STA 2210 Homework 8 (Due on Mon 7/20 by 11:59pm)

The data set BM, a .csv file, contains data on percent body fat and other various measurements of body size, for a sample of 252 men. Write your R codes, in addition to your answer, to the following problems. (Don’t forget to refer to the R reference card to find helpful commands.)

1. **Create a scatterplot for the variables “WEIGHT” and “BODYFAT,” and describe the relationship between the variables.**

plot\_ss(BM$WEIGHT, BM$BODYFAT)

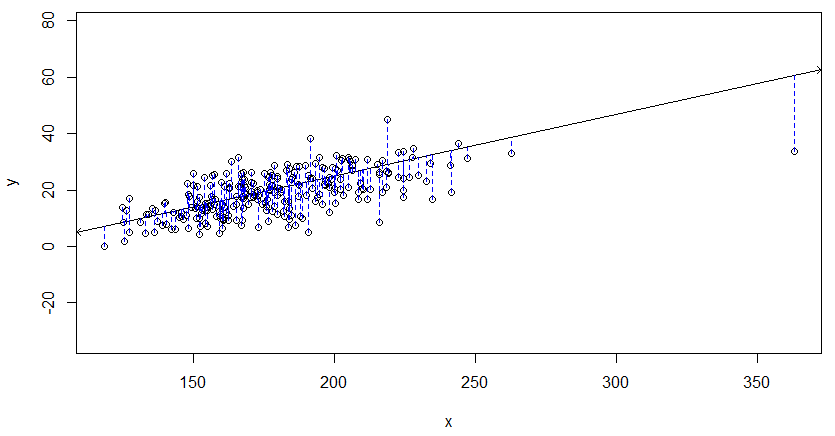
Call:

lm(formula = y ~ x, data = pts)

Coefficients:

(Intercept) x

-18.5580 0.2177

Sum of Squares: 10629.85

There is a moderate-to-strong positive, linear correlation between weight and bodyfat. There appears to only be one outlier, and both sides of the line of best fit appear to have uniform variance.

1. **Compute the correlation between the variables “WEIGHT” and “BODYFAT.”**

cor(BM$BODYFAT, BM$WEIGHT) = 0.6131561

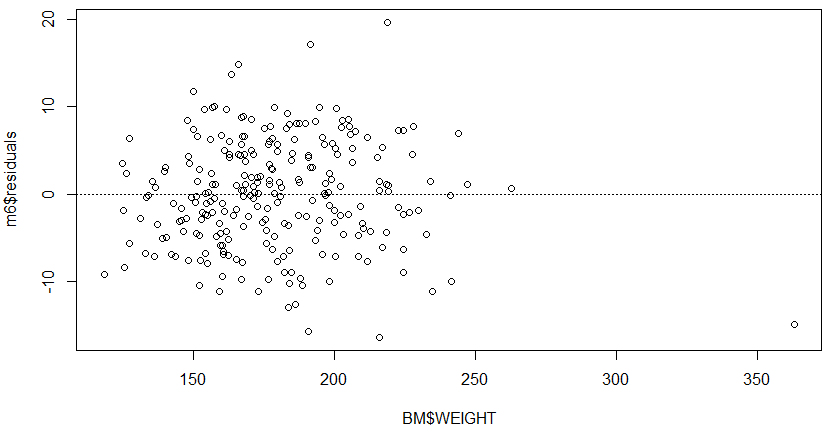
The correlation between weight and bodyfat is 0.61, which is moderately strong and positive.

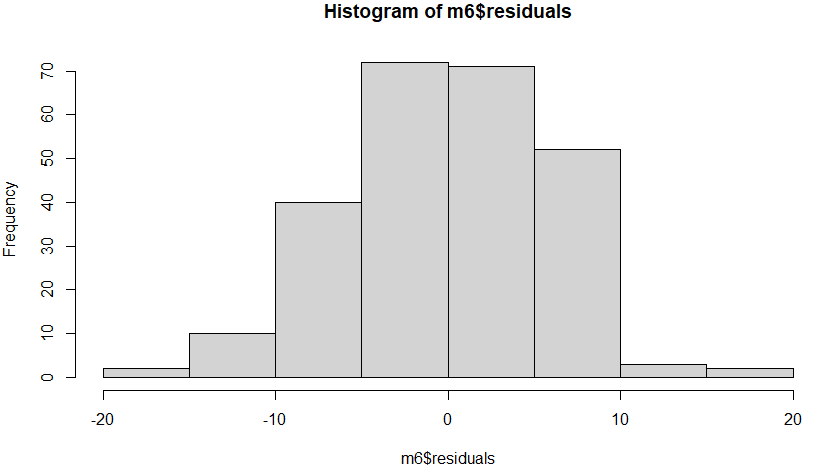
1. **Check each of (1) linearity, (2) nearly normal residuals, and (3) constant variability for WEIGHT as a linear predictor for BODYFAT. Based on your analysis, would a linear model be reliable?**

m6 <- lm(BODYFAT ~ WEIGHT, data = BM)

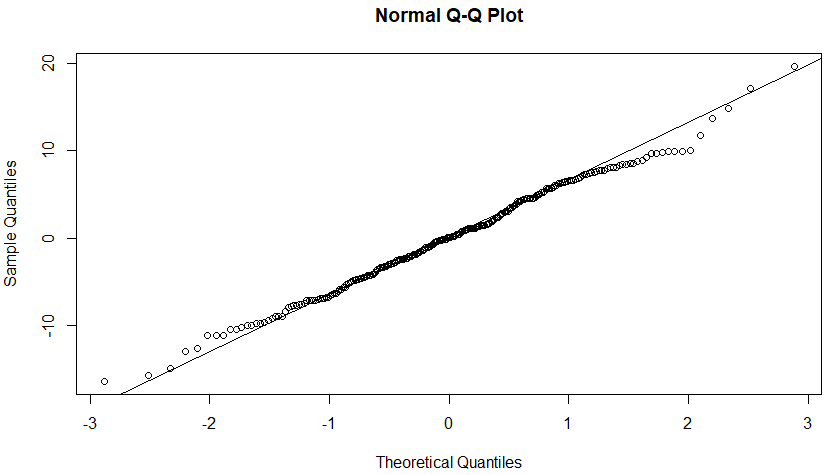
plot(m6$residuals ~ BM$WEIGHT)

abline(h = 0, lty = 3)



hist(m6$residuals)

qqnorm(m6$residuals)

qqline(m6$residuals)

For linearity, the scatterplot should appear linear, and the residual plot should have no apparent patterns. This is true for both cases, as the scatterplot has a positive and moderately strong linear relationship, and the residual plot has no pattern.

For nearly normal residuals, the histogram appears normally distributed, and the qq plot also appears nearly normal. Based on these two cases being checked off, the relationship of residuals is nearly normal.

For constant variance, we look at the residual plot and make sure that both sides of the y=0 line have residuals that are consistent in standard error and distance from the line. The variance of the predicted error is very similar on both sides, so we say that constant variability is met. Since all three conditions are met, a linear model would be reliable.

1. **Find the equation of the least-squares line for the model for “WEIGHT” and “BODYFAT.”**

plot\_ss(BM$WEIGHT, BM$BODYFAT)

Call:

lm(formula = y ~ x, data = pts)

Coefficients:

(Intercept) x

-18.5580 0.2177

Sum of squares: 10629.85

Equation of least-squares model = -18.56 + 0.2177 \* WEIGHT

summary(m6)

Call:

lm(formula = BODYFAT ~ WEIGHT, data = BM)

Residuals:

Min 1Q Median 3Q Max

-16.434 -4.315 0.079 4.540 19.681

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -9.99515 2.38906 -4.184 3.97e-05 \*\*\*

WEIGHT 0.16171 0.01318 12.273 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.135 on 250 degrees of freedom

Multiple R-squared: 0.376, Adjusted R-squared: 0.3735

F-statistic: 150.6 on 1 and 250 DF, p-value: < 2.2e-16

Equation of regression model = -10.00 + 0.162 \* WEIGHT

1. **Use your model to predict the body fat of a man that weighs 154.25 lbs.**

which(BM$WEIGHT == 154.25) = 1

BM$BODYFAT[1] = 12.6

-18.56 + (0.2177 \* 154.25) = 15.022 = BODYFAT PREDICTED

1. **Case 1 in the data frame weighs 154.25 and has a body fat of 12.6. What is the residual for this case?**

Residual = Data – Fit = 12.6 – 15.02 = -2.42; The least squares equation overestimated the body fat by 2.35.