Final Project

Zhengran Ji

6/3/2022

problem 1.a

```
plot(data$Sugar,data$Calories, main = "Suagr vs. Calories")
```

By the plot, I would expect there are positive relation between calories and sugar.

problem 1.b

```
plot(data$Sugar,data$Calories, main = "Suagr vs. Calories")
```

By the scatterplot, I would expect the relation to be linear.

problem 1.c

Calories/serving = beta_0 + beta_1 * suagr/serving

probelm 1.d

```
model1 <- lm(data$Calories~data$Sugar)
summary(model1)
```

```
##
## Call:
## lm(formula = data$Calories ~ data$Sugar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.428  -9.832   0.245   8.909  40.322
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  87.4277     5.1627  16.935   <2e-16 ***
## data$Sugar    2.4808     0.7074   3.507   0.0013 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.27 on 34 degrees of freedom
```

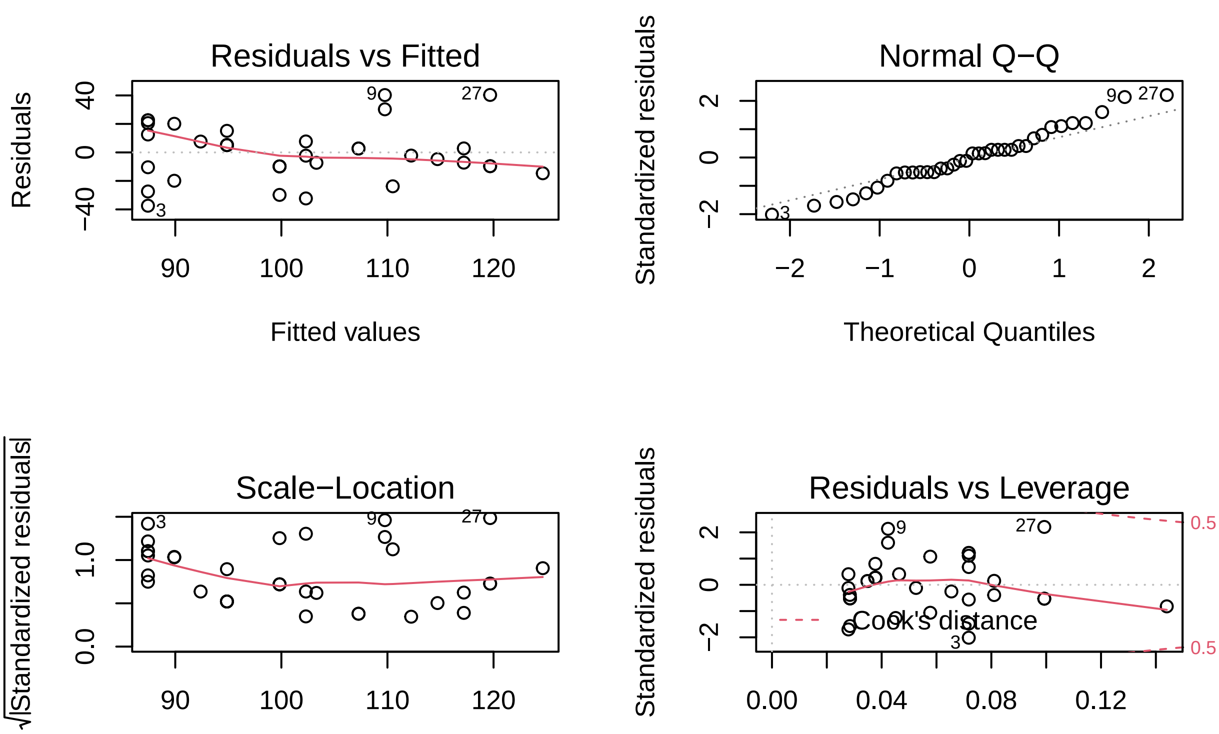
```
## Multiple R-squared:  0.2656, Adjusted R-squared:  0.244
## F-statistic:  12.3 on 1 and 34 DF,  p-value: 0.001296
```

Estimation of intercption is 87.4277 and its standard error is 5.1627, estimation of slope is 2.4808 and its standard error is 0.7074.

probelm 1.e

By the Residuals vs. Fitted plot, we can see that the variance are almost constant. By the qq plot, we can see that the data is fairly normal distributed. And by the Scale-location adn residual vs leverage plot, we can see that the variance are fairly constant and the relation is fairly linear.

```
par(mfrow = c(2,2))
plot(model1)
```



problem 1.f

Interpretation of beta_0_hat is when the cereal has 0 sugar, it is estimated to have 87.4277 Calories per serving. The interpretation of beta_1_hat is as the sugar per serving increase by 1 unit, the cereal's Calories per severing is estimated to increase by 2.4808.

problem 1.g

Method 1: Conduct a hypothesis testing. H0: beta_1_hat = 0, h1: beta_1_hat not equal to 0. The p-value of the slope is smaller than 0.05, so by hypothesis testing we can conclude the the slope is not zero, so there is a significant association between amount of suagr per serving and amount of calories per serving.

Method 2: The confidence interval of the slope is (1.04319, 3.91841) which does not include 0. Therefore, there is a significant association between amount of suagr per serving and amount of calories per serving.

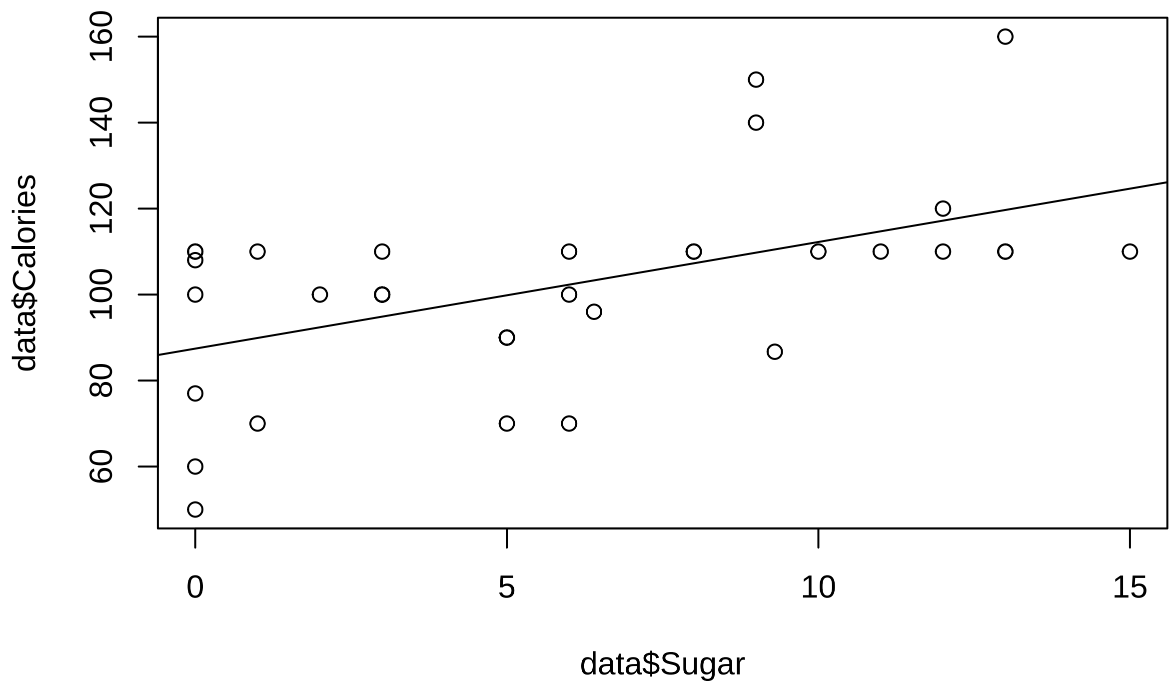
```r
cl <- c(2.4808-qt(0.975,34)*0.7074,2.4808+qt(0.975,34)*0.7074)
cl
```

```
## [1] 1.04319 3.91841
```

problem 1.h

Mueslix Crispy Blend has the largest residual. The interpretation of the residual is that the difference between the prediction calories per serving and the actually value is 40.321746 for Mueslix Crispy Blend cereal.

```r
plot(data$Calories~data$Sugar)
abline(model1)
```




```r
data$residuals <- abs(model1$residuals)  ## Add the residuals to the data.frame
```

problem 1.i

The estimation is 98.5913 calories.

```r
87.4277+4.5*2.4808
```

```
## [1] 98.5913
```

problem 1.j

Calories/serving = beta_0 + beta_1 * suagr/serving + beta_2 * fiber/serving

beta_0_hat = 109.3082, beta_1_hat = 1.0050, and beta_2_hat = -3.7442.

The residual for Mueslix Crispy Blend is 48.8595712 which is bigger. I didn't expect that because as we add one more variable in the model, the model should fit the data better.

```
model2 <- lm(data$Calories~data$Sugar+data$Fiber)
summary(model2)
```

```
##
## Call:
## lm(formula = data$Calories ~ data$Sugar + data$Fiber)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.599  -9.321  -4.435  -0.029  48.860
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 109.3082     6.3913  17.103  < 2e-16 ***
## data$Sugar    1.0050     0.6546   1.535    0.134
## data$Fiber   -3.7442     0.8346  -4.486 8.31e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.42 on 33 degrees of freedom
## Multiple R-squared:  0.5438, Adjusted R-squared:  0.5162
## F-statistic: 19.67 on 2 and 33 DF,  p-value: 2.375e-06
```

```
data$residuals_m <- abs(model2$residuals)
```

problem 2.1

```r
data2 <- readr::read_csv("Simulation_ver2.csv")
```

```
## Rows: 2000 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): x, y
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
sub1 <- data2[1:200,]
sub2 <- data2[201:400,]
sub3 <- data2[401:600,]
sub4 <- data2[601:800,]
sub5 <- data2[801:1000,]
sub6 <- data2[1001:1200,]
sub7 <- data2[1201:1400,]
sub8 <- data2[1401:1600,]
sub9 <- data2[1601:1800,]
sub10 <- data2[1801:2000,]
model1 <- lm(sub1$y~sub1$x)
model2 <- lm(sub2$y~sub2$x)
model3 <- lm(sub3$y~sub3$x)
model4 <- lm(sub4$y~sub4$x)
model5 <- lm(sub5$y~sub5$x)
model6 <- lm(sub6$y~sub6$x)
model7 <- lm(sub7$y~sub7$x)
model8 <- lm(sub8$y~sub8$x)
model9 <- lm(sub9$y~sub9$x)
model10 <- lm(sub10$y~sub10$x)
```

problem 2.2

1. The estimation for beta_0 is 0.29221, and for beta_1 is 2.33153. beta_0 and beta_1 in * are both in the 90% confidence interval constructed by the dataset.

```r
summary(model1)
```

```
##
## Call:
## lm(formula = sub1$y ~ sub1$x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5929 -1.4001  0.0602  1.3840  6.9778
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.29221    0.79078    0.37    0.712
## sub1$x       2.33153    0.05058   46.10   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.224 on 198 degrees of freedom
## Multiple R-squared:  0.9148, Adjusted R-squared:  0.9143
## F-statistic:  2125 on 1 and 198 DF,  p-value: < 2.2e-16
```

```r
confint(model1,level = 0.9)
```

```
##                    5 %    95 %
## (Intercept) -1.014629 1.59905
## sub1$x       2.247944 2.41511
```

2. The estimation for beta_0 is 0.38115, and for beta_1 is 2.33066. beta_0 and beta_1 in * are both in the 90% confidence interval constructed by the dataset.

```r
summary(model2)
```

```
##
## Call:
## lm(formula = sub2$y ~ sub2$x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0265 -1.3993 -0.0325  1.2157  6.0278
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.38115    0.70061   0.544    0.587
## sub2$x       2.33066    0.04518  51.580   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.872 on 198 degrees of freedom
## Multiple R-squared:  0.9307, Adjusted R-squared:  0.9304
## F-statistic:  2661 on 1 and 198 DF,  p-value: < 2.2e-16
```

```r
confint(model2,level = 0.9)
```

```
##                     5 %     95 %
## (Intercept) -0.7766587 1.538967
## sub2$x       2.2559916 2.405336
```

3. The estimation for beta_0 is 1.91428, and for beta_1 is 2.22112. beta_0 and beta_1 in * are both in the 90% confidence interval constructed by the dataset.

```r
summary(model3)
```

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

```
## lm(formula = sub3$y ~ sub3$x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2811 -1.2974 -0.2081  1.2827  8.3905
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.91428    0.75126   2.548   0.0116 *
## sub3$x       2.22112    0.04953  44.846   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.031 on 198 degrees of freedom
## Multiple R-squared:  0.9104, Adjusted R-squared:  0.9099
## F-statistic:  2011 on 1 and 198 DF,  p-value: < 2.2e-16
```

```r
confint(model3,level = 0.9)
```

```
##                   5 %     95 %
## (Intercept) 0.6727506 3.155802
## sub3$x      2.1392693 2.302969
```

4. The estimation for beta_0 is 1.63739, and for beta_1 is 2.24306. beta_0 and beta_1 in * are both in the 90% confidence interval constructed by the dataset.

```r
summary(model4)
```

```
##
## Call:
## lm(formula = sub4$y ~ sub4$x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.185 -1.450  0.020  1.481  6.018
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.63739    0.75137   2.179   0.0305 *
## sub4$x       2.24306    0.04809  46.641   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.128 on 198 degrees of freedom
## Multiple R-squared:  0.9166, Adjusted R-squared:  0.9162
## F-statistic:  2175 on 1 and 198 DF,  p-value: < 2.2e-16
```

```r
confint(model4,level = 0.9)
```

```
##                   5 %     95 %
## (Intercept) 0.3956901 2.879091
## sub4$x      2.1635862 2.322538
```

5. The estimation for beta_0 is -0.89020, and for beta_1 is 2.41614. beta_0 in * is not in the 90% confidence interval constructed by the dataset and beta_1 in * is in the interval.

```
summary(model5)
```

```
##
## Call:
## lm(formula = sub5$y ~ sub5$x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0515 -1.3271 -0.0639  1.2666  7.9730
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.89020    0.85795  -1.038    0.301
## sub5$x       2.41614    0.05595  43.181   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.172 on 198 degrees of freedom
## Multiple R-squared:  0.904,  Adjusted R-squared:  0.9035
## F-statistic:  1865 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
confint(model5,level = 0.9)
```

```
##                   5 %      95 %
## (Intercept) -2.308040 0.5276407
## sub5$x       2.323675 2.5086137
```

6. The estimation for beta_0 is -0.13145, and for beta_1 is 2.3546. beta_0 and beta_1 in * are both in the 90% confidence interval constructed by the dataset.

```
summary(model6)
```

```
##
## Call:
## lm(formula = sub6$y ~ sub6$x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.810 -1.287 -0.026  1.516  4.315
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.13145    0.72360  -0.182    0.856
## sub6$x       2.35460    0.04697  50.133   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.924 on 198 degrees of freedom
## Multiple R-squared:  0.927,  Adjusted R-squared:  0.9266
## F-statistic:  2513 on 1 and 198 DF,  p-value: < 2.2e-16
```

```r
confint(model6,level = 0.9)
```

```
##                   5 %     95 %
## (Intercept) -1.327250 1.064358
## sub6$x       2.276983 2.432217
```

7. The estimation for beta_0 is 2.35366, and for beta_1 is 2.18559. beta_0 and beta_1 in * are both not in the 90% confidence interval constructed by the dataset.

```r
summary(model7)
```

```
##
## Call:
## lm(formula = sub7$y ~ sub7$x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3048 -1.3699 -0.0154  1.4445  5.1923
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.35366    0.86837    2.71  0.00731 **
## sub7$x       2.18559    0.05762   37.93  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.066 on 198 degrees of freedom
## Multiple R-squared:  0.879,  Adjusted R-squared:  0.8784
## F-statistic:  1439 on 1 and 198 DF,  p-value: < 2.2e-16
```

```r
confint(model7,level = 0.9)
```

```
##                   5 %     95 %
## (Intercept) 0.9186059 3.788715
## sub7$x      2.0903722 2.280800
```

8. The estimation for beta_0 is 2.03995, and for beta_1 is 2.20278. beta_0 and beta_1 in * are both not in the 90% confidence interval constructed by the dataset.

```r
summary(model8)
```

```
##
## Call:
## lm(formula = sub8$y ~ sub8$x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9170 -1.2136  0.0094  1.4706  5.8885
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.03995    0.79790   2.557   0.0113 *
## sub8$x       2.20278    0.05241  42.030   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.01 on 198 degrees of freedom
## Multiple R-squared:  0.8992, Adjusted R-squared:  0.8987
## F-statistic:  1766 on 1 and 198 DF,  p-value: < 2.2e-16
```

```r
confint(model8,level = 0.9)
```

```
##                   5 %     95 %
## (Intercept) 0.7213547 3.358554
## sub8$x      2.1161639 2.289388
```

9. The estimation for beta_0 is 1.82781, and for beta_1 is 2.21834 . beta_0 and beta_1 in * are both in the 90% confidence interval constructed by the dataset.

```r
summary(model9)
```

```
##
## Call:
## lm(formula = sub9$y ~ sub9$x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6850 -1.3282  0.1277  1.4268  5.4652
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.82781    0.79087   2.311   0.0219 *
## sub9$x       2.21834    0.05118  43.346   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.112 on 198 degrees of freedom
## Multiple R-squared:  0.9047, Adjusted R-squared:  0.9042
## F-statistic:  1879 on 1 and 198 DF,  p-value: < 2.2e-16
```

```r
confint(model9,level = 0.9)
```

```
##                   5 %     95 %
## (Intercept) 0.5208293 3.134782
## sub9$x      2.1337697 2.302920
```

10. The estimation for beta_0 is 0.91345, and for beta_1 is 2.28313. beta_0 and beta_1 in * are both in the 90% confidence interval constructed by the dataset.

```r
summary(model10)
```

```
##
## Call:
## lm(formula = sub10$y ~ sub10$x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4640 -1.4177  0.2732  1.3321  5.5245
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.91345    0.81682   1.118    0.265
## sub10$x      2.28313    0.05191  43.981   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.128 on 198 degrees of freedom
## Multiple R-squared:  0.9071, Adjusted R-squared:  0.9067
## F-statistic:  1934 on 1 and 198 DF,  p-value: < 2.2e-16
```

```r
confint(model10,level = 0.9)
```

```
##                    5 %     95 %
## (Intercept) -0.4364219 2.263322
## sub10$x      2.1973382 2.368917
```

problem 2.3

The average of beta_0_hat is 1.033821, and the average for beta_1_hat is 2.278695. I was expecting the average for beta_0_hat to be 0.7 and the average for beta_1_hat to be 2.3. The calculation was close for both.

```r
(0.29221+0.381115+1.91428+1.63739-0.8902-0.13145+2.35366+2.03995+1.82781+0.91345)/10
```

```
## [1] 1.033821
```

```r
(2.33153+2.33066+2.22112 +2.24306+2.41614 +2.35460+2.18559+2.20278+2.21834+2.28313)/10
```

```
## [1] 2.278695
```

problem 2.4

70% of the interval contain interception, 80% contain slope. I was expecting 80% of both contain for both. The result is close enough.

problem 3