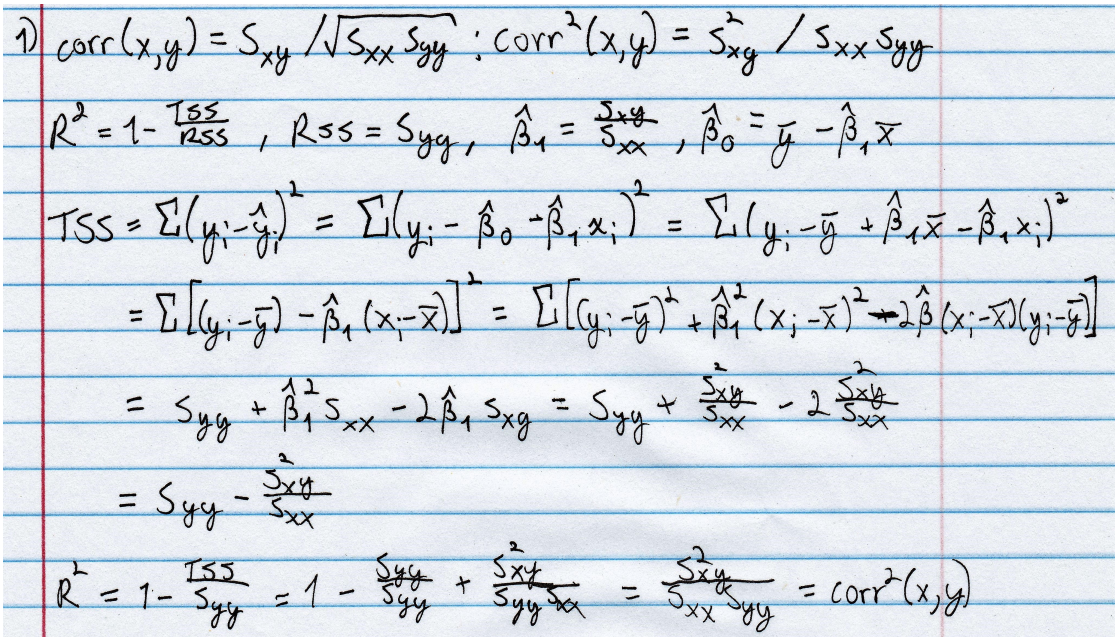
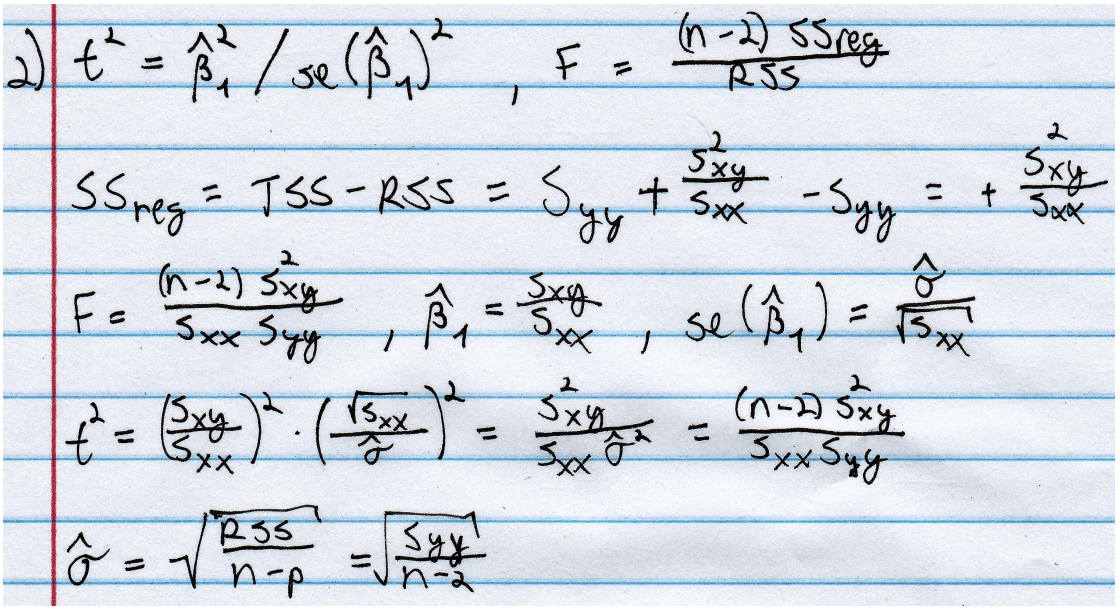
Kyle Salitrik (kps168)
STAT 462
Homework 3 Report

Theoretical Solutions





Applied Analysis

Part A

Same reduction on the data (as in HW1 and HW2) was performed:

- Samples below 2% and above 40% body fat were excluded. The minimum was obtained through the chart on the (ACE) website and maximum was set to limit outliers.
- The minimum height accepted was cut to be over 50 inches to remove outliers.
- The maximum weight was also clipped to be below 300 pounds. Although the outlier may have been legitimate, it could have also been an error.

The following table lists the requested quantiles of the Abdomen Circumference. Of these estimates, the 10% and 90% quantiles are the least reliable as they have the least amount of data near them to inform the regression model due to the fact that they are far from the mean.

	10%	25%	50%	75%	90%
Quantile	79.65	84.75	90.90	99.05	105.40
Estimate	10.85454	14.17711	18.18373	23.49333	27.63025

Table 1: Quantiles and Point Estimates

Part B

The tables below list the confidence intervals and prediction intervals for each quantile of the abdomen circumference.

	10%	25%	50%	75%	90%
Lower	9.908185	13.44332	17.59416	22.78713	26.66297
Fit	10.85454	14.17711	18.18373	23.49333	27.63025
Upper	11.8009	14.9109	18.77331	24.19952	28.59752

Table 2: Confidence Intervals

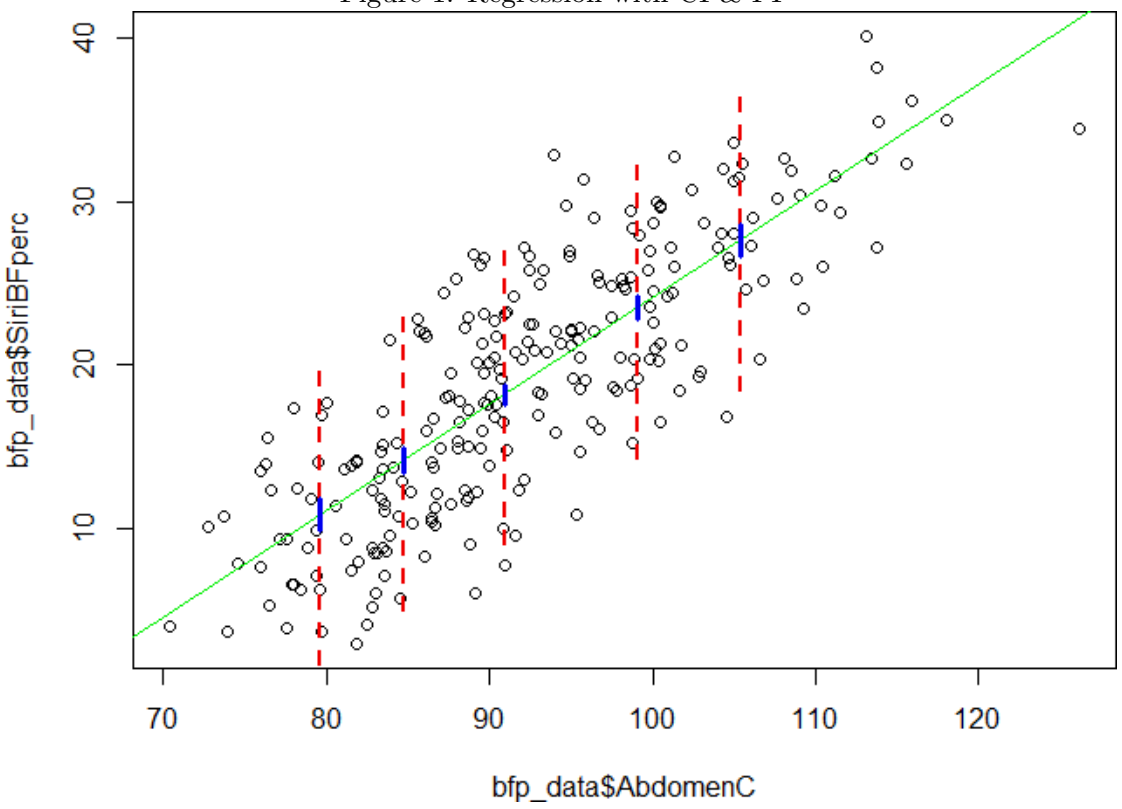
	10%	25%	50%	75%	90%
Lower	1.649528	4.991515	9.008532	14.30989	18.42306
Fit	10.85454	14.17711	18.18373	23.49333	27.63025
Upper	20.05956	23.36271	27.35894	32.67676	36.83744

Table 3: Prediction Intervals

This plot superimposes the regression line (green), confidence interval at each quantile (blue), and prediction interval at each quantile (red). The confidence interval indicates that we are 95% confident that the mean response at each quantile will be within the lower and upper bound shown in Table 2. Similar logic follows for the prediction interval; we are 95% confident that the response at the given quantiles will be within the lower and upper bounds probided by Table 3.

The prediction intervals are larger due to the fact that the confidence intervals are only concerned with the mean of the response at that point. The prediction intervals attempt to take into account the full expected variance of the response at that given point. This means that the prediction interval attempts to predict the maximum and minimum values that will be seen at a given point as opposed to the mean.

Figure 1: Regression with CI & PI



Part C

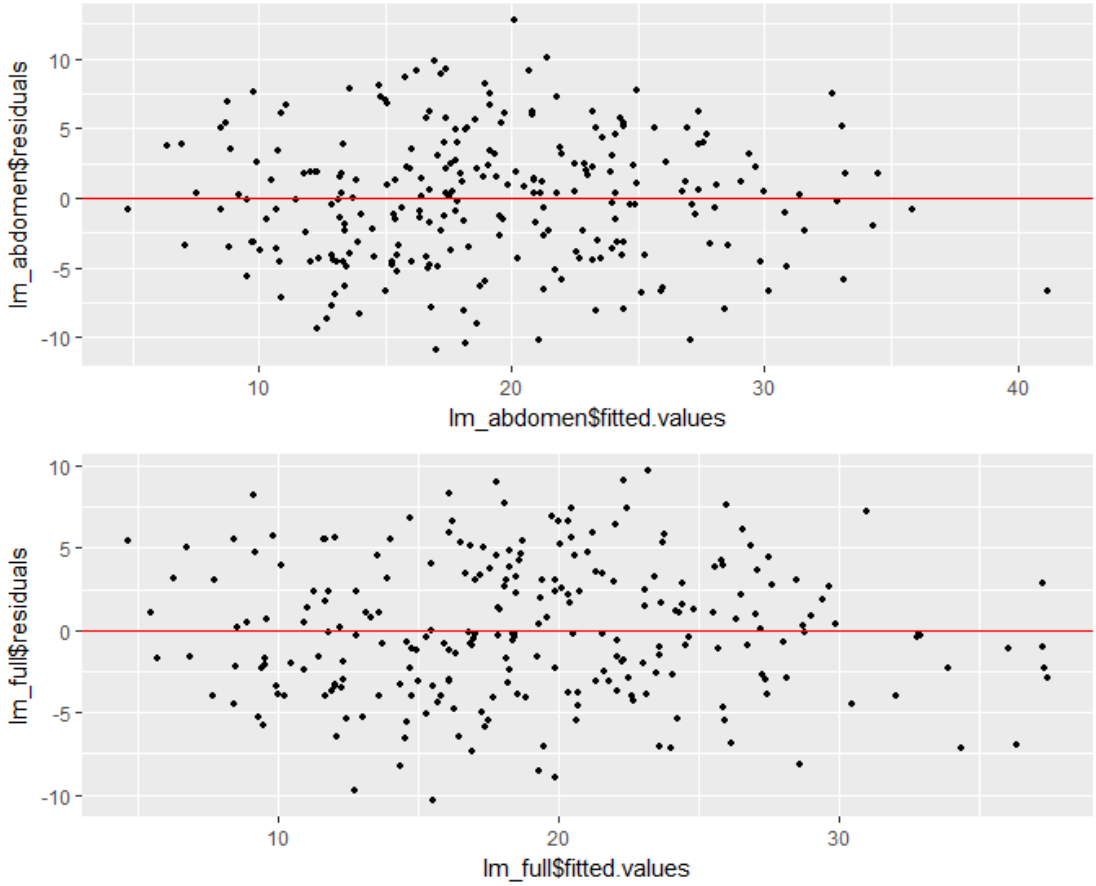
We then fitted a linear regression model with the predictors being: abdomen, neck, chest, hip, thigh, knee, ankle, bicep, forearm, and wrist circumferences, in addition to weight and height. This model will be referred to as the full model. Following this, we can compare the plots of residuals vs fitted values, residual Q-Q plots and perform a Shapiro-Wilk test for both models.

There are four assumptions that we make in linear regression that must be checked by examining the aforementioned plots and the Shaprio-Wilk tests:

- The response and predictors have a linear relationship
- The errors are independent
- The errors are normally distributed
- The errors have equal variance

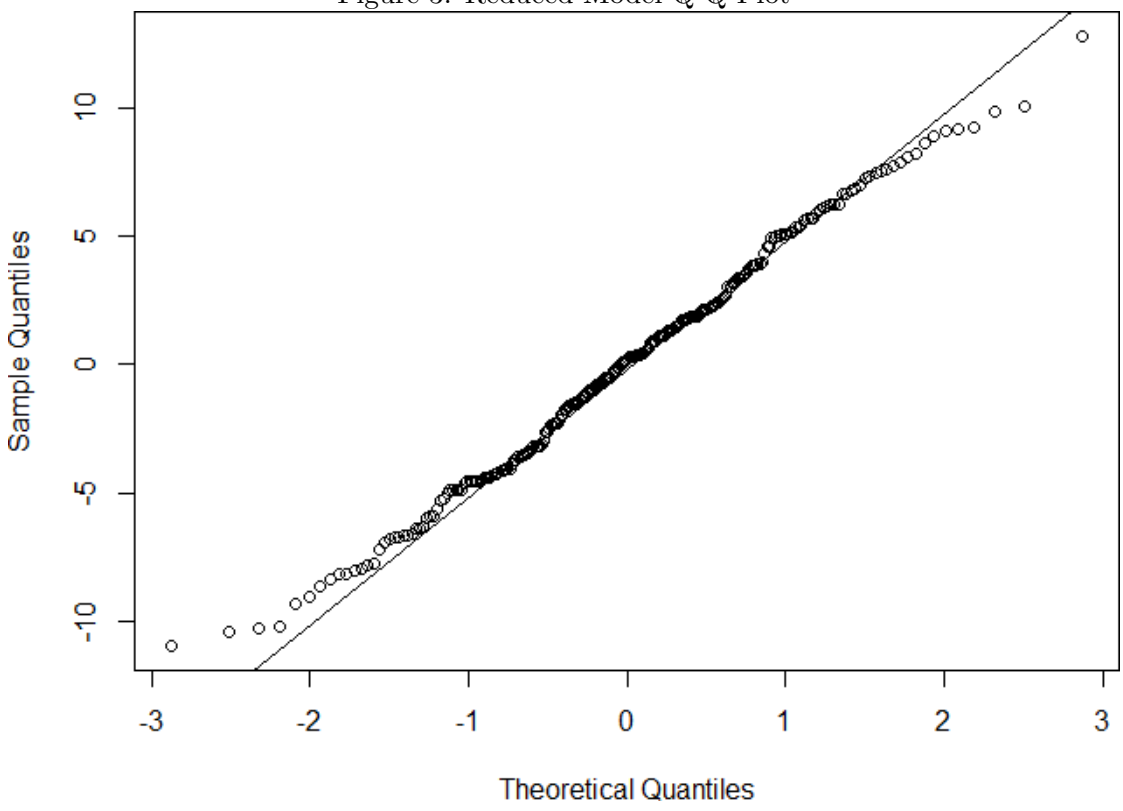
First is the examination of the residuals vs fitted values of each model. As can be seen in figure 2, there does not seem to be a nonlinear trend in either plot, indicating that our first assumption is valid. Examining the spread of the residuals shows that the variance is relatively constant for the reduced model with exception of less than 10% body fat. The full model, however appears to have significantly less variance for measurements over 30% body fat.

Figure 2: Residual vs Fitted Values Plots



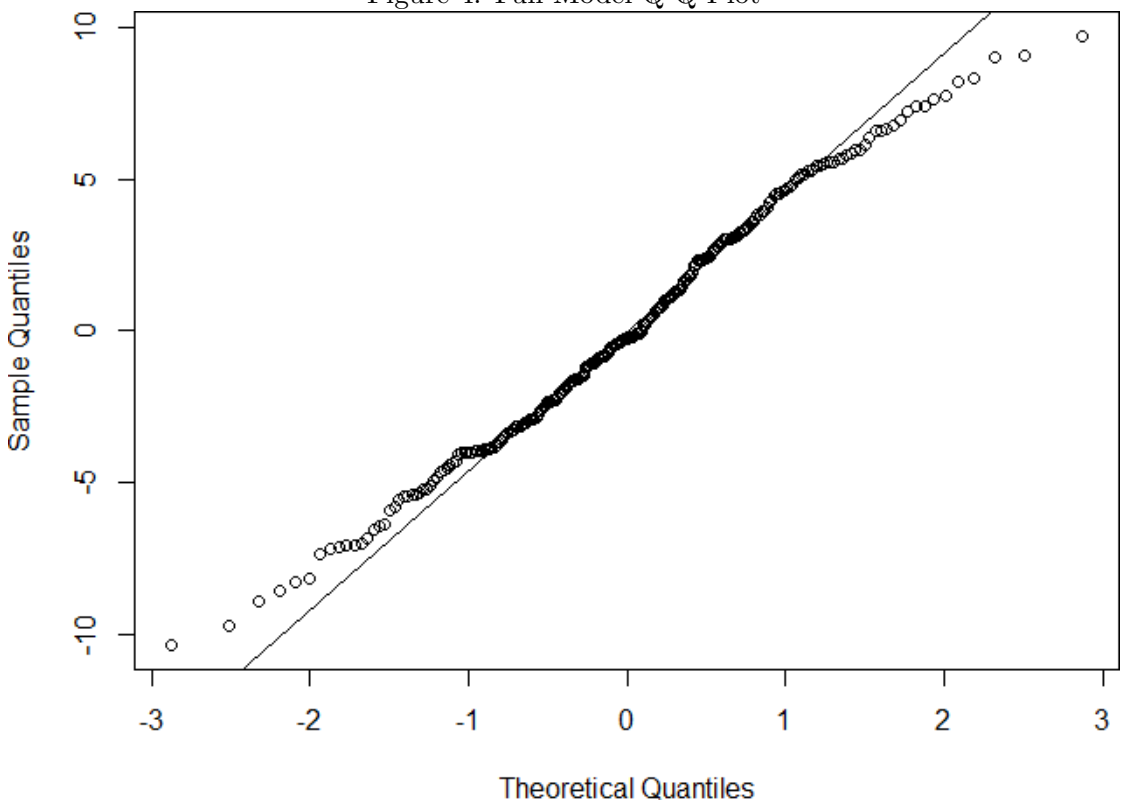
Second, to determine the normality of the errors, the Q-Q plots of the residuals for each model are examined. The plot for the reduced model that follows shows a light tail on both ends, but is otherwise relatively normal.

Figure 3: Reduced Model Q-Q Plot



The Q-Q plot of the full model shows significantly lighter tails than the reduced model, but still does not appear to be skewed indicating that the distribution of the errors is mostly normal, with potential variance issues indicated by the tails.

Figure 4: Full Model Q-Q Plot



Next we perform a Shapiro-Wilk test on both models. This test is for the following hypotheses:

$$H_0 : \text{The errors are normally distributed}$$
$$H_1 : \text{The errors are not normally distributed}$$

As we can see from the following table, the p-values for both models are significantly large enough to not reject the null hypothesis. This indicates that the errors are, in fact, normally distributed. It is worth noting that as we expected from the Q-Q plots, the reduced model has a higher p-value, meaning that the possibility for the errors being normal is higher.

	Full Model	Reduced Model
W	0.99087	0.99373
P	0.1271	0.3954

Table 4: Shapiro-Wilk Tests

The final assumption – that the errors must be independent – is assumed to be true as the measurements were taken from a sample of the population. Also, logically, the likelihood of one person's body fat percentage affecting another person's seems improbable.