Correlation and Least Squares Regression Worksheet

Week 9 Lecture 2 Lab 07

YOUR NAME HERE

March 4th, 2024

Correlation

Direction: This is your worksheet. PLEASE DO NOT SUBMIT THIS AS A LAB ASSIGNMENT!

Please run the code chunk below and load your data set as well as utilizing `library()` functions that you need.

```r
library(openintro)
library(tidyverse)
data("babies")
glimpse(babies)
```

```
Rows: 1,236
Columns: 8
$ case      <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1~
$ bwt       <int> 120, 113, 128, 123, 108, 136, 138, 132, 120, 143, 140, 144, ~
$ gestation <int> 284, 282, 279, NA, 282, 286, 244, 245, 289, 299, 351, 282, 2~
$ parity    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ age       <int> 27, 33, 28, 36, 23, 25, 33, 23, 25, 30, 27, 32, 23, 36, 30, ~
$ height    <int> 62, 64, 64, 69, 67, 62, 62, 65, 62, 66, 68, 64, 63, 61, 63, ~
$ weight    <int> 100, 135, 115, 190, 125, 93, 178, 140, 125, 136, 120, 124, 1~
$ smoke     <int> 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, ~
```

Question 1: Please explain what you see in this dataset.

Calculate Correlation Coefficient

Question 2: Please run the code chunk below and comment on the correlation coefficient.

```
cor(babies$gestation, babies$bwt, use = "na.or.complete")
```

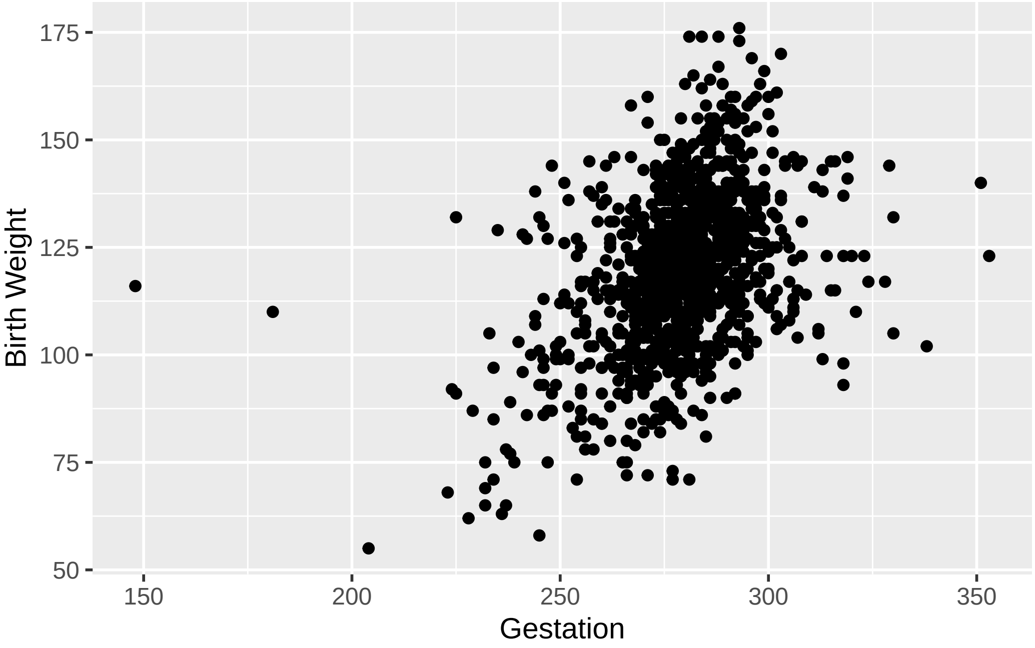
```
[1] 0.407854
```

Outliers

Explain briefly here what an outlier is.

Question 3: Let's examine scatterplot to look for potential outliers. Do we have potential outliers in this scatterplot?

```
ggplot(na.omit(babies), aes(x = gestation, y = bwt)) +
  geom_point() +
  labs(x = "Gestation", y = "Birth Weight")
```



Scatterplot - 1

We see two outliers on the left-hand side of the scatterplot. Let's find them and remove these outliers. (Run the code chunk below.)

```
babies <- babies %>%
  filter(case != 261 & case != 870 & case !=979) # this removes three outliers whose case
```

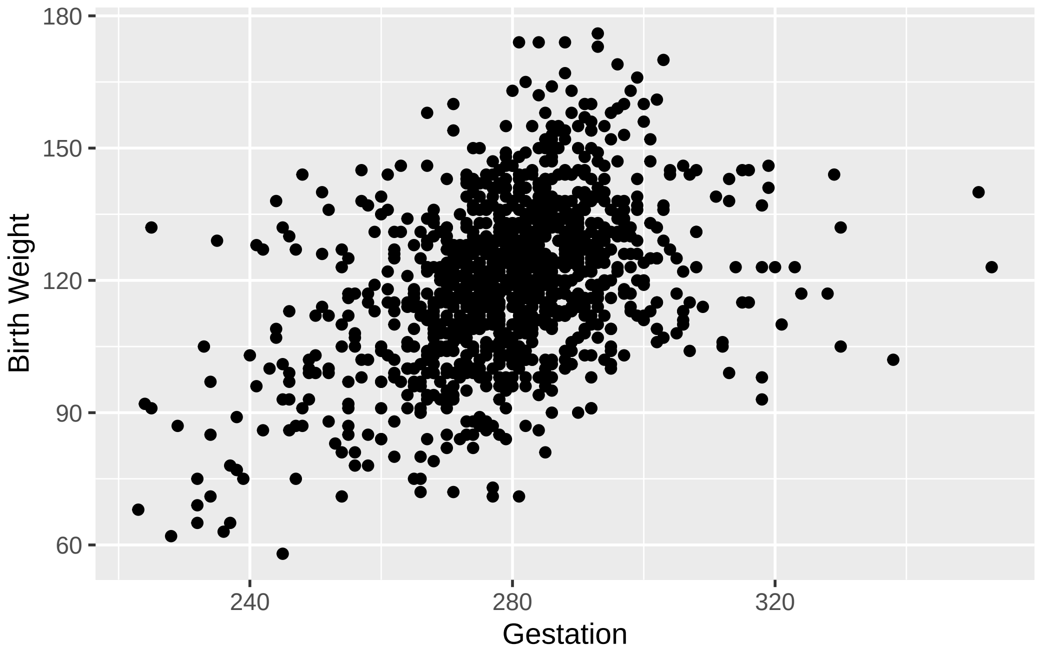
Type your lecture notes here.

Question 4: Let's check the correlation coefficient and scatterplot again. Interpret the outputs.

```
cor(babies$gestation, babies$bwt, use = "na.or.complete")
```

```
[1] 0.4144271
```

```
ggplot(na.omit(babies), aes(x = gestation, y = bwt)) +
  geom_point() +
  labs(x = "Gestation", y = "Birth Weight")
```



Question 5: Do we have potential outliers in this scatterplot given above?

Scatterplot - 2

We see some outliers on the right-hand side of the scatterplot. Let’s remove those outliers.

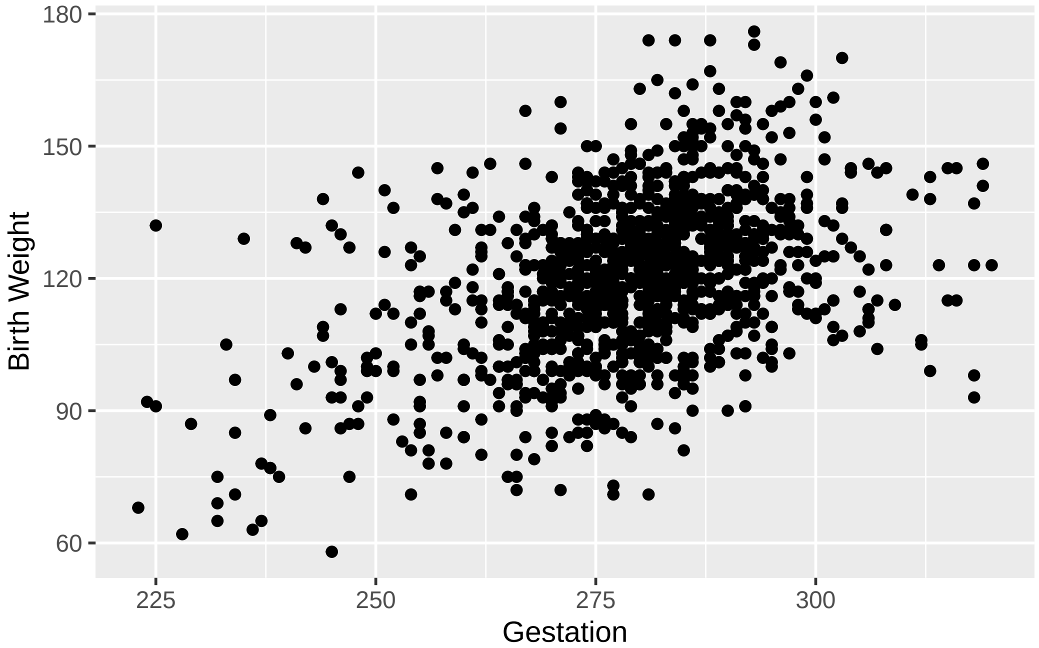
```
babies <- babies %>%
  filter(case != 1173 & case != 11 & case != 1200 & case != 153 & case != 762 & case != 97
```

Question 6: Let’s check the correlation coefficient and scatterplot again. Interpret the outputs.

```
cor(babies$gestation, babies$bwt, use = "na.or.complete")
```

```
[1] 0.4387897
```

```
ggplot(na.omit(babies), aes(x = gestation, y = bwt)) +
  geom_point() +
  labs(x = "Gestation", y = "Birth Weight")
```



Question 7. Compare 3 correlation coefficients and 3 scatterplots. Interpret the potential outliers in terms of being leverage points / influential points.

Type your answer here.

Question 8. What is your final conclusion? Do you delete those outliers or not? Explain your reasoning.

Bivariate Regression

Case of this Lab

Understanding Birth Weight and Gestation: A Least Squares Regression Analysis

In the area of prenatal care and childbirth, understanding the relationship between gestation period and birth weight is crucial. We often speculate about how the duration of pregnancy might be related with the weight of a newborn.

In the `babies` dataset, each observation includes information about gestation period and birth weight. We want to investigate if there's a linear relationship between these two variables.

Steps for Conducting Hypothesis Testing for This Test

Step 1. Formulate Hypotheses

Question 9. Write out the null & the alternative hypothesis in words, in the context of this study:

Type here.

Step 2. Generate Your Model

Question 10. Run the code chunk below and explain what each argument does in this function.

```r
# data(babies) # if you decided to keep the outliers, reload the dataset again.
fit <- lm(babies$bwt ~ babies$gestation)
summary(fit)
```

```
Call:
lm(formula = babies$bwt ~ babies$gestation)

Residuals:
    Min      1Q  Median      3Q     Max 
-49.751 -11.047   0.046   9.902  53.249 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -36.43018    9.21722  -3.952 8.19e-05 ***
babies$gestation   0.55936    0.03299  16.958  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.38 on 1206 degrees of freedom
  (13 observations deleted due to missingness)
Multiple R-squared:  0.1925,	Adjusted R-squared:  0.1919 
F-statistic: 287.6 on 1 and 1206 DF,  p-value: < 2.2e-16
```

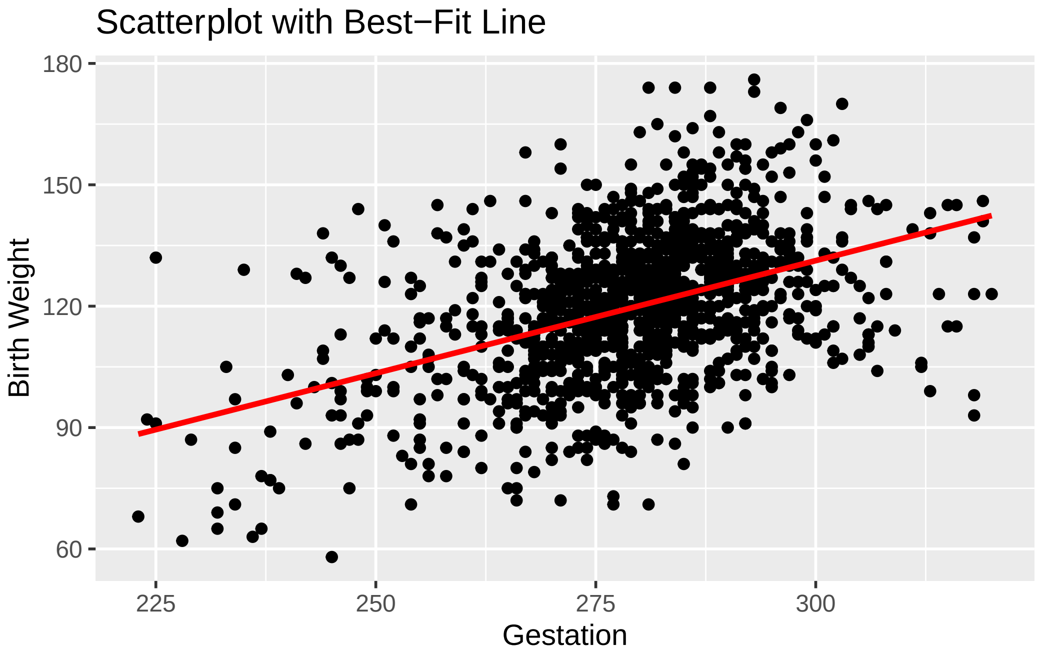
Question 11. Check the conditions/assumptions for this study and interpret these assumptions overall.

Step 3.1 Checking Conditions/Assumptions

Linearity with Scatterplot

```
ggplot(na.omit(babies), aes(x = gestation, y = bwt)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Scatterplot with Best-Fit Line",
       x = "Gestation",
       y = "Birth Weight")
```

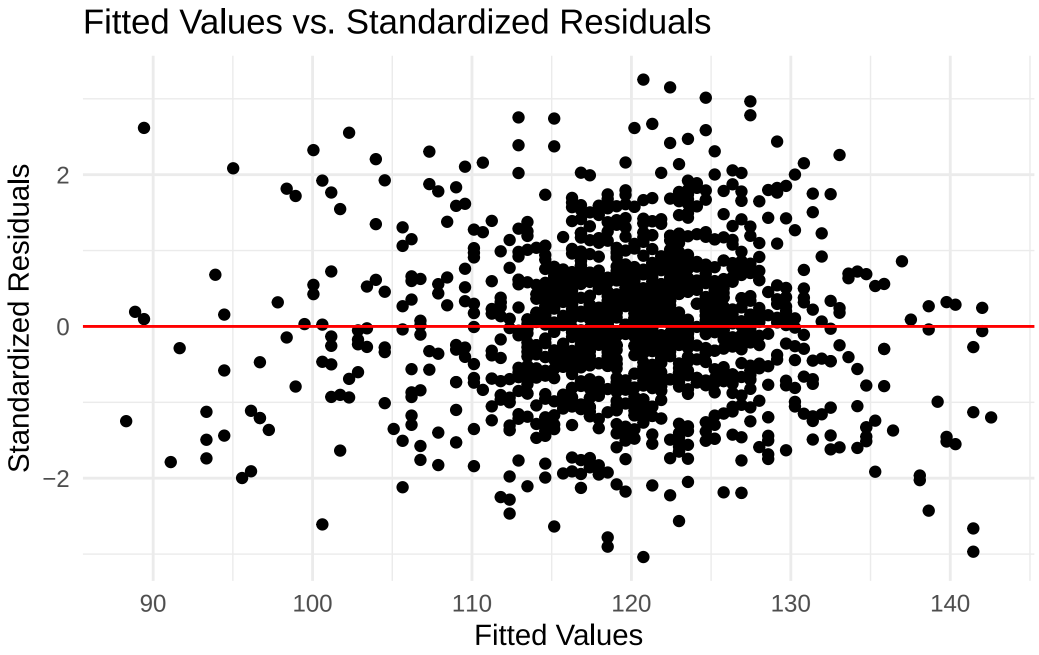
```
`geom_smooth()` using formula = 'y ~ x'
```



Linearity with Residuals Plot

```
# get list of residuals
res <- resid(fit)
res.stdres <- rstandard(fit) # standardized residuals

# produce residual vs. fitted plot
ggplot(data.frame(fitted = fitted(fit), stdres = res.stdres),
       aes(x = fitted, y = stdres)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "solid", color = "red") +
  labs(title = "Fitted Values vs. Standardized Residuals",
       x = "Fitted Values",
       y = "Standardized Residuals") +
  theme_minimal()
```

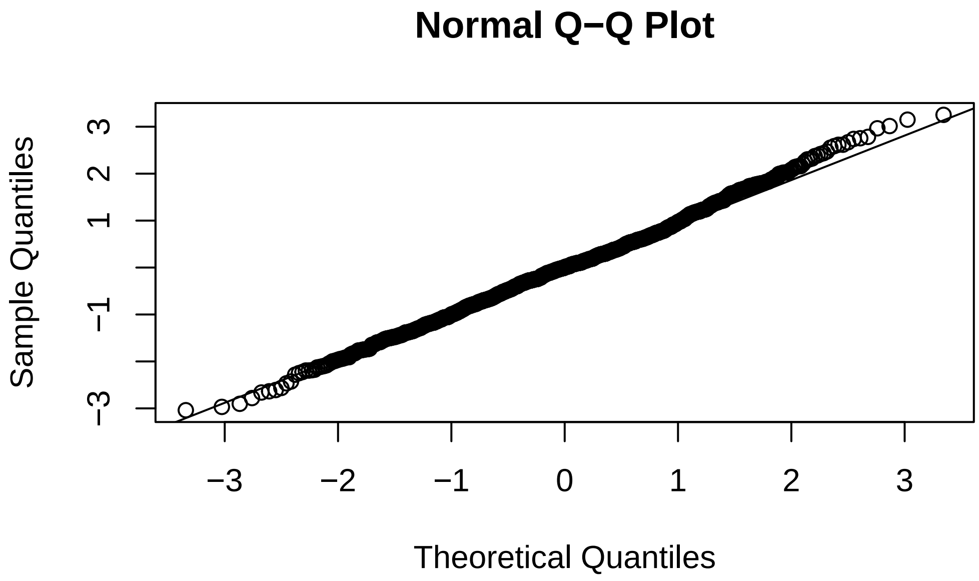


Type your interpretation here.

Step 3.2. Checking Conditions/Assumptions - Normality

```
# create Q-Q plot for residuals
qqnorm(res.stdres) # by using standardized residuals

# add a straight diagonal line to the plot
qqline(res.stdres)
```

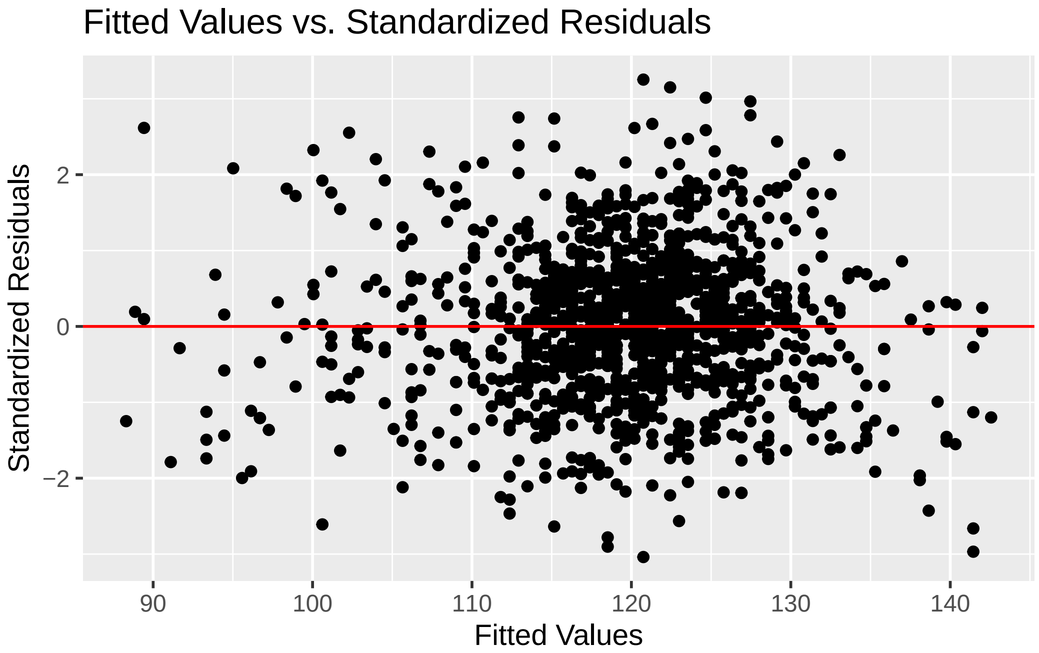


Type your interpretation here.

Step 3.3. Checking Conditions/Assumptions - Homoscedasticity

```
# get list of residuals
res <- resid(fit)
res.stdres <- rstandard(fit) # standardized residuals

# produce residual vs. fitted plot
ggplot(data.frame(fitted = fitted(fit), stdres = res.stdres),
       aes(x = fitted, y = stdres)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "solid", color = "red") +  # Add a horizontal line
  labs(title = "Fitted Values vs. Standardized Residuals",
       x = "Fitted Values",
       y = "Standardized Residuals")
```



Type your interpretation here.

Draw conclusion

Question 12. Re-run the code chunk below. Interpret Estimate, p-value of the slope, and Adjusted R-Squared.

```
fit <- lm(babies$bwt ~ babies$gestation)
summary(fit)
```

```
Call:
lm(formula = babies$bwt ~ babies$gestation)

Residuals:
    Min      1Q  Median      3Q     Max
-49.751 -11.047   0.046   9.902  53.249

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -36.43018    9.21722  -3.952 8.19e-05 ***
```

```
babies$gestation   0.55936    0.03299  16.958  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.38 on 1206 degrees of freedom
  (13 observations deleted due to missingness)
Multiple R-squared:  0.1925,    Adjusted R-squared:  0.1919
F-statistic: 287.6 on 1 and 1206 DF,  p-value: < 2.2e-16
```

Conclusion Statement:

Question 13. Type the regression equation and interpret the slope.