**Group 6­ - Executive Summary (Chunyu Luo, Kechen Lu, Steven Yang, Sabrina Liu)**

**Introduction/Motivation:**

Our goal of the project is to come up with a cost-efficient way to estimate the percentage of body fat based on a real data set from 252 men. We plan to visualize the data through histogram and scatter plot, comparing variables and using regression analysis to select significant predictors. Our final rule of thumb is Body Fat Pct = 0.61544\*Abdomen - 37.95511

**Background Information/Data Cleaning**

1. Based on the data, we could observe that the median of body fat is 19%, with a mean 18.94%,

2. Outliers detection: we excluded three suspicious data which rarely occur in real life.

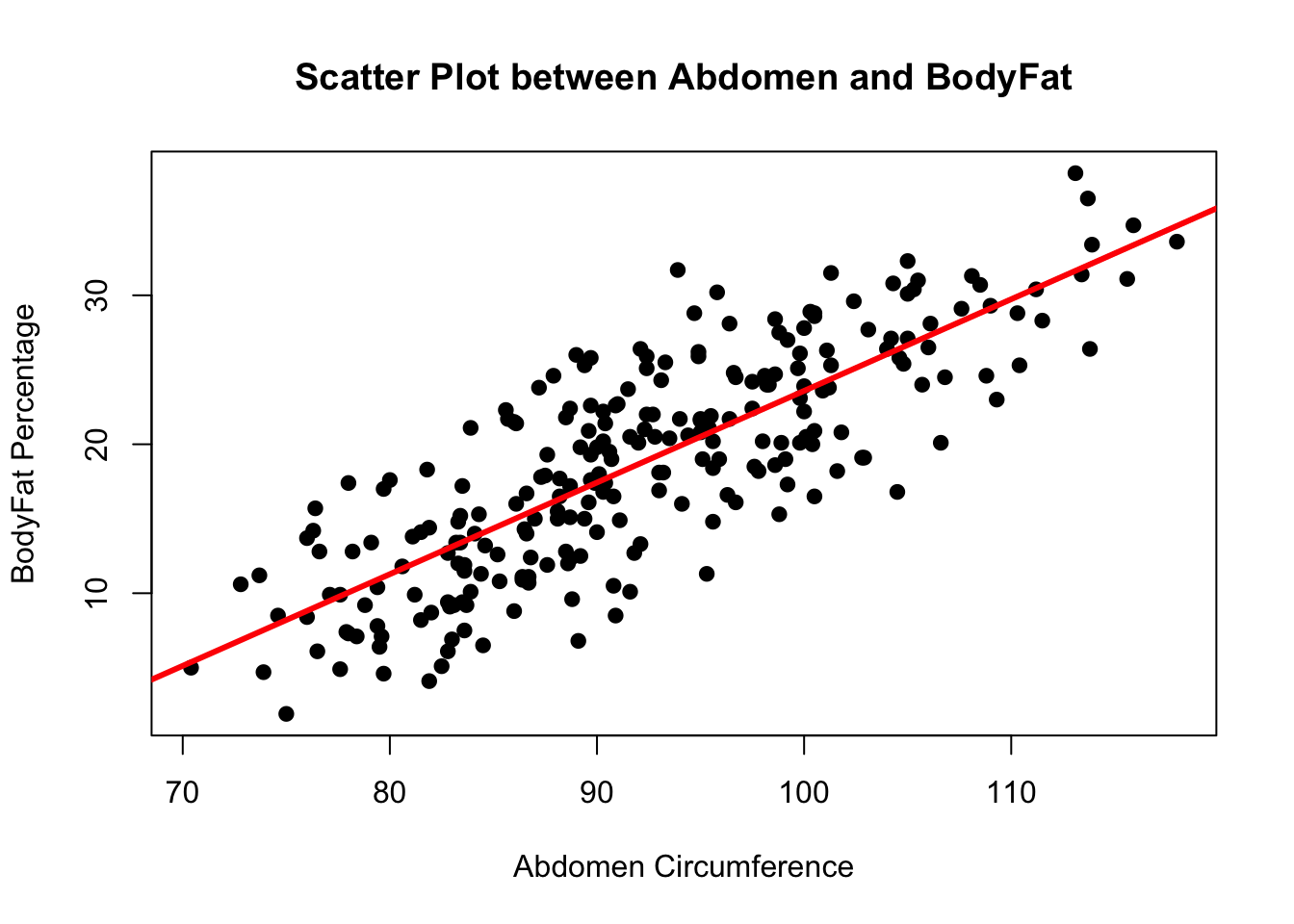
1. We removed the 182nd individual with body fat 0%, because it is impossible in real life.
2. We removed the 42nd individual with 29.5 inches in height, because it is highly unlikely.
3. We removed the 39th individual because it has unusually large weight, abdomen, adiposity, and hip compared to the rest of our data.

**Motivation for Model/Choosing Model/Final Model/Rule of Thumb**

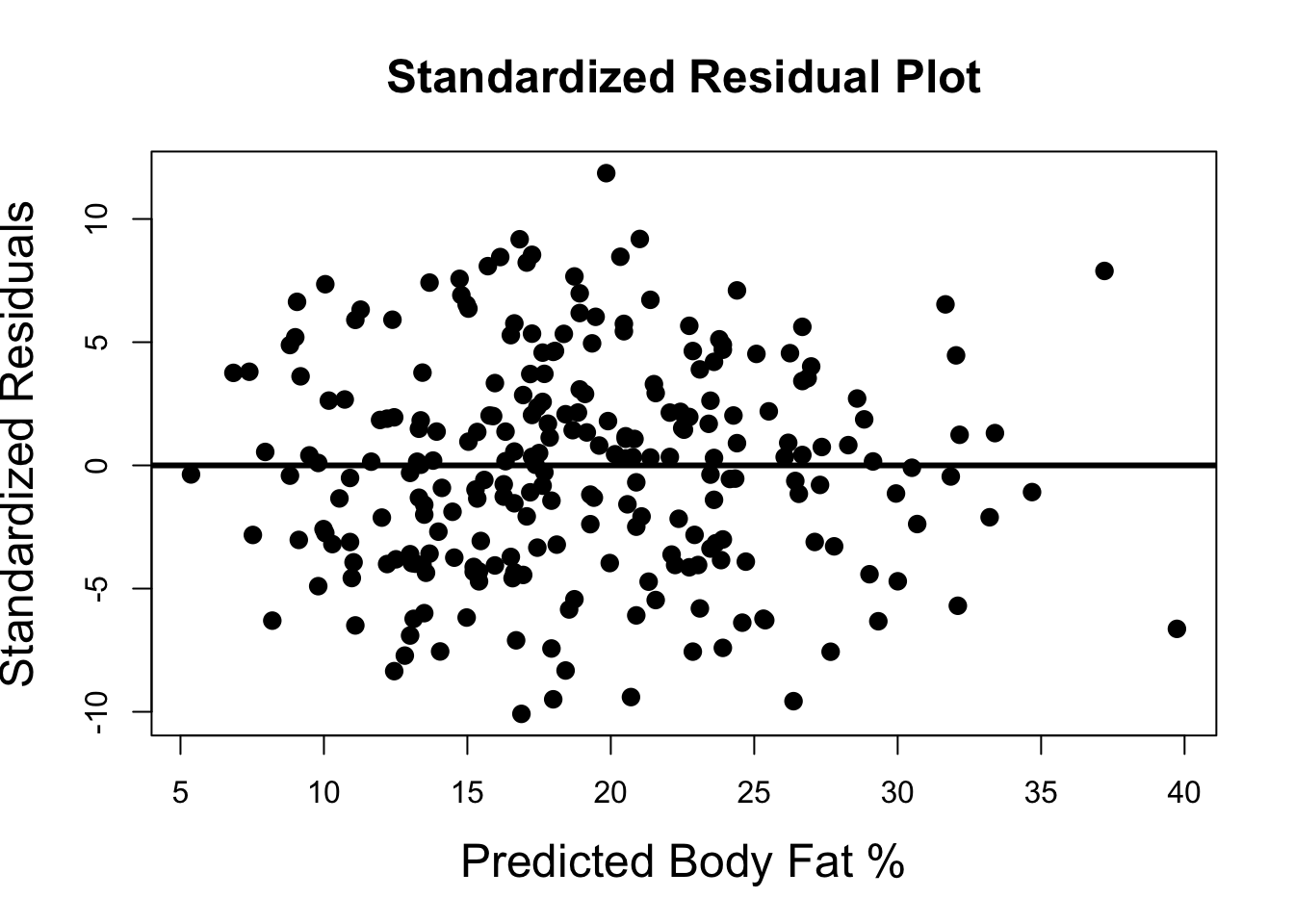
1. Rule of thumb: **Body Fat Pct = 0.61544\*Abdomen - 37.95511**

2. Example usage of the final model: A man with 120cm abdomen circumference is expected to have a body fat percentage of 35.89769% based on our model. His 95% prediction interval is between 27.216% and 44.579%.

3. Model interpretation: Our estimated coefficients are 0.6154 and -37.9551, which are in the units of percentage per centimeter and percentage, respectively. This means that for every 1 centimeter increase in abdomen circumference, the model predicts that body fat % will increase, on average, by 0.6154 percentage.

4. Reasons for choosing the final model: We chose this model because of the following reasons. First, from background research, we noticed that the percentage of body fat increases evidently as abdomen circumference increases. Second, from analysing the SLR model, R2 value for abdomen (0.6725) is the highest among all predictors, which indicates 67.25% of the variability in percentage of body fat is accounted for by abdomen circumference. Finally, we create an MLR model with all these three predictors. We see from the t-statistics (10.472) and p-value (<2e-16) of Abdomen is most statistically significant compared to the other two. Then we remove each one of them to build a second MLR model with two predictors. We see that when we remove the abdomen circumference, the R2 decreases the most (about 0.139). Therefore, we can determine that abdomen circumference is the best predictor.

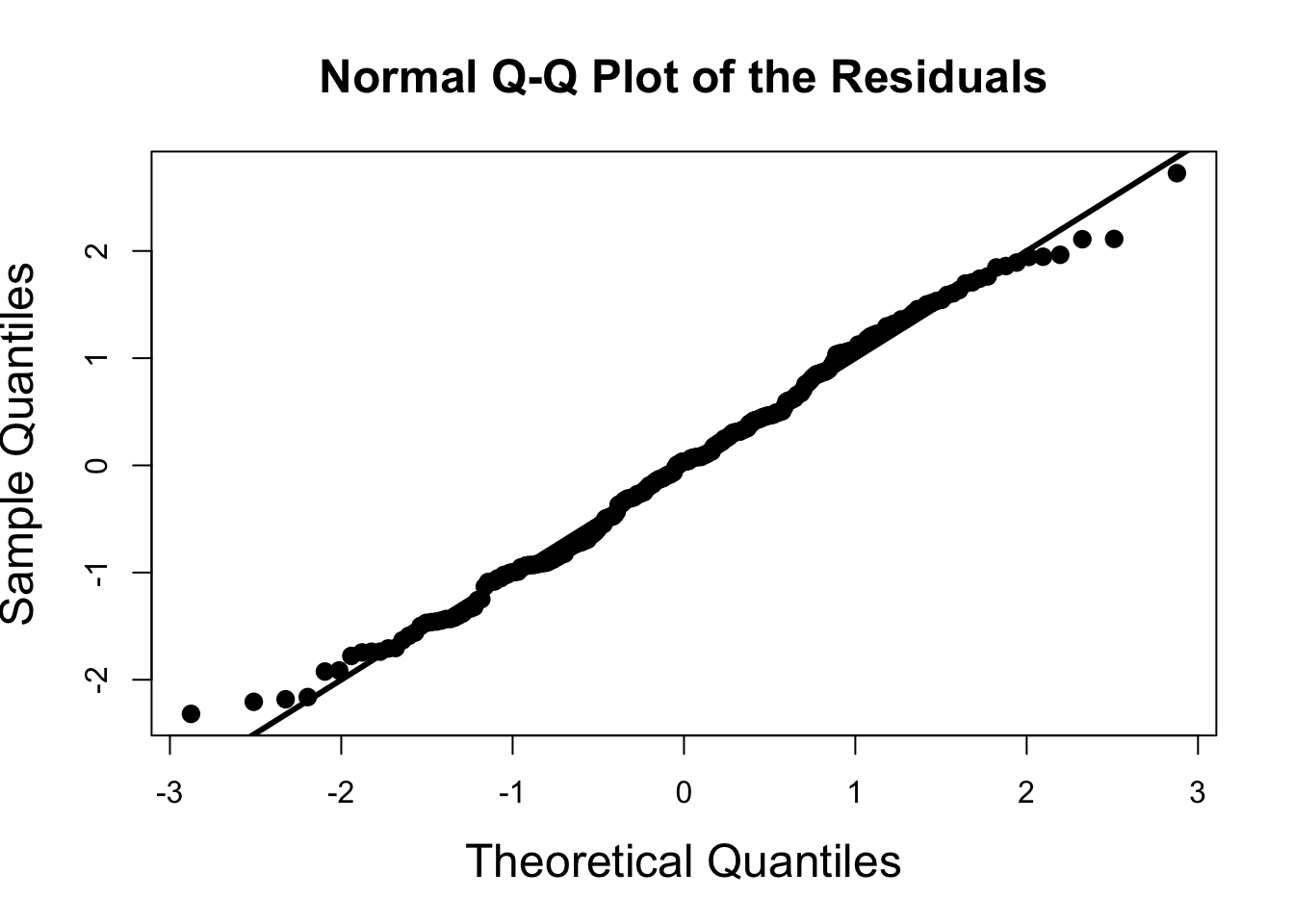
**Statistical Analysis/Hypothesis Testing/Inference/**

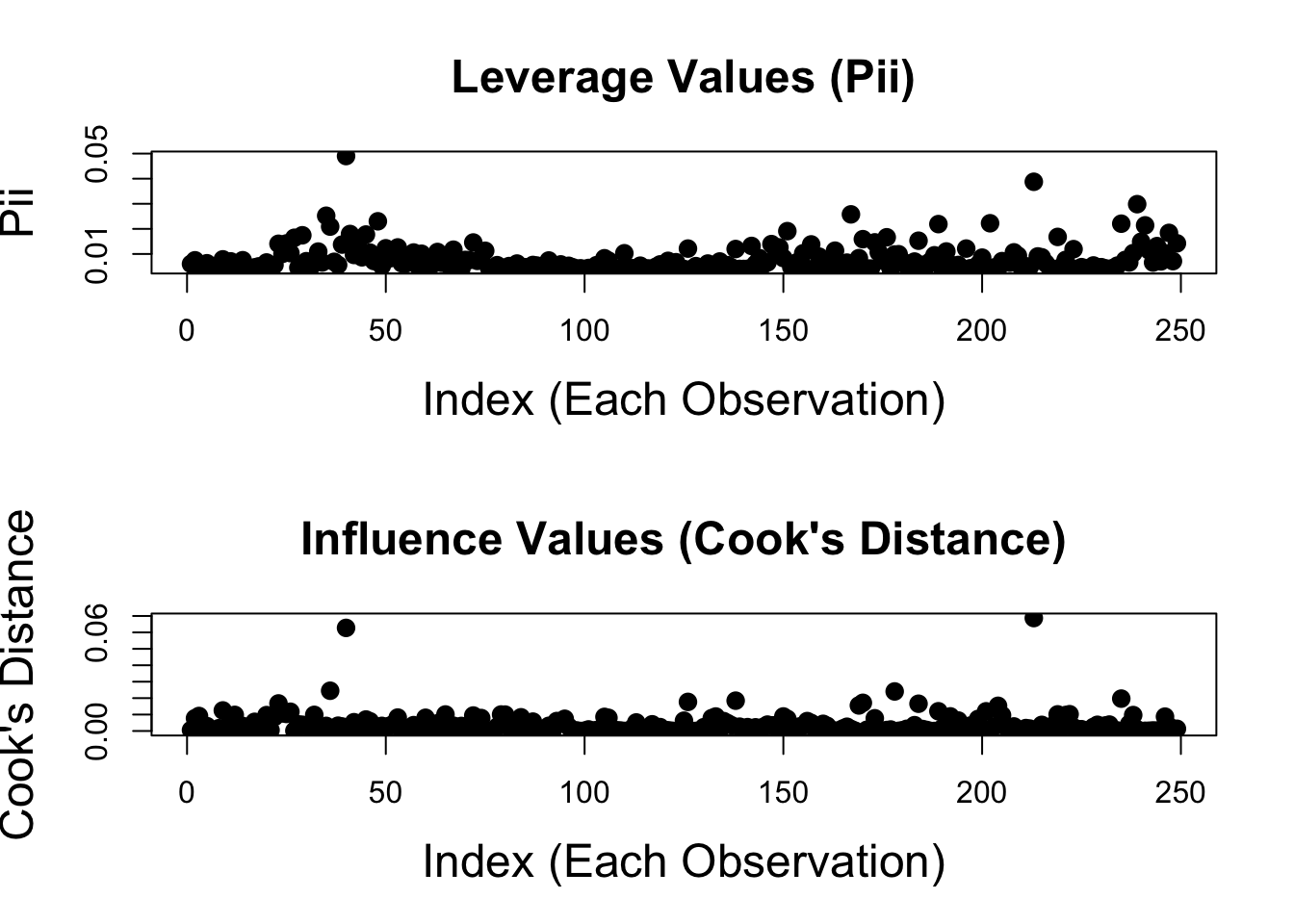
1. We conducted the following test to see whether the predictor abdomen circumference is significant in predicting the outcome. The null hypothesis states the slope is equal to 0, and alternative hypothesis states the slope is not equal to 0. We performed a t-test and observed that t-statistic is 21.74, p-value is less than 2e-16. From our statistical test, we have strong evidence to reject the null hypothesis. We can conclude that there is a linear relationship between Abdomen circumference and Body Fat Percentage.
2. We found our R2 to be 0.6586, which implies 65.86% of the variability in percentage of body fat is accounted for by abdomen circumference.
3. The estimated slope and intercept are 0.61544 and -37.95511. We are 95% confident that the true Intercept for our model is between(-43.1199043, -32.7903198) and the true slope coefficient is between (0.5596846, 0.6711946). Also, based on the 95% CI, we can reject the null hypothesis. In other words, there is a linear relationship between Abdomen circumference and Body Fat Percentage.

**Model Diagnostics**

1. Model diagnostics: we diagnose the SLR assumptions with a residual plot, cook’s distance and a QQ plot.

2. Model diagnostics explanation:

**a.** We checked the following four assumptions for SLR. First, we checked linearity using residual plot. Because there are no obvious non-linear trends in the residual plot, we believe linearity is reasonable. Second, we checked normality using the qq plot. Because the points in the QQ plot hug the 45 degree line very closely, we conclude normality is reasonable. However, there may be possibly skinny tail or right skew issues (see -1 to 0 region). Third, we checked equal variance of errors based on residual plot. There is no obvious pattern in the residual plot, so homoscedasticity is plausible.

**b.** We also looked at three types of outliers in regression models: Outliers in Y, Leverage and Influential Points. We used the pii measures to detect leverage points. We found two leverage points: the 40th and 213th observation (index after first deletion of outliers) and we removed these outliers. After rerunning the diagnostics checks, we found dramatic changes in the estimated slope and intercept terms.

**Model Strengths/Weaknesses**

1. Linearity seems reasonable based on both residual plot and scatter plot.

2. Normally distributed errors seem plausible from the QQ plot diagnostic. However, there may be possibly skinny tail or right skew issues. The skewness in the QQ plot could be one weakness.

3. Our model is straightforward and easy to understand. It can be easily updated with new data.

**Conclusion/Discussion**

The result of our analysis shows that the percentage of body fat has a positive linear relationship with abdomen circumference. Our model provides a simple, convenient way to estimate the body fat based on abdomen circumference. Our best possible rule of thumb is **Body Fat Pct = 0.6\*Abdomen-38.**

**Reference: N/A**

**Contributions:**

Chunyu Luo:

Presentation Slides: Diagnostic, Strength and Weaknesses

Executive Summary: Reviewed and revised **Strength and Weaknesses**

Sabrina Liu:

Presentation Slides: Histograms explanation and Outliers detecting

Executive Summary: Check and review work

Kechen Lu:

Presentation Slides: Analysing raw data and correlation analysis

Executive Summary: Edit summary

Steven Yang:

Presentation Slides: Statistical Modeling, Final Model and Use of Model, Hypothesis Testing

Executive Summary: Reviewed and revised **Choosing Model/Final Model, Background Information/Data Cleaning, Model Diagnostics**