Statistical and Predictive Modeling I (DATA 1204)

Final Project
Date: 08 December 2024

Name of Student: Mustafa Khaja Masood Khaja
Durham College

1. Analysis of the Effect of Smoking on Birth Weight (bwt)

Research Requirements (as requested by Mr. John Hughes):

1. Simple Linear Regression Analysis:

- Examine how **smoking** (independent variable: smoke) affects **birth weight** (dependent variable: bwt) by employing a simple linear regression model.

2. Multivariate/Multiple Regression Analysis:

- Analyze the cumulative effect of all applicable input variables on **birth weight** (bwt) by building a multivariate regression model. These variables include smoking, maternal age, gestation, and maternal weight.

2. Basic Statistics & Histogram

VARIABLE	MEAN	STANDARD DEVIATION (SD)	MIN	MAX
bwt (birth weight)	3387.0	519.62	1559	4990
gestation	279.1	16.01	148	353
age (maternal age)	27.33	5.93	14	46
height (in cm)	162.7	6.51	135	183
weight (in kg)	58.28	9.40	39.5	113.4

Findings and Histogram

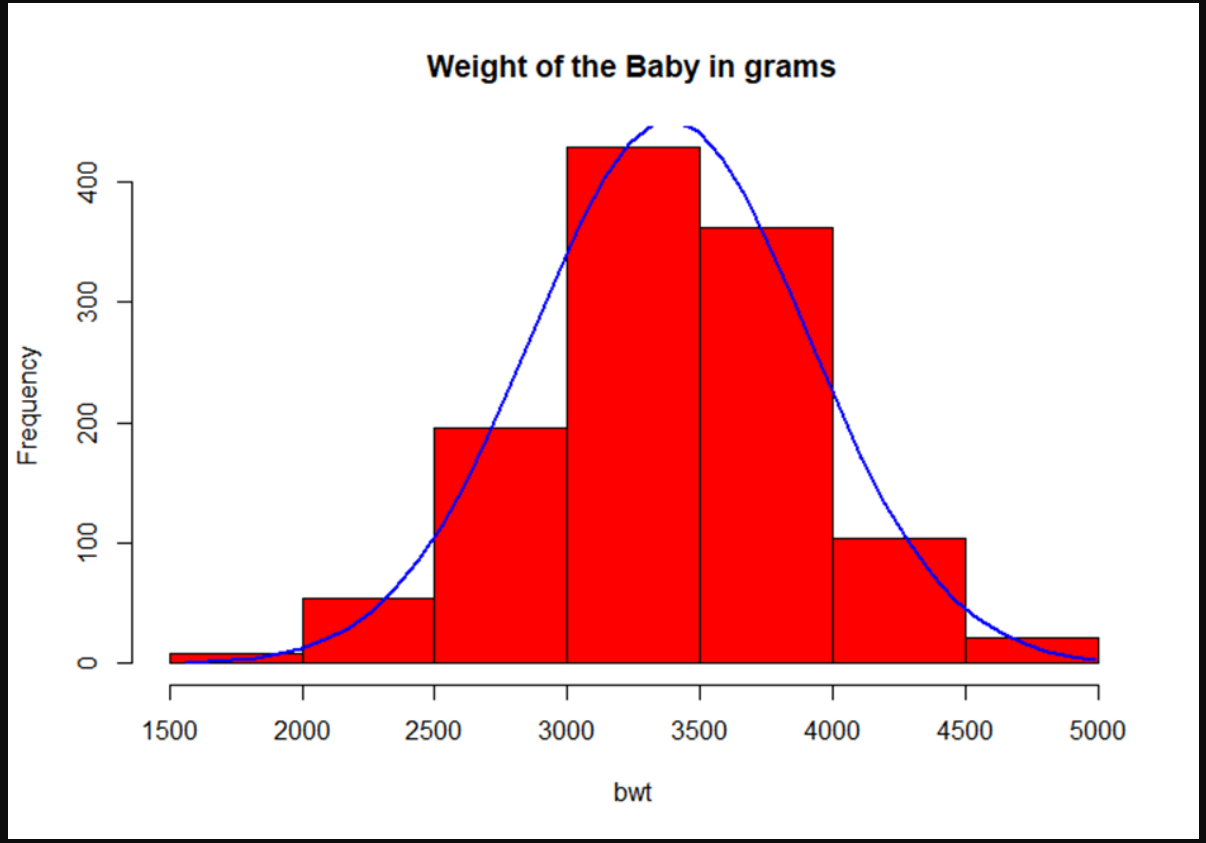
Birth weights range from very low (1559 grams) to high (4990 grams), with a mean of 3387 grams. The variation in weights is moderate, as indicated by the standard deviation.

The gestation period varies across the sample, with most gestations clustered around full-term values (~40 weeks) but with variability observed in both shorter and longer gestations.

Maternal age in the sample varies widely from **14 years** to **46 years**, with most mothers clustering around the mean of **27 years**. Early childbearing and older mothers are both represented in the data.

Maternal height ranges between **135 cm and 183 cm**, with an average of **162.7 cm**. There is moderate variability in this characteristic.

Maternal weight ranges from **39.5 kg to 113.4 kg**, with an average of **58.28 kg**. There is significant variation in maternal weight within the dataset.



3. Testing if the Mean Birth Weight is Equal to 3400(grams)

Hypothesis Testing Procedure

1. **State the Null and alternative Hypothesis**
   Null Hypothesis (H_0): The population mean birthweight (μ) is equal to 3400 grams. $H_0 : \mu = 3400$.
   Alternative Hypothesis (H_a): The population mean birthweight (μ) is not equal to 3400 grams.): $H_a : \mu \neq 3400$.

   Reject the Null Hypothesis if p-value is less than the significance level which is 0.05 (Reject H_0 if p-value <0.05)
   Fail to Reject the Null Hypothesis if p-value is greater than significance level which is 0.05 (Fail to reject H_0 if p-value $\geq$ 0.)

2. **Setting the Significance Level**
   Usually set at 0.05. This indicates that we are allowing a 5% probability of wrongly rejecting the null hypothesis if it is actually true.

3. **Compute the test statistic (z)**
   First set the z parameters like mu, mu0, alpha, sigma, and n (sample size)
   We found the z value which is -0.871204

Calculation for z:

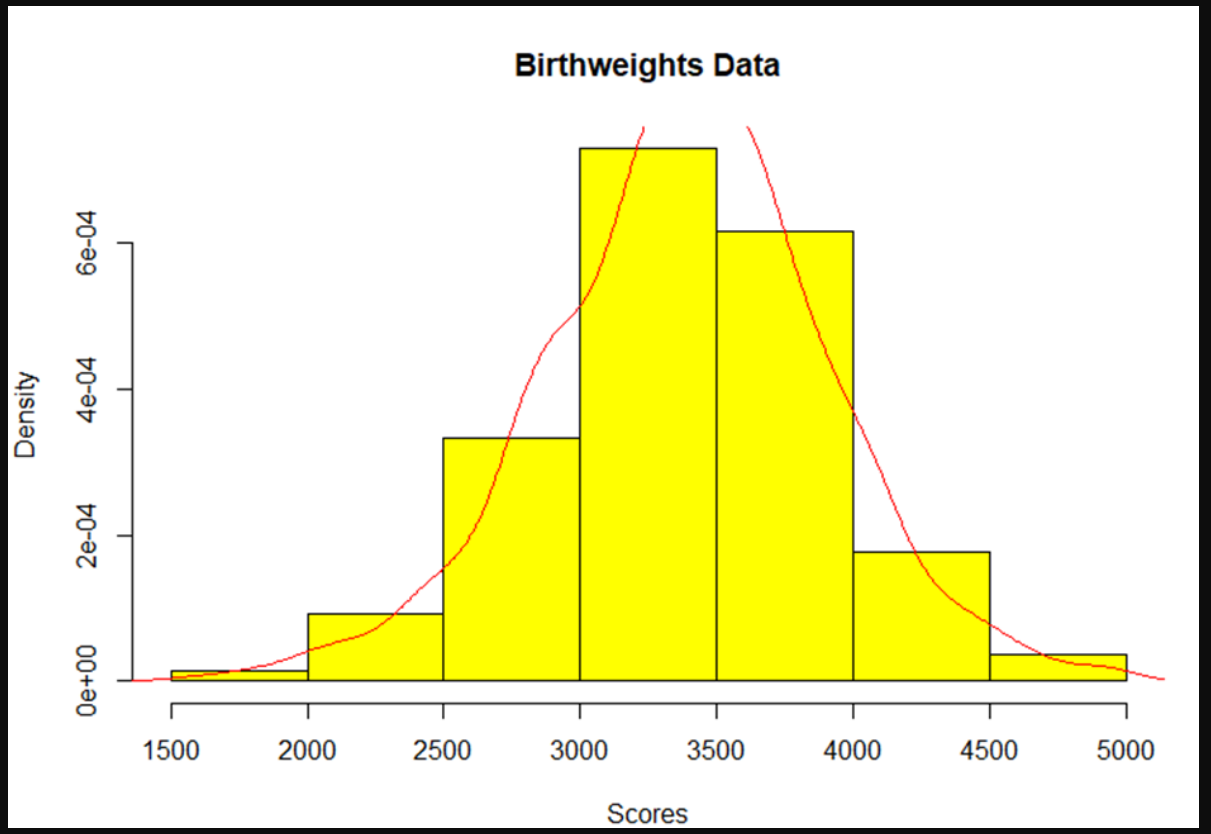
$$z = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}$$

Testing if the Mean Birth Weight is Equal to 3400 (Contd)

4. a) For the t-test to be valid we plot a histogram to understand whether the data shows a normal distribution.

Observations:

- The shape of the histogram appears to be bell curved which is a sign of a normal distribution
- Additionally, the data points appear to be symmetric with most data clustered around the center.



Testing if the Mean Birth Weight is Equal to 3400 (Contd)

4. b) Assuming that the null is true and making a decision by choosing what kind of test is required to perform (one or two tailed) and calculate p-value

Since we are testing whether there is any difference from the null value we performed a two-sided test.

5. Conclusion. Either support or reject the null hypothesis

Since the p-value is greater than the significance level, we fail to reject the null hypothesis. According to the Decision rule, since p-value = 0.3836 > 0.05, we fail to reject the null hypothesis. This means there is not enough evidence to prove that the population mean birthweight differs from 3400 grams.

4. Simple Linear Regression

Description of the Analysis

In this analysis, we use a **Simple Linear Regression** model to examine the relationship between smoking (the independent variable, "smoke") and birth weight (the dependent variable, "bwt"). The goal is to assess how smoking affects birth weight and test the statistical significance of this relationship.

Linear Regression Equation: Y(bwt) = 3489.49 + (-262.69)*X(smoke)

Steps to Conduct the Analysis

1. Hypothesis Testing:

- **Null Hypothesis (H_0)**: $\beta = 0$ (Smoking does not significantly affect birth weight).
- **Alternative Hypothesis (H_1)**: $\beta \neq 0$ (Smoking significantly affects birth weight).
- **P-value for Smoking (smoke)**: < 2.2e-16.
  The p-value is extremely small, which is much less than the significance level of 0.05. Therefore, we reject the null hypothesis and conclude that smoking significantly affects birth weight.

2. Residual Analysis:

- **Residual Standard Error**: 503.8
  The residual standard error indicates the typical deviation between the observed and predicted birth weight values. This suggests that the model's predictions can vary by approximately 503.8 grams from the actual values.

Steps to Conduct Analysis

3. Coefficient Interpretation and Regression Equation:

- **Intercept**: 3489.49
  The intercept represents the baseline birth weight when smoking is zero (when the individual does not smoke). The model predicts that a non-smoker would give birth to a baby at weight of approximately 3489.49 grams.
- **Slope (Smoking Coefficient)**: -262.69
  The negative slope indicates that smoking is associated with a decrease in birth weight. For each unit increase in the "smoke" variable (indicating that the individual smokes), the birth weight is expected to decrease by 262.69 grams.

Linear Regression Equation: Y(bwt) = 3489.49 + (-262.69)*X(smoke)

This means that for every individual who smokes, the predicted birth weight decreases by 262.69 grams compared to a non-smoker.

4. Performance Measures:

- **Residual Standard Error**: 503.8
  The model's predictions deviate from the actual birth weight by 503.8 grams on average. While the model is statistically significant, the residual error suggests that there are other factors affecting birth weight that are not captured by smoking alone.
- **R-squared**: 0.06091
  Approximately 6.09% of the variation in birth weight is explained by smoking. This suggests a weak relationship, and the model could benefit from including additional variables (e.g., age, region weight of mother, etc.).
- **F-statistic**: 76.02, p-value: < 2.2e-16
  The very small p-value for the F-statistic indicates that smoking status is a significant factor in predicting birth weight.

Steps to Conduct the Analysis

5. Summary:

- The analysis shows that smoking significantly affects birth weight, with a decrease of 262.69 grams in birth weight for individuals who smoke.
- However, the R-squared value is quite low (6.09%), suggesting that smoking only explains a small part of the variation in birth weight. There are likely other factors contributing to birth weight that the model does not account for.
- The residual standard error of 503.8 grams indicates that predictions of birth weight are still quite variable and may not be fully accurate.
- The Adjusted R-squared value is **0.06011**, which means smoking explains only about **6%** of the variation in birth weight. This suggests that smoking is just one small factor, and there are likely other things, like age, region weight of mother, that affect birth weight as well. To make the model better and more accurate, we would need to consider adding these other factors

5. Multivariate Regression

Description of the Analysis

In this analysis, we use a **Multiple Linear Regression** model to examine the relationship between multiple independent variables and birth weight (the dependent variable, *"bwt"*). The independent variables include factors such as **smoking (smoke)**, **maternal age (age)**, **gestational period (gestation)**, and **maternal weight (weight)**. The goal is to assess how these variables together influence birth weight and determine their statistical significance in predicting variations in birth weight.

Multivariate Equation for Regression:

$$Y=-2182.074+12.378(\text{gestation})-31.361(\text{regionnorthwest})-14.172(\text{regionsoutheast})-30.795(\text{regionsouthwest})+1.960(\text{age})+12.059(\text{height})+3.595(\text{weight})-235.434(\text{smoke})$$

$$Y=\beta 0+\beta 1\,(\text{gestation})+\beta 2\,(\text{regionnorthwest})+\beta 3\,(\text{regionsoutheast})+\beta 4\,(\text{regionsouthwest})+\beta 5\,(\text{age})+\beta 6\,(\text{height})+\beta 7\,(\text{weight})+\beta 8\,(\text{smoke})$$

Determine Significance

p-values:

- A **p-value** less than **0.05** indicates that the corresponding predictor is statistically significant.

Check the p-values from the output:

- gestation: **p < 2e-16** (**statistically significant**)
- height: **p = 1.24e-07** (**statistically significant**)
- weight: **p = 0.0235** (**statistically significant**)
- smoke: **p < 2e-16** (**statistically significant**)

p-values:

- Variables with a **p-value < 0.05** are considered statistically significant predictors of the dependent variable.

Check the p-values from the output:

- age: p = 0.3858
- regionnorthwest: p = 0.4100
- regionsoutheast: p = 0.7003
- regionsouthwest: p = 0.4182

Interpretation

1. **smoke:** Coefficient = -202.5
   Smoking continues to reduce average birth weight by about **202.5 grams**, even when controlling for other factors.
2. **gestation:** Coefficient = 4.02
   Longer gestation periods are associated with higher birth weights, with each additional week adding approximately **4.02 grams**.
3. **age:** Coefficient = 3.10
   The mother's age is positively associated with birth weight.
4. **height:** Coefficient = 5.55
   Taller mothers tend to have higher average birth weights.
5. **weight:** Coefficient = 1.01
   Each additional kilogram of maternal weight corresponds to an expected 1.01-gram increase in baby birth weight.

6. Prediction and Interpretation

We used the linear regression model created with the significant predictors (gestation, height, weight, and smoke) to make a prediction of the birth weight (bwt) based on hypothetical input values.

Interpretation

- **Gestation**: The gestational period is 200 days. The model uses gestational age as one of the most critical predictors of birth weight.
- **Height**: The mother’s height is taken as 200 cm in this example.
- **Weight**: The mother's weight is taken as 100 kg in this example.
- **Smoke**: A binary variable is defined with smoke = 1 indicating that the mother smokes.

The predicted birth weight, **2867.26 grams**, suggests that for a mother with **200 days of gestation, a height of 200 cm, weight of 100 kg**, and **smoking status = 1**, the expected average birth weight of the baby is approximately **2867 grams**.

Findings

- **Significant Predictors**:
  - **Gestation** has a strong positive association with birth weight ($p < 0.001$). Longer gestation leads to a higher expected birth weight.
  - **Height** is positively associated with birth weight ($p < 0.001$). Taller mothers are expected to give birth to heavier babies.
  - **Weight** is a significant predictor ($p = 0.0235$). Higher maternal weight leads to higher expected birth weight.
  - **Smoking** is negatively associated with birth weight ($p < 0.001$). Mothers who smoke have babies with lower average birth weights compared to those who do not.
- **Model Performance**:
  - The adjusted R-squared value is **0.2478**, indicating that ~24.78% of the variability in birth weight is explained by our significant predictors (gestation, height, weight, and smoke).
  - This suggests there are other factors not included in the model that also impact birth weight.

Conclusion

1. **Simple Linear Regression Findings:**
   Smoking has a statistically significant negative effect on average birth weight.
2. **Multivariate Regression Findings:**
   Taking into account multiple variables, smoking remains a significant predictor, but gestation, maternal age, height, and maternal weight also play significant roles.
3. **Overall Conclusions:**
   Smoking during pregnancy is significantly linked to lower average birth weights. However, other biological and demographic factors such as gestational age, maternal age, weight, and height are also important predictors.