GOUP 10 ASSIGMENT I

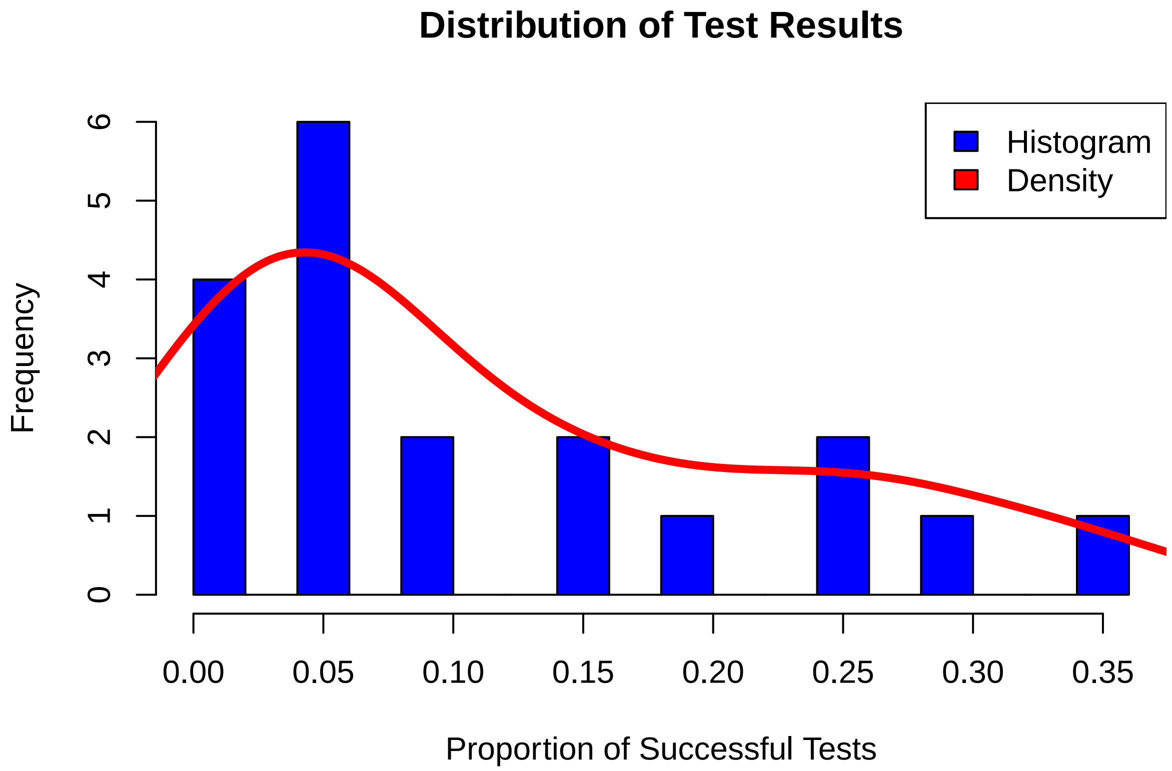
Group-Homework-statistics for Data science

2024-01-16

```
setwd("C:/Users/HP/Documents")
```

#1) Two methods are applied to train patients with senile dementia to care for themselves. After the completion of the training, patients are asked to take 20 tests involving activities of daily living. The response from each patient is the proportion of his or her tests that are successful.

```
Group1=c(0.05, 0.15, 0.35, 0.25, 0.20, 0.05, 0.10, 0.05, 0.30, 0.05, 0.25)
Group2=c(0, 0.15, 0, 0.05, 0, 0, 0.05, 0.10)
# Combine the data into a single vector to visualize  Distribution of tests result
all_data<- c(Group1, Group2)
hist(all_data, breaks = 14, col = "blue", xlab = "Proportion of Successful Tests", main = "Distribution

# Add a density plot
lines(density(all_data), col = "red", lwd = 4)

# Add a legend
legend("topright", legend = c("Histogram", "Density"), fill = c("blue", "red"))
```



```
# Perform the independent sample t-test
t_test_result <- t.test(Group1, Group2)
t_test_result
```

```
##
##  Welch Two Sample t-test
##
## data:  Group1 and Group2
## t = 3.0583, df = 15.485, p-value = 0.007736
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.03655926 0.20321347
## sample estimates:
## mean of x mean of y
## 0.1636364 0.0437500
```

The difference between the two groups is statistically significant. the p-value is less than a chosen Moderately significance level (commonly 0.05).which is p-value = 0.007736.So we reject the null hypothesis and conclude that there is a significant difference between the two groups.

##Transform the data to Y = sqrt(X).

```
#Original data()
Group1=c(0.05, 0.15, 0.35, 0.25, 0.20, 0.05, 0.10, 0.05, 0.30, 0.05, 0.25)
Group2=c(0, 0.15, 0, 0.05, 0, 0, 0.05, 0.10)
# Transformed Data
Group1_sqrt<-sqrt(Group1)
Group2_sqrt<-sqrt(Group2)
```

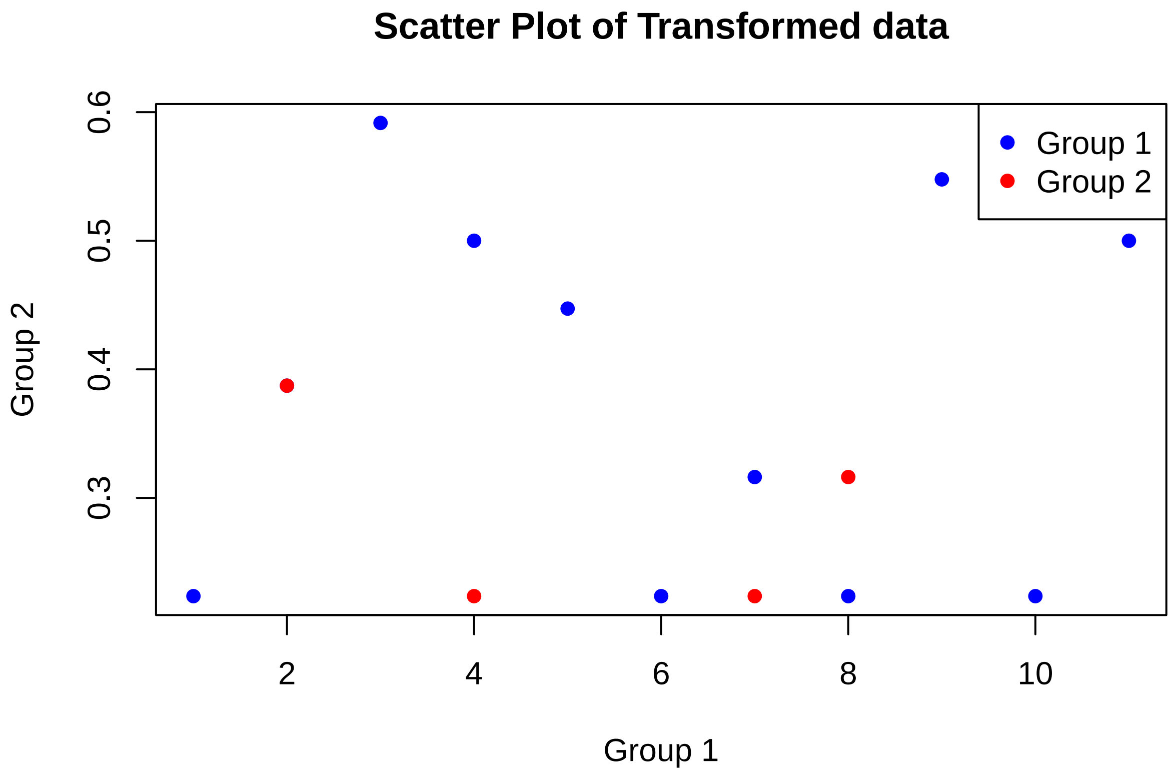
```r
cat("Group1: ", Group1_sqrt, "\n")
```

```
## Group1:  0.2236068 0.3872983 0.591608 0.5 0.4472136 0.2236068 0.3162278 0.2236068 0.5477226 0.223606
```

```r
cat("Group2: ", Group2_sqrt, "\n")
```

```
## Group2:  0 0.3872983 0 0.2236068 0 0 0.2236068 0.3162278
```

```r
# Assuming Group1_sqrt and Group2_sqrt contain the transformed data for Group 1 and Group 2 respectivel

# Create a scatter plot for Group 1
plot(Group1_sqrt, main = "Scatter Plot of Transformed data", xlab = "Group 1", ylab = "Group 2", col =

# Add points for Group 2 to the existing plot
points(Group2_sqrt, col = "red", pch = 16)

# Add a legend
legend("topright", legend = c("Group 1", "Group 2"), col = c("blue", "red"), pch = 16)
```



##Coduct t-test for Transformed data

```r
t_test_resul2=t.test(Group1_sqrt,Group2_sqrt)
t_test_resul2
```

```
##
##  Welch Two Sample t-test
##
## data:  Group1_sqrt and Group2_sqrt
## t = 3.2848, df = 14.097, p-value = 0.005381
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.08220025 0.39093254
## sample estimates:
## mean of x mean of y
## 0.3804089 0.1438425
```

##Compare The t_test result of original and Trasformed Data The p-value obtained after conducting the t-test on the original data (0.007736) and the transformed data (0.005381) indicates the probability of observing the observed difference in means (or more extreme) under the null hypothesis that there is no difference between the two groups.

A lower p-value suggests stronger evidence against the null hypothesis. In this case, both p-values (0.007736 and 0.005381) are less than the commonly used significance level of 0.05. Therefore, we would reject the null hypothesis and conclude that there is a statistically significant difference between the two groups, both in the original data and the transformed data.

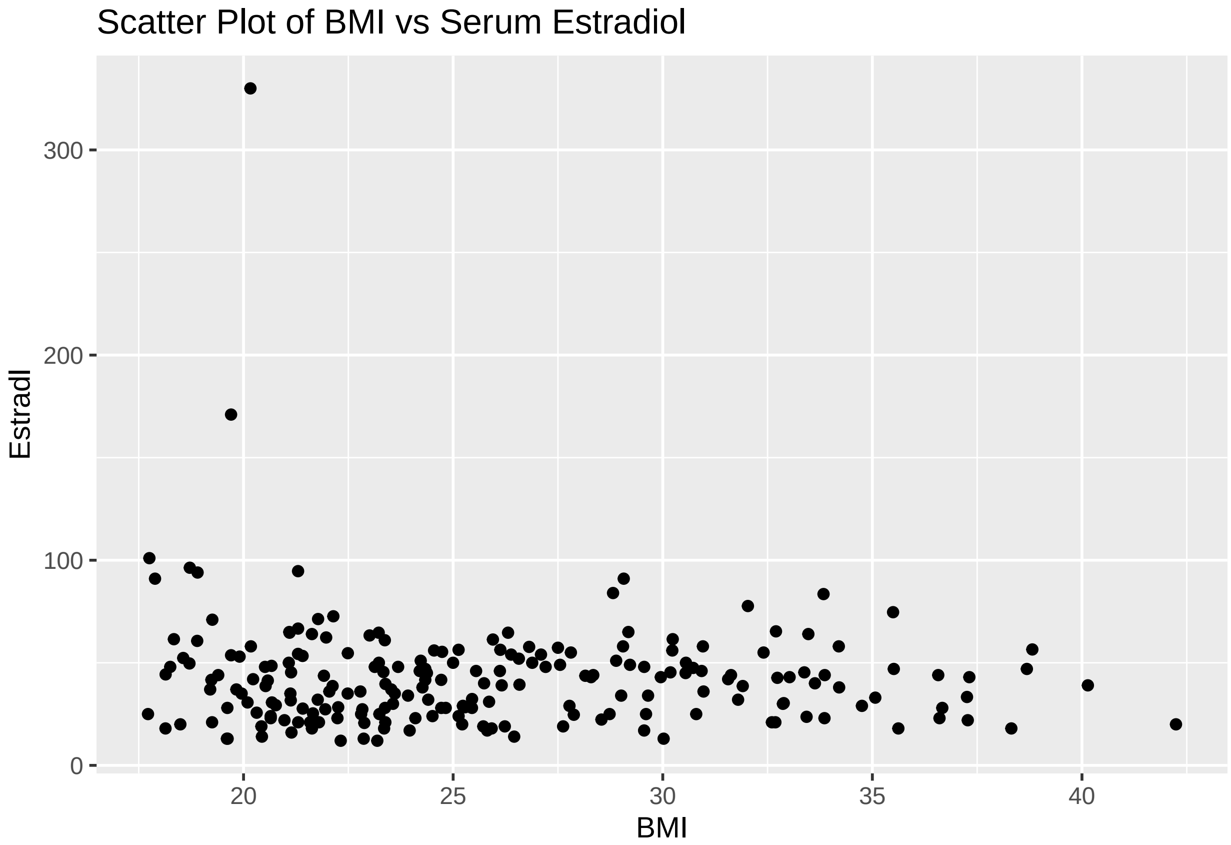
```r
data=read.csv("E:\\Data Science Course\\Advanced Statistics\\ESTRADL.csv")
head(data)
```

```
##   Id Estradl Ethnic Entage Numchild Agefbo Anykids Agemenar     BMI  WHR
## 1  2   94.00      0     30        0      0       0       11 18.9038 0.70
## 2  2   14.00      0     23        0      0       0       15 20.4386 0.70
## 3  3   28.33      0     21        0      0       0       13 22.2578 0.75
## 4  6   38.67      0     33        0      0       0       14 20.5265 0.73
## 5  8   41.67      0     31        0      0       0       13 24.3356 0.75
## 6  9   44.33      0     36        2     27       1       11 18.1416 0.71
```

```r
BMI <- data$BMI
Estradl <- data$Estradl
correlation <- cor(BMI, Estradl, method = "pearson")
correlation
```

```
## [1] -0.09670012
```

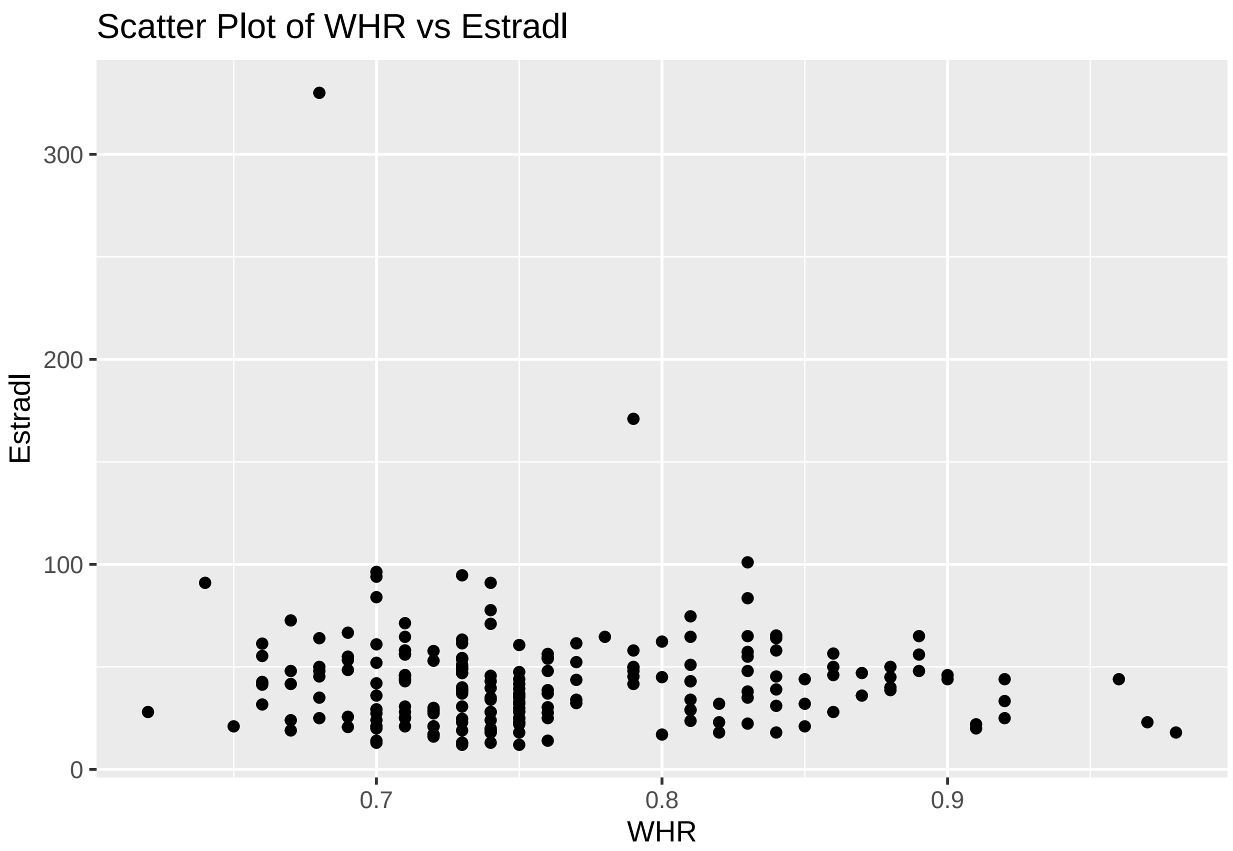
```r
library(ggplot2)

# Assuming our data frame is called 'data' and the columns are named 'BMI' and 'serum_estradiol'
# Create a scatter plot
ggplot(data, aes(x = BMI, y = Estradl)) +
  geom_point() +  # Add points
  labs(x = "BMI", y = "Estradl") +  # Label axes
  ggtitle("Scatter Plot of BMI vs Serum Estradiol")  # Add title
```



```
WHR <- data$WHR
Estradl <- data$Estradl
correlation <- cor(WHR, Estradl, method = "pearson")
correlation
```

```
## [1] -0.05114826
```

```
# Create a scatter plot
ggplot(data, aes(x = WHR, y = Estradl)) +
  geom_point() +  # Add points
  labs(x = "WHR", y = "Estradl") +  # Label axes
  ggtitle("Scatter Plot of WHR vs Estradl")  # Add title
```



## In summary, based on these correlation coefficients, we can conclude that there is a weak, but not significant, negative linear relationship between WHR/BMI and serum estradiol, hence the relationship is not strong, and the correlation coefficient being close to zero suggests that there may not be a significant linear relationship between a given variables.so there is no a crude association between either measure of adiposity (BMI, WHR), considered separately, and serum estradiol.

#(b) Are these relationships similar for Caucasian and African-American women?

```
# Subset the data for Caucasian and African-American women
caucasian_data <- subset(data, Ethnic == 0)  # Assuming 0 represents Caucasian
african_american_data <- subset(data, Ethnic == 1)  # Assuming 1 represents African-American
# Correlation analysis for Caucasian women
correlation_caucasian <- cor(caucasian_data$WHR, caucasian_data$Estradl)

# Correlation analysis for African-American women
correlation_african_american <- cor(african_american_data$WHR,african_american_data$Estradl)
# Print correlation coefficients
print(paste("Correlation coefficient for Caucasian women:", correlation_caucasian))
```

```
## [1] "Correlation coefficient for Caucasian women: -0.129943455591677"
```

```
print(paste("Correlation coefficient for African-American women:", correlation_african_american))
```

```
## [1] "Correlation coefficient for African-American women: 0.124898275192638"
```

.For Caucasian women, the correlation coefficient is approximately -0.1299. This suggests a weak negative correlation between WHR and serum estradiol among Caucasian women, meaning that as WHR increases, serum estradiol tends to decrease slightly.

.For African-American women, the correlation coefficient is approximately 0.1249. This indicates a weak positive correlation between WHR and serum estradiol among African-American women, implying that as WHR increases, serum estradiol tends to increase slightly. ##These results suggest that the relationship between WHR and serum estradiol differs between Caucasian and African-American women

```
# Correlation analysis for Caucasian women
correlation_caucasian <- cor(caucasian_data$BMI, caucasian_data$Estradl)

# Correlation analysis for African-American women
correlation_african_american <- cor(african_american_data$BMI,african_american_data$Estradl)
# Print correlation coefficients
print(paste("Correlation coefficient for Caucasian women:", correlation_caucasian))
```

```
## [1] "Correlation coefficient for Caucasian women: -0.205060527334884"
```

```
print(paste("Correlation coefficient for African-American  women:",correlation_african_american))
```

```
## [1] "Correlation coefficient for African-American  women: 0.108588273256186"
```

.For Caucasian women, the correlation coefficient of approximately -0.205 suggests a moderate negative correlation between BMI and serum estradiol. This indicates that as BMI increases, serum estradiol tends to decrease moderately.

.For African-American women, the correlation coefficient of approximately 0.109 suggests a weak positive correlation between BMI and serum estradiol. This implies that as BMI increases, serum estradiol tends to increase slightly. ##Similar to the WHR results, these findings indicate that there is a difference in the relationship between BMI and serum estradiol among Caucasian and African-American women.

Do the relationships between the adiposity measures and serum estradiol persist after controlling for the other breast-cancer risk factors in list items 1 to 6?

To assess whether the relationships between the adiposity measures (BMI and WHR) and serum estradiol persist after controlling for other breast cancer risk factors, we perform multiple linear regression

```
# Fit a multiple linear regression model
model <- lm(Estradl~BMI+WHR+Ethnic + Entage + Numchild + Agefbo + Anykids + Agemenar, data=data)
summary(model)
```

```
##
## Call:
## lm(formula = Estradl ~ BMI + WHR + Ethnic + Entage + Numchild +
##     Agefbo + Anykids + Agemenar, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.111 -14.857  -4.646  10.300 269.490
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.2737    22.5206   1.611 0.108809
## BMI          -0.1967     0.4349  -0.452 0.651595
## WHR           7.9121    34.0061   0.233 0.816256
## Ethnic      -15.9056     4.5085  -3.528 0.000518 ***
## Entage        0.6503     0.3863   1.683 0.093829 .
## Numchild      0.5181     2.8442   0.182 0.855628
## Agefbo       -0.3570     0.2614  -1.366 0.173482
## Anykids       2.9781     4.3871   0.679 0.498023
## Agemenar      0.0908     0.1785   0.509 0.611635
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.52 on 202 degrees of freedom
## Multiple R-squared:  0.09567,    Adjusted R-squared:  0.05986
## F-statistic: 2.671 on 8 and 202 DF,  p-value: 0.008252
```

Interpretation of Coefficients:

‘BMI’ and ‘WHR’

Neither BMI nor WHR appears to be statistically significant predictors of serum estradiol. This suggests that after controlling for other variables, there is no significant relationship between these adiposity measures and serum estradiol. ‘Ethnic’: Ethnicity (African-American vs. Caucasian) is a statistically significant predictor of serum estradiol. AfricanAmerican ethnicity is associated with a decrease in serum estradiol levels compared to Caucasian ethnicity. ‘Entage’: Age is not statistically significant at the conventional significance level of 0.05, ‘Numchild’, ‘Agefbo’, ‘Anykids’, ‘Agemenar’: None of these variables are statistically significant predictors of serum estradiol. Overall, the analysis suggests that after controlling for other variables, BMI and WHR are not significantly associated with serum estradiol levels in this dataset. However, ethnicity (African-American vs. Caucasian) does appear to be a significant predictor of serum estradiol levels.

d). One debate in the breast-cancer literature is whether overall adiposity (BMI) or central (abdominal) adiposity (WHR) is a better indicator of breast-cancer risk. Perform analyses to inform the debate as to which measure of adiposity is more closely related to serum estradiol either crudely or after adjusting for other breast-cancer risk factors.

Separate regression analyses as crude and adjusted

```r
# Crude Analysis for BMI
# Simple linear regression for BMI
model_bmi <- lm(Estradl~ BMI, data = data)
summary(model_bmi)
```

```
##
## Call:
## lm(formula = Estradl ~ BMI, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.993 -16.077  -2.404   9.132 284.291
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  55.9255     9.5005   5.887 1.55e-08 ***
## BMI          -0.5067     0.3607  -1.405    0.162
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.32 on 209 degrees of freedom
## Multiple R-squared:  0.009351,   Adjusted R-squared:  0.004611
## F-statistic: 1.973 on 1 and 209 DF,  p-value: 0.1616
```

```r
#Crude Analysis for WHR
# Simple linear regression for WHR
model_whr <- lm(Estradl ~ WHR, data = data)
summary(model_whr)
```

```
##
## Call:
```

```
## lm(formula = Estradl ~ WHR, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.482 -17.231  -2.962   9.287 285.461
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    58.92      21.77   2.706  0.00736 **
## WHR           -21.15      28.56  -0.740  0.45988
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.41 on 209 degrees of freedom
## Multiple R-squared:  0.002616,   Adjusted R-squared:  -0.002156
## F-statistic: 0.5482 on 1 and 209 DF,  p-value: 0.4599
```

```
# Adjusted analysis for BMI
lm_BMI_adjusted <- lm(Estradl~BMI+WHR+Ethnic + Entage + Numchild + Agefbo + Anykids + Agemenar, data=da
summary(lm_BMI_adjusted)
```

```
##
## Call:
## lm(formula = Estradl ~ BMI + WHR + Ethnic + Entage + Numchild +
##     Agefbo + Anykids + Agemenar, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.111 -14.857  -4.646  10.300 269.490
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.2737    22.5206   1.611 0.108809
## BMI          -0.1967     0.4349  -0.452 0.651595
## WHR           7.9121    34.0061   0.233 0.816256
## Ethnic      -15.9056     4.5085  -3.528 0.000518 ***
## Entage        0.6503     0.3863   1.683 0.093829 .
## Numchild      0.5181     2.8442   0.182 0.855628
## Agefbo       -0.3570     0.2614  -1.366 0.173482
## Anykids       2.9781     4.3871   0.679 0.498023
## Agemenar      0.0908     0.1785   0.509 0.611635
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.52 on 202 degrees of freedom
## Multiple R-squared:  0.09567,    Adjusted R-squared:  0.05986
## F-statistic: 2.671 on 8 and 202 DF,  p-value: 0.008252
```