −2182.074 + 12.378 (gestation) − 31.361 (regionnorthwest) − 14.172 (regionsoutheast) − 30.795 (regionsouthwest) + 1.960 (age) + 12.059 (height) + 3.595 (weight) − 235.434 (smoke)

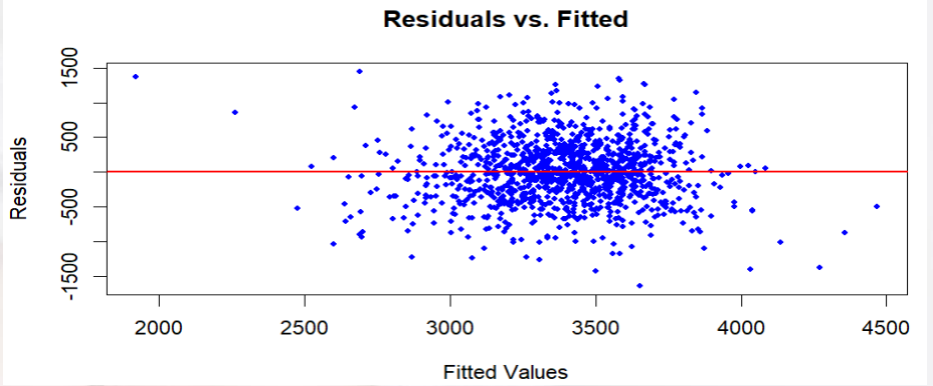
INTERPRETATION OF MULTIPLE LINEAR REGRESSION

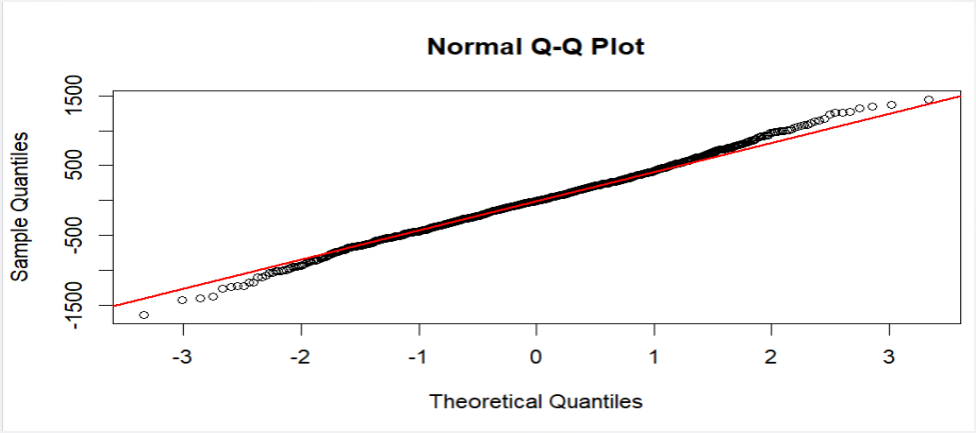
- Results Explanation:
  1. **Intercept (-2182.074) :** The intercept is the baseline predicted birthweight when all predictors are set to zero. While zero values for predictors like gestation and weight may not make practical sense, this term serves as a mathematical anchor for the equation.
  2. **Coefficient for gestation (+12.378):** For every additional day of gestation, birthweight increases by approximately 12.378 grams, holding all other factors constant.
  3. **Coefficients for region (northwest: -31.361) (southeast: -14.172) (southwest: -30.795) :** The region variable is a categorical factor with multiple levels, the reference level (base category) is assumed to be northeast. Birthweight in other regions (northwest, southeast, southwest) is predicted to be lower compared to the reference region (northeast).
  4. **Coefficient for age (+1.960):** For every additional year of maternal age, birthweight increases by approximately 1.960 grams, holding other factors constant.
  5. **Coefficient for height (+12.059):** For every additional centimeter of maternal height, birthweight increases by 12.059 grams, holding other factors constant.
  6. **Coefficient for weight (+3.595):** For every additional kilogram of maternal weight, birthweight increases by 3.595 grams, holding other factors constant.
  7. **Coefficient for smoke (-235.434):** If the mother smokes, birthweight decreases by approximately 235.434 grams, holding other factors constant. This is a significant negative effect.
- Significant Variables:
  - **gestation**: Positive impact ($p < 0.001$).
  - **height**: Positive impact ($p < 0.001$).
  - **weight**: Positive but less significant ($p = 0.0235p$).
  - **smoke**: Significant negative impact ($p < 0.001$).
- Non-Significant Variables:
  - **region**, **age**: P-values > 0.05, not statistically significant in this model. Removing these variables reduces complexity without affecting model accuracy
- Goodness-of-Fit:
  - **R-squared**: 0.2529 indicates that 25.29% of the variance in birthweight is explained by the predictors included in the model. While this is a moderate fit, other unmeasured factors likely influence birthweight. (e.g., other factors like nutrition, genetics, or healthcare access are not included in the model).
  - **Adjusted R-squared**: 0.2478, adjusted R-squared accounts for the number of predictors in the model, preventing overfitting. A slight reduction from R-squared suggests the model is not overfit and predictors contribute meaningfully.
  - **F-statistic** is significant ($p<0.001$), Indicates that the model as a whole is statistically significant, meaning the predictors collectively explain the variance in birthweight effectively.
- Backward Stepwise Regression: It simplifies the model by removing predictors that do not significantly improve the model's performance (measured by AIC - Akaike Information Criterion). This model with variables: gestation, height, weight, smoke. R-squared and Adjusted R-squared values are similar to the full model, indicating these four predictors carry most of the explanatory power. Backward regression is preferable for interpretability and ease of application.

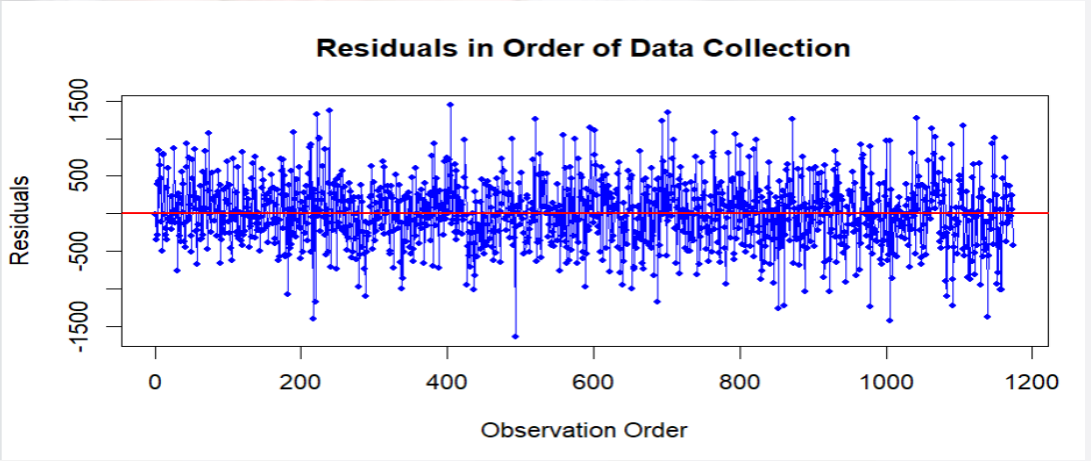
CONT.

- Forward Stepwise Regression: It starts with no predictors and iteratively adds significant variables to optimize model performance. The forward model includes additional predictors (e.g., region and age), but they do not contribute significantly. The model performance is comparable to the backward regression model.
- Predicted Birthweight: Using the original model (multi_model), we predict the birthweight of a baby for a new mother: For a new observation (gestation=280, region=southwest, age=30, height=165, weight=70, smoke=1), the predicted birthweight is: **3317.615 grams**.
- Residuals Vs. Fitted Plot : This plot helps check the assumption of homoscedasticity, that is the constant variance of residuals. In our plot, the residuals are fairly scattered without a clear pattern, suggesting homoscedasticity.
- Normal Q-Q Plot: Assesses if the residuals are normally distributed. In our plot, most points align well, suggesting residuals are approximately normal. A few deviations at the tails indicate minor issues with normality, which might not severely impact the regression's validity.
- Residuals in Order of Data Collection: Checks the independence of residuals and reveals any patterns over time. The plot shows no clear pattern or trend, suggesting residuals are independent, and there's no temporal correlation.
- These diagnostics collectively validate the key regression assumptions, indicating the model is reliable and interpretable.
- Conclusion: The regression model is significant (F-statistic = 49.29, $p < 0.001$), with gestation, height, weight, and smoking as key predictors. Region and age are not significant. We reject the null hypothesis (H_0), confirming the model's effectiveness for insights into birthweight and prenatal care.

METRIC	BACKWARD MODEL	FORWARD MODEL
Significant Predictors	Gestation, Height, Weight, Smoke	Gestation, Height, Weight, Smoke
Non-Significant Predictors	Removed (Region, Age)	Retained (Region, Age)
R-squared	0.2518	0.2529
Adjusted R-squared	0.2492	0.2478
Residual Std. Error	450.2	450.7







CONCLUSION

Linear Regression Analysis : Smoking has a statistically significant, negative relationship with birth weight, but it explains only a small proportion of the variability in birth weight.

- **Key Observations:**
  - p-value for smoking coefficient: Less than 0.001.
  - Significance level (α): 0.05.

- **Final Remark:**
  - **We reject the null hypothesis (H_0)** because the p-value is significantly less than 0.05. This indicates a statistically significant relationship between maternal smoking and birth weight. The negative coefficient (-262.69 grams) shows that, on average, birth weight decreases by approximately 262.69 grams for mothers who smoke compared to non-smokers.

T-Test Analysis : There is a statistically significant difference in birth weight between babies born to mothers who smoke and those who do not.

- **Key Observations:**
  - t-statistic: 8.42
  - p-value: Less than 0.001.
  - Significance level (α): 0.05.

- **Final Remark:**
  - Since the p-value is significantly less than 0.05, **we reject the null hypothesis (H_0)**, concluding that smoking status has a significant effect on birth weight. The results suggest that babies born to mothers who smoke tend to have lower birth weights on average compared to those born to non-smoking mothers.

CONT.

Multiple Regression Analysis : Smoking, along with other factors such as maternal weight, gestation period, and region, has a statistically significant relationship with birth weight. However, the model explains only a moderate proportion of the variability in birth weight.

- **Key Observations:**
  - **Adjusted R^2:** 0.35 (indicating moderate explanatory power).
  - **p-values for key coefficients (smoking, weight, gestation):** All less than 0.05.
- **Final Remark:**
  - **We reject the null hypothesis (H_0)** because the F-statistic is highly significant ($p<0.001$), indicating that the independent variables collectively influence birthweight. Furthermore, individual significance tests highlight that gestation, height, weight, and smoking have significant effects, while region and age are not statistically significant. Removing non-significant variables simplifies the model without losing predictive power. This confirms the model's utility in identifying critical factors influencing birthweight and supports the use of multiple regression for actionable insights in prenatal care.

In Conclusion:

Overall, The results of the t-test and linear regression support the analysis's conclusion that **smoking has a statistically significant negative effect on birth weight**. Compared to babies born to non-smokers, babies born to moms who smoke typically weigh less at birth. Furthermore, multiple regression shows that other factors including **gestational age and mother weight have a considerable impact on birth weight in addition to smoking**. The multifaceted nature of the factors influencing birth weight is shown by the fact that the overall model only accounts for a minor amount of variability. In order to enhance delivery outcomes, our data highlight the significance of targeting smoking and other important factors during prenatal care.

REFERENCES

i. Chen, J. (2024, September 21). *How a Histogram Works to Display Data*. Investopedia. https://www.investopedia.com/terms/h/histogram.asp

ii. Hayes, A. (2024, July 16). *Multiple Linear Regression (MLR) definition, formula, and example*. Investopedia. https://www.investopedia.com/terms/m/mlr.asp

iii. Kavita. (2024, November 27). *Linear Regression: A Comprehensive Guide*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/

iv. Wikipedia contributors. (2024, February 12). *One- and two-tailed tests*. Wikipedia. https://en.wikipedia.org/wiki/One-_and_two-tailed_tests