Stat 462 HW3 Jiaqi Li

1.

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

$$R^2 = 1 - \frac{RSS}{Total\ SS} = \frac{Total\ SS - RSS}{Total\ SS}$$

$$= \frac{ESS}{Total\ SS}$$

$$= \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

$$\sum_i (\hat{y}_i - \bar{y})^2 = \sum_i (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2$$

$$= \sum_i (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y})^2$$

$$= \hat{\beta}_1^2 \sum_i (x_i - \bar{x})^2$$

$$= \frac{[\sum_i (x_i - \bar{x})(y_i - \bar{y})]^2 \sum_i (x_i - \bar{x})^2}{[\sum_i (x_i - \bar{x})^2]^2}$$

$$= \frac{[\sum_i (x_i - \bar{x})(y_i - \bar{y})]^2}{\sum_i (x_i - \bar{x})^2}$$

$$R^2 = \frac{[\sum_i (x_i - \bar{x})(y_i - \bar{y})]^2}{\sum_i (x_i - \bar{x})^2} \Big/ \sum_i (y_i - \bar{y})^2$$

$$= \frac{[\sum_i (x_i - \bar{x})(y_i - \bar{y})]^2}{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}$$

$$= \left(\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \right)^2 = \left(\frac{\frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_i (x_i - \bar{x})^2 \frac{1}{n} \sum_i (y_i - \bar{y})^2}} \right)^2$$

$$= \left(\frac{Cov(x, y)}{\sqrt{Var(x)\, Var(y)}} \right)^2$$

$$= [corr(x, y)]^2$$

Note: $Cov(x,y) = \frac{1}{n} \sum_i (x_i - E(x))(y_i - E(y)) = \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y})$

2. For $H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$, we have $p = 2$

$$F = \frac{SS_{reg}/(p-1)}{RSS/(n-p)} = \frac{SS_{reg}}{RSS/(n-2)}$$

$$= \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\hat{\sigma}^2}$$

$$= \frac{\sum_i (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2}{\hat{\sigma}^2}$$

$$= \frac{\sum_i (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y})^2}{\hat{\sigma}^2}$$

$$= \frac{\hat{\beta}_1^2 \sum_i (x_i - \bar{x})^2}{\sigma^2}$$

$$= \frac{\hat{\beta}_1^2}{\hat{\sigma}^2 / \sum_i (x_i - \bar{x})^2}$$

$$= \frac{\hat{\beta}_1^2}{se(\hat{\beta}_1)^2} = \left(\frac{\hat{\beta}_1}{se(\hat{\beta}_1)}\right)^2 = t^2$$

Note: $se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_i (x_i - \bar{x})^2}}$ based on $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}\right)$.

Application

A.

```
> Abd.q=quantile(bodyfat$AbdomenC,  probs = c(0.1, 0.25, 0.5, 0.75, 0.9))
> Abd.q
   10%    25%    50%    75%    90%
 79.50  84.55  90.90  99.20 105.70
```

At 0.1, 0.25, 0.5, 0.75, 0.9 quantiles, Abdomen Circumference are 79.50, 84.55, 90.90, 99.20, 105.70.

```
> yhat.q=beta0.Abd+beta1.Abd*Abd.q
> yhat.q
     10%      25%      50%      75%      90%
10.93033 14.08238 18.04586 23.22647 27.28358
```

At Abdomen Circumference equal to 79.50, 84.55, 90.90, 99.20, 105.70, fitted values are 10.93033, 14.08238, 18.04586, 23.22647, 27.28358.

```
> mean(bodyfat$bodyfat.percent)
[1] 19.0498
```

The sample mean of the body fat percentage is 19.0198.

The least estimate of response is 10.93033 and the largest estimate of response is 27.28358. 0.5 quantile of Abdomen Circumference has the best estimate for the body fat percentage due to the sample mean of body fat percentage.

Since the fitted value of 0.9 quantile has the largest distance with the sample mean, so it is the least reliable estimate for body fat percentage.

(|19.0198-27.28358|=8.26378 > |19.0198-10.93033|>|19.0198-14.08238|>|19.0198-23.22647|>|19.0198-18.04586|)

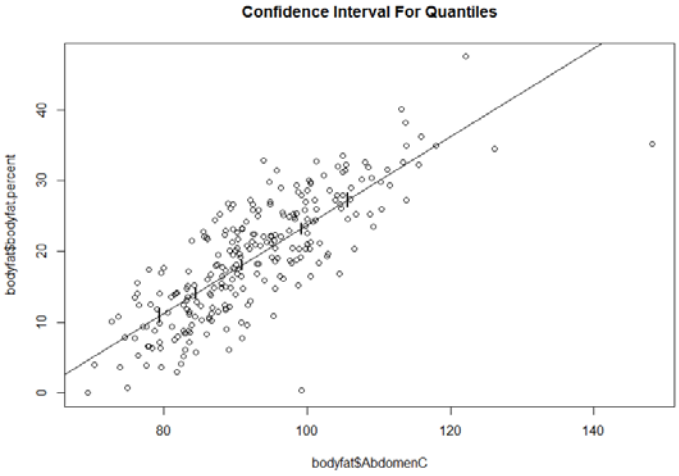
B.

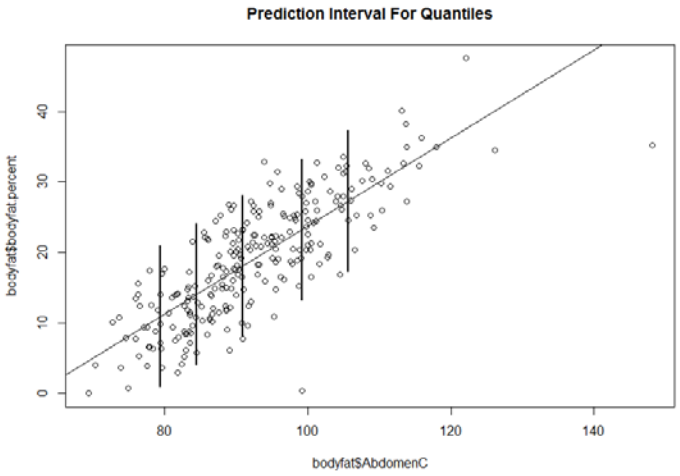
```
> CI
         fit      lwr      upr
10% 10.92993  9.94563 11.91424
25% 14.08197 13.30169 14.86225
50% 18.04542 17.41118 18.67965
75% 23.22599 22.48730 23.96468
90% 27.28306 26.29057 28.27555
```

The 95% confidence intervals for the fitted value of 0.1, 0.25, 0.5, 0.75, 0.9 quantiles of Abdoment Circumference are (9.94563, 11.91424), (13.30169, 14.86225), (17.41118, 18.67965), (22.48730, 23.96468), (26.29057, 28.27555).

```
> PI
          fit        lwr      upr
10% 10.92993  0.9436601 20.91621
25% 14.08197  4.1137361 24.05020
50% 18.04542  8.0875518 28.00328
75% 23.22599 13.2609245 33.19105
90% 27.28306 17.2959762 37.27015
```

The 95% confidence intervals for the fitted value of 0.1, 0.25, 0.5, 0.75, 0.9 quantiles of Abdoment Circumference are (0.9436601, 20.91621), (4.1137361, 24.05020), (8.0875518, 28.00328), (13.2609245, 33.19105), (17.2959762, 37.27015).



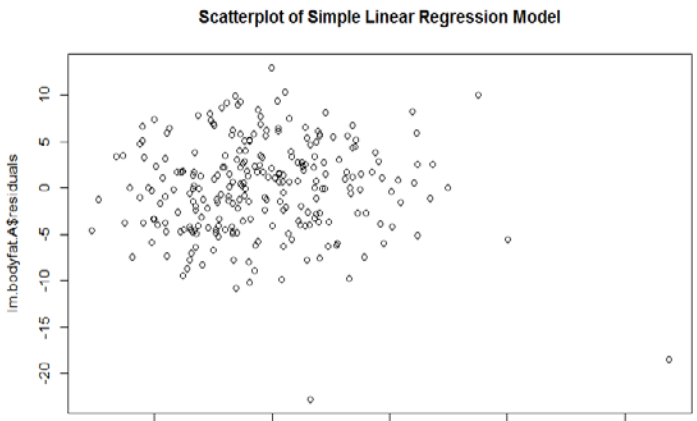


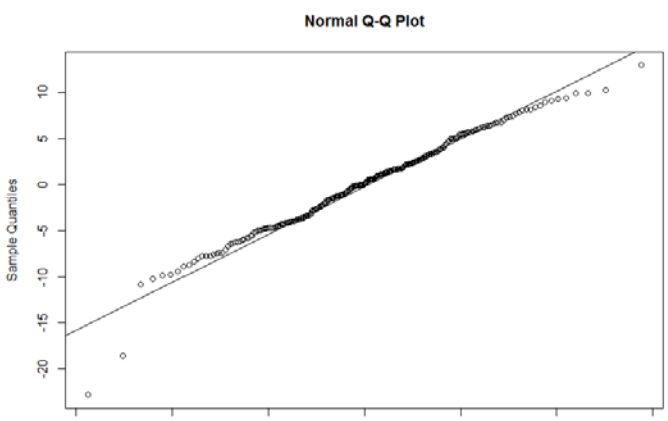
For the <u>confidence interval</u>: we are 95 percent confident that for every 1 cm increase of 0.1, 0.25, 0.5, 0.75, 0.9 quantiles of Abdomen Circumference, there will be increases of 9.94563 to 11.91424, 13.30169 to 14.86225, 17.41118 to 18.67965, 22.48730 to 23.96468, 26.29057 to 28.27555 percentage in the mean of body fat percentage.

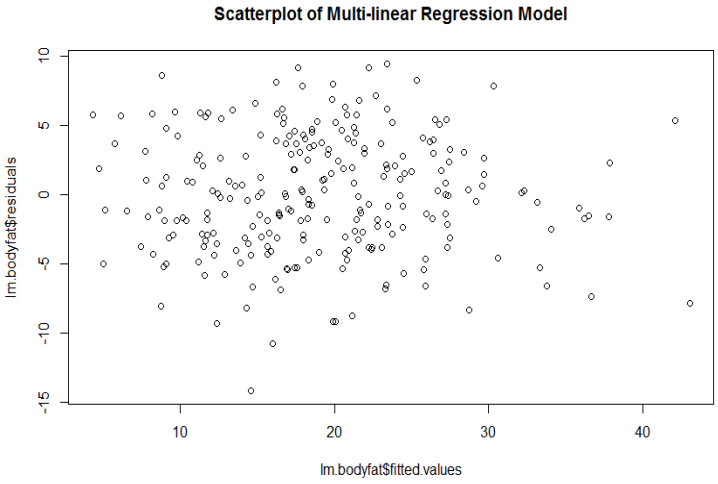
For the <u>prediction interval</u>: we are 95 percent confident that for every 1 cm increase of 0.1, 0.25, 0.5, 0.75, 0.9 quantiles of Abdomen Circumference, there will be increases of 0.9436601 to 20.91621, 4.1137361 to 24.05020, 8.0875518 to 28.00328, 13.2609245 to 33.19105, 17.2959762 to 37.27015 percentage in the mean of body fat percentage.

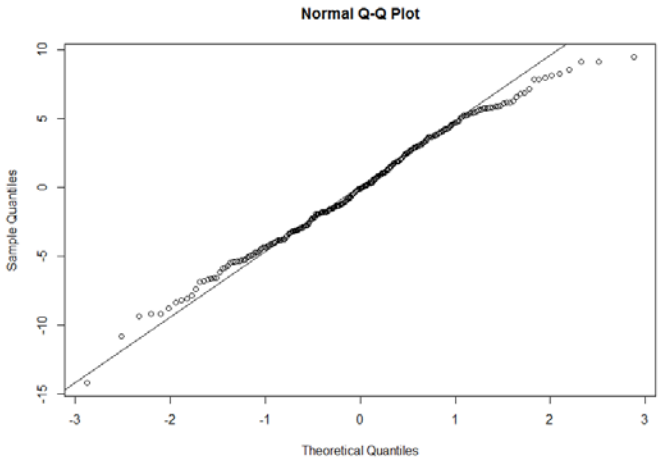
Prediction intervals are larger than the confidence intervals. When we try to calculate the prediction intervals, we are considering more error terms. Thus, prediction intervals have larger variance than confidence intervals.

C.









· Test for Simple Linear Regression Model:

```
        Shapiro-Wilk normality test

data:  lm.bodyfat.A$residuals
W = 0.97997, p-value = 0.001304
```

H_0: Residuals of the simple linear regression model are normally distributed.

H_1: Residuals of the simple linear regression model are <u>NOT</u> normally distributed.

p-value = 0.001304 < α = 0.05

Thus, we reject null hypothesis and conclude that residuals of the simple linear regression model are <u>NOT</u> normally distributed.

By the scatter plot, we see a linear trend, which indicates that residuals of the simple linear regression model satisfies the linearity; however, variance are not constant; by the Q-Q plot, we see that at the left corner, some data do not fit the line well enough, which indicates that the residuals violate the normality.

Thus, the assumptions that errors are normally distributed is <u>NOT</u> adequate.

· Test for Multi-linear Regression Model:

```
        Shapiro-Wilk normality test

data:  lm.bodyfat$residuals
W = 0.9919, p-value = 0.1829
```

H_0: Residuals of the multi-linear regression model are normally distributed.

H_1: Residuals of the multi-linear regression model are <u>NOT</u> normally distributed.

p-value = 0.1829 > α = 0.05

Thus, we fail to reject null hypothesis and conclude that residuals of the simple linear regression model are normally distributed.

By the scatter plot, we see a linear trend, which indicates that residuals of the multi-linear regression model satisfies the linearity; also, variance are generally constant; by the Q-Q plot, we see that all data fit the line pretty good, which indicates that the residuals are normally distributed.

Thus, the assumptions that errors are normally distributed is adequate.