Homework 1 Stat 587

Jennifer Weichenrieder

2/20/2019

Problem 1

Latent heat of fusion data from two methods, Method A and Method B, are compared.

Part A

What are the means of each treatment? The sample mean $\overline{X}$ is calculated as:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

We have n_a equaling 13 and n_b equal to 8. We find sample means $\overline{X}_a$ and $\overline{X}_b$ for the two methods.

```
## [1] "Sample mean of method A: 80.0207692307692"
```

```
## [1] "Sample mean of method B: 79.97875"
```

Sample variances for each treatment, s_a and s_b, are also computed:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

```
## [1] "Sample variance of method A: 0.00057435897435901"
```

```
## [1] "Sample variance of method B: 0.000983928571428553"
```

Now assuming normality of this data, we can run a T-test to compare the two means, μ_a and μ_b. Our null hypothesis: are the means of Method A and Method B equal? Let's use a two-sided test with $\alpha/2$ of 0.025.

$$H_0 : \mu_a = \mu_b$$

$$H_1 : \mu_a \neq \mu_b$$

First, run the test without assuming equal variances, which is default in R. This is Welch's t-test, which DOES NOT use pooled variance/standard deviation. This is our test statistic t:

$$t = \frac{\overline{X}_a - \overline{X}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}$$

```
##
##  Welch Two Sample t-test
##
## data:  treata and treatb
## t = 3.2499, df = 12.027, p-value = 0.006939
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.01385526 0.07018320
## sample estimates:
## mean of x mean of y
##  80.02077  79.97875
```

This is a very small P-value of 0.00694, which is below the $\alpha/2$ value. We would reject the null hypothesis of equal means based on this test. Also note the value of μ_{H_0}, a difference in means equaling 0, does not fall within the 95% confidence interval of (0.01385526,0.07018320).

Now, we run a slightly different t-test. This time, we will be using the option to specify equal variance assumption. Our new test statistic will be calculated using a pooled standard deviation, s_p. The formula for t is:

$$t = \frac{\overline{X}_a - \overline{X}_b}{s_p\sqrt{\frac{1}{n_a} + \frac{1}{n_b}}}$$

Our formula for s_p, the pooled sample standard deviation, is:

$$s_p = \sqrt{\frac{(n_a - 1)s_a^2 + (n_b - 1)s_b^2}{n_a + n_b - 2}}$$

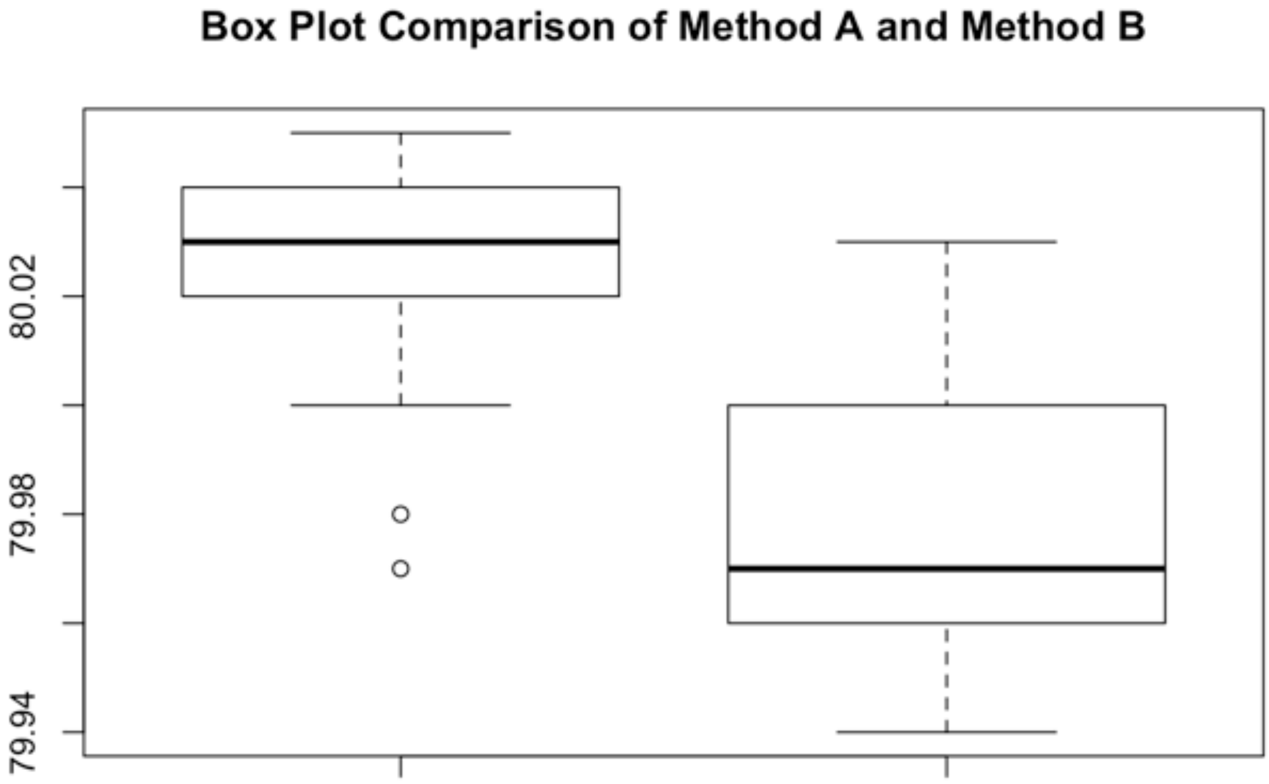
```
##
##  Two Sample t-test
##
## data:  treata and treatb
## t = 3.4722, df = 19, p-value = 0.002551
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.01669058 0.06734788
## sample estimates:
## mean of x mean of y
##  80.02077  79.97875
```

This is an even smaller P-value of 0.002551, so still below the $\alpha/2$ value. Our 95% confidence interval still doesn't contain μ_{H_0}, which equals 0. Though this test gives a different P-value and confidence interval, we would still reject the null hypothesis.

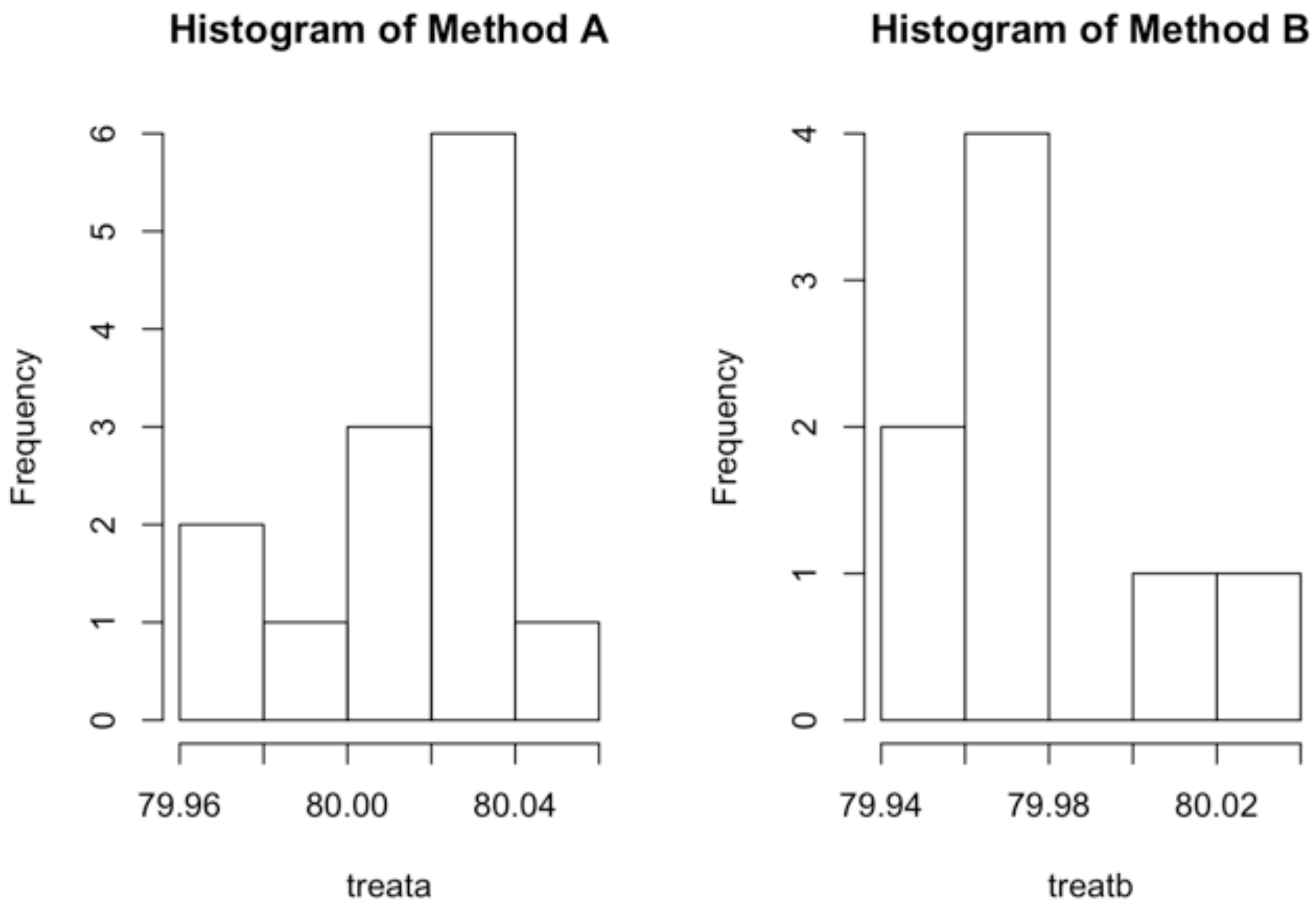
Part B

Here are some plots that illustrate the subtle differences between the two methods for fusion of ice.

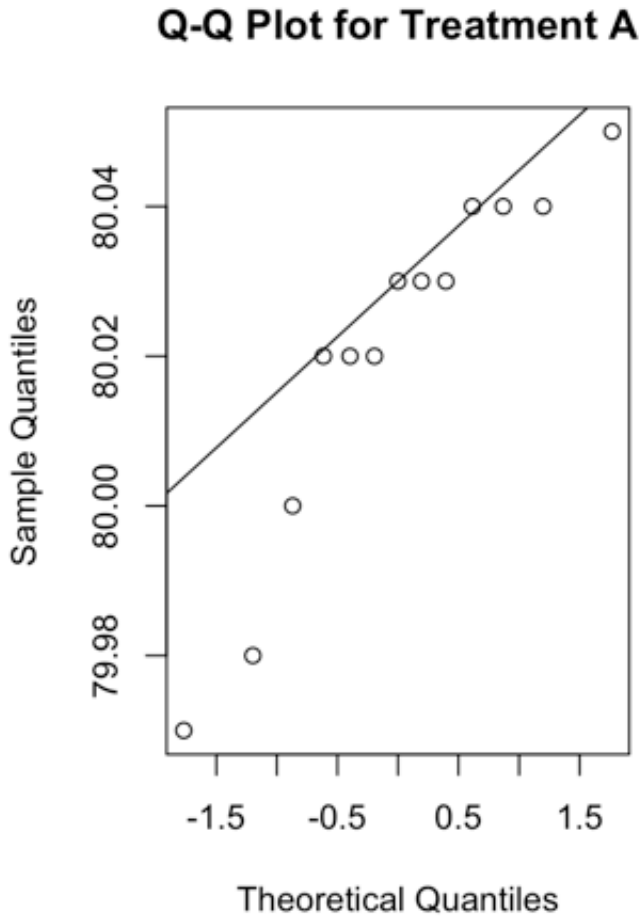
For instance, a box plot:

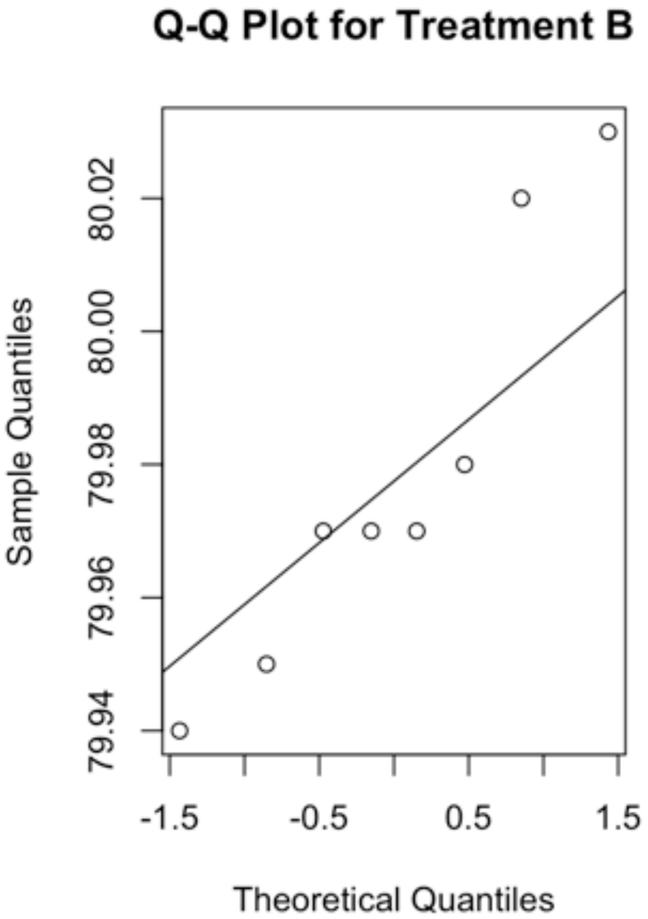


Or a pair of histograms:



Or a Q-Q plot, with a regression line:





It seems there are some differences between the two treatment methods, but they are so subtle that there is no statistical significance. The box-and-whisker plot shows that the mean observation from Method A is about equal to the maximum observation from Method B. Coincidentally, the mean observation from Method B is about equal to the minimum observation from Method A. The histograms also expose the differing skews of each method; Method A appears to skew left, while Method B appears to skew right. The Q-Q plots also offer supporting evidence that perhaps the observations are skewed.

Part C

After pooling the data, we will create a kernel density estimation for the combined observations. The formula for the kernel density estimator is:

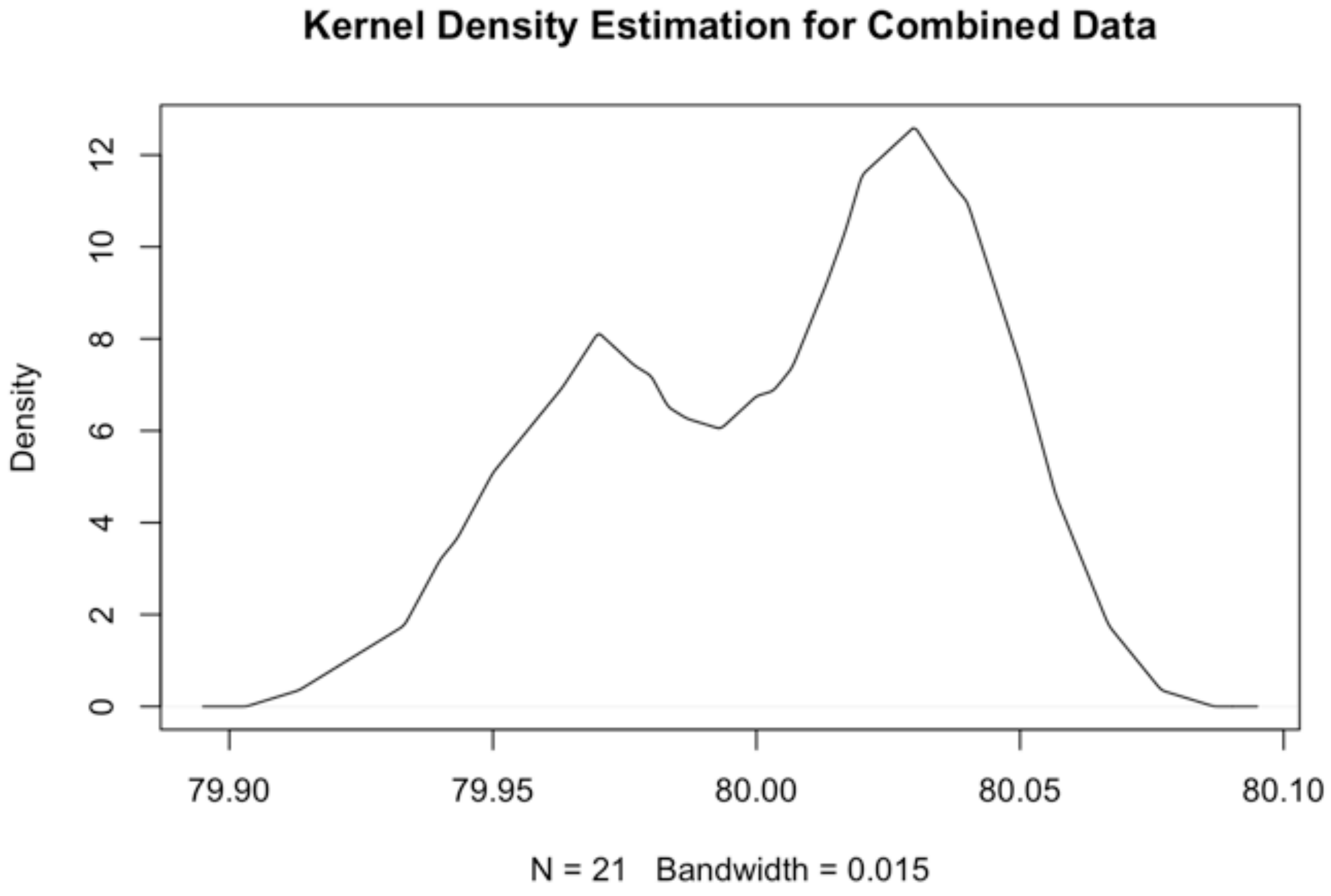
$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h})$$

K is the kernel function. We are using the triangle kernel, which has the formula:

$$K(t) = \begin{cases} (1+t) \text{ if t} \in (-1,0) \\ (1-t) \text{ if t} \in (0,1) \\ 0 \text{ otherwise} \end{cases}$$

We have $n = 21$ observations combined. We will use the triangle kernel and set the width bandwidth h = 0.015.

```
##
## Call:
##  density.default(x = treatpool, bw = 0.015, kernel = "triangular")
##
## Data: treatpool (21 obs.);   Bandwidth 'bw' = 0.015
##
##        x               y
##  Min.   :79.89   Min.   : 0.0000
##  1st Qu.:79.94   1st Qu.: 0.9807
##  Median :80.00   Median : 5.6456
##  Mean   :80.00   Mean   : 4.9954
##  3rd Qu.:80.05   3rd Qu.: 7.5952
##  Max.   :80.09   Max.   :12.5814
```



What is the estimated value given by this density at $x = 80$?

```
## $x
## [1] 80
##
## $y
## [1] 6.743372
```

This model estimates 6.743372 observations at $x = 80$, but since this number is supposed to represent a count, it should be rounded to 7 for a better approximation.

Problem 2

Our data, which has no missing observations, consists of one output, Y, which is the degree of liking a brand, and two indicators, X_1 for moisture content and X_2 for sweetness. Perhaps related to cake?

Part A

We can fit a regression model of the form $Y = X\beta + \epsilon$. First we build a complete model, featuring an interaction term:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_1 x_{2i} + \epsilon_i$$

β_0 is the intercept, β_1 is the amount the favorability changes with moisture content, β_2 is the amount the favorability changes with sweetness level, and β_{12} is a coefficient for the interaction of moisture content and sweetness. We are looking for $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$, and $\hat{\beta}_{12}$. These are our least squared estimators and will minimize the total of squared errors. The sum of squared errors is:

$$\sum_{i=1}^{n}(Y_i - (\beta_0 \ + \beta_1 x_{1_i} + \beta_2 x_{2i} + \beta_{12} x_1 x_{2i}))^2$$

```
## 
## Call:
## lm(formula = y ~ x1 + x2 + (x1 * x2), data = cake)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -4.150 -1.488  0.125  1.700  3.700 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  27.1500     6.4648   4.200  0.00123 ** 
## x1            5.9250     0.8797   6.735 2.09e-05 ***
## x2            7.8750     2.0444   3.852  0.00230 ** 
## x1:x2        -0.5000     0.2782  -1.797  0.09749 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.488 on 12 degrees of freedom
## Multiple R-squared:  0.9622, Adjusted R-squared:  0.9528 
## F-statistic: 101.9 on 3 and 12 DF,  p-value: 8.379e-09
```

It appears that x_1 is the most significant variable in this model, followed by x_2 . $\hat{\beta}_1$ can be interpreted as telling us we can estimate that for every unit the moisture content increases, the favorability of the brand increases by 5.925 points. Using the $\hat{\beta}$ values from the model, the formula can be written:

$$Y_i = 27.15 + 5.925x_{1_i} + 7.875x_{2_i} - 0.5x_1 x_{2\,i} + \epsilon_i$$

Now for a main effects model, removing the interaction term:

$$Y_i = \beta_0 + \beta_1 x_{1_i} + \beta_2 x_{2i} + \epsilon_i$$

Here we are only looking for $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$. The coefficient for the interaction term, $\hat{\beta}_{12}$, is omitted.

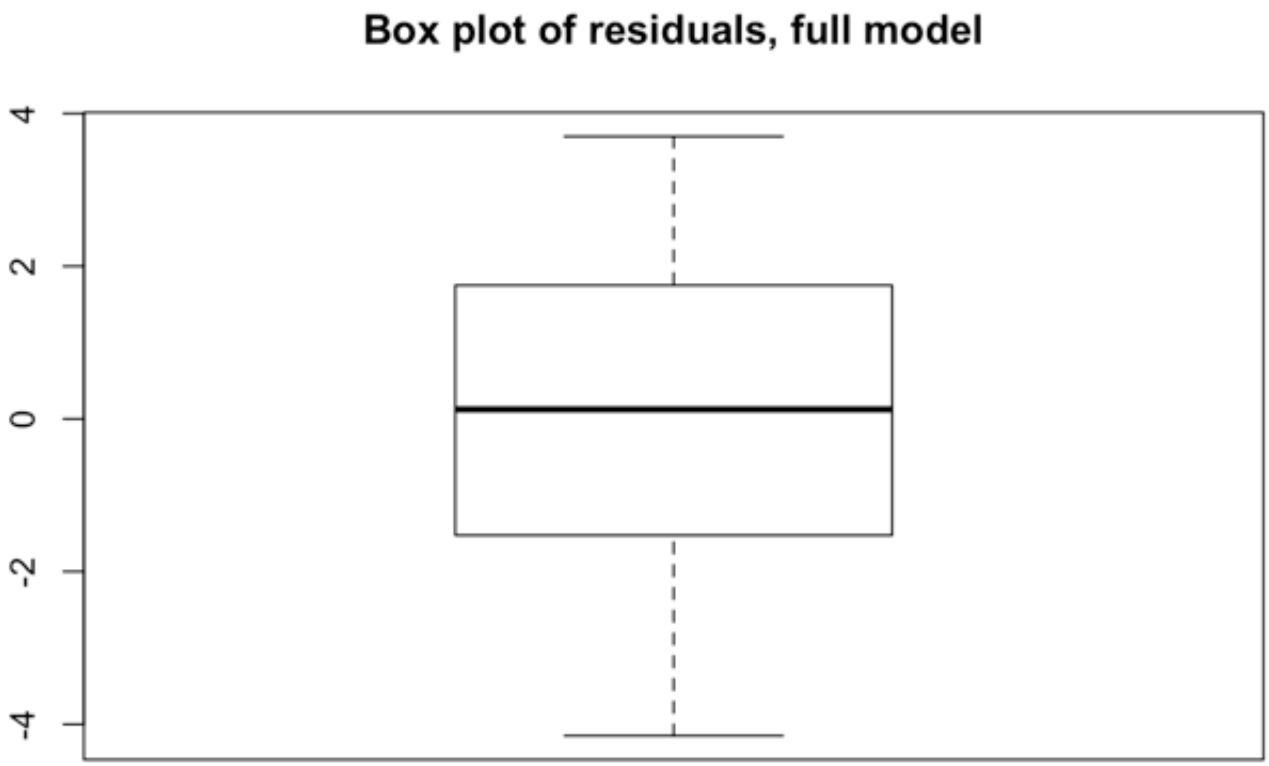
```
##
## Call:
## lm(formula = y ~ x1 + x2, data = cake)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.400 -1.762  0.025  1.587  4.200
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.6500     2.9961  12.566 1.20e-08 ***
## x1            4.4250     0.3011  14.695 1.78e-09 ***
## x2            4.3750     0.6733   6.498 2.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.693 on 13 degrees of freedom
## Multiple R-squared:  0.9521, Adjusted R-squared:  0.9447
## F-statistic: 129.1 on 2 and 13 DF,  p-value: 2.658e-09
```

At $\alpha = 0.05$, both input variables are significant. $\hat{\beta}_1$ can be interpreted as telling us we can estimate that for every unit the moisture content increases, the favorability of the brand increases by 4.425 points. Using the $\hat{\beta}$ values from the model, the formula can be written:

$$Y_i = 37.65 + 4.425 x_{1_i} + 4.375 x_{2_i} + \epsilon_i$$

Part B

What about the residuals? A boxplot of the residuals from the full model, which includes the interaction term:



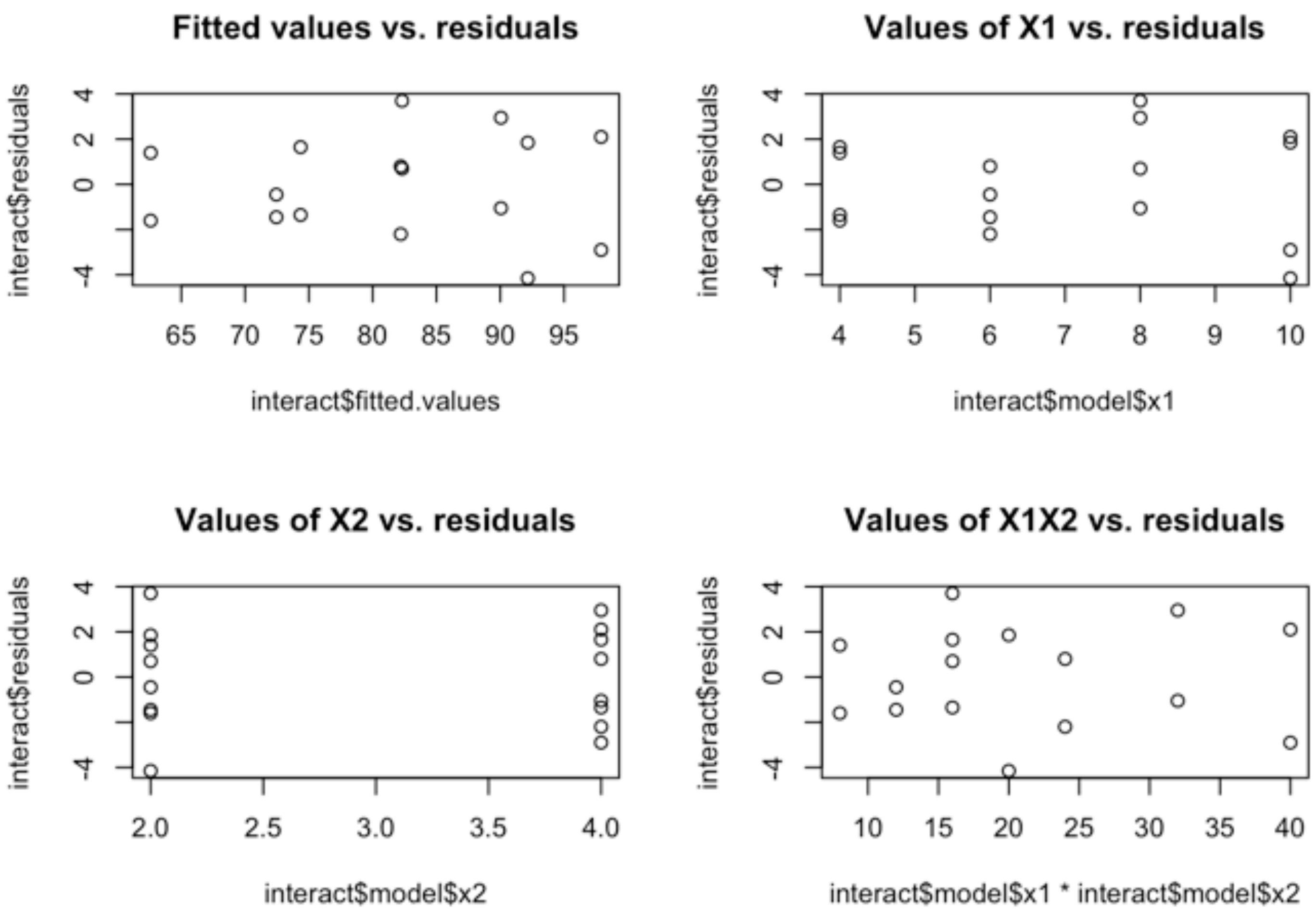
```
## [1] "Mean:  -1.04083408558608e-17"
```

```
## [1] "Standard deviation:  2.22560853101648"
```

The boxplot shows a fairly symmetrical spread of residuals, with a central mean that is close to the median, and quartiles that are also fairly even. None of the residuals appear to be outliers. The mean is very close to 0, and all observations fall within two standard deviations of the mean.

Part C

Plotting the residuals for all terms in the full model against Y, x_1, x_2, and x_{12}.



From these graphs, it looks like the residuals are scattered evenly over values of Y, x_1, x_2, and $x_1 x_2$. Because there are only two distinct values of x_2, this graph is the least useful.

Part D

Lack of fit can be evaluated by ANOVA. We use the F-statistic and examine the residuals. A small F statistic means we don't reject H_0 (the hypothesis that all βs equal 0), but a large F statistic is evidence to reject H_0 (and thus have a regression model with at least one meaningful effect, and thus a non-zero β.)

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x1         1 1566.45 1566.45 215.947 1.778e-09 ***
## x2         1  306.25  306.25  42.219 2.011e-05 ***
## Residuals 13   94.30    7.25
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we see that $\alpha = 0.01$, both x_1 and x_2 are significant. This is evaluating the main effects model. The null hypothesis is that both βs equal zero, and the alternative is that at least one β does not equal zero. The large F-statistic values indicate that both coefficients belong in the model, and that our current model fits well.

Problem 3

Evaluating the hat matrix diagonal values created from the data and regression of the previous question. The $\hat{\beta}$ values found were calculated using linear algebra. In matrix notation,

$$\mathbf{Y} = \mathbf{X}\hat{\beta} + \epsilon$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Recall the hat matrix turns β into $\hat{\beta}$.

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

```
##      1      2      3      4      5      6      7      8      9     10 
## 0.2375 0.2375 0.2375 0.2375 0.1375 0.1375 0.1375 0.1375 0.1375 0.1375 
##     11     12     13     14     15     16 
## 0.1375 0.1375 0.2375 0.2375 0.2375 0.2375
```

Part A

Eight of the diagonal entries equal 0.137, and the other eight equal 0.237. The symmetry is due to linear algebra. $(\mathbf{X}^T\mathbf{X})^{-1}$ is a symmetric matrix. To find the variance of a particular β, each diagonal/pivot value belongs to a β and its entry is multiplied by the variance. The hat matrix is an orthogonal projection matrix.

Part B

The hat matrix helps us detect values in the data which may be influential or outliers. A larger value in $\mathbf{H}$ means a stronger influence. As far as outliers by values of x, there seem to be none in this data set. Every value of x_1 and x_2 in the set is represented multiple times.

Part C

What about point 14? Here x_1 equals 10, x_2 equals 4, but Y equals 95. Is it an outlier? We need DFFITS, DFBETAS, and Cook's Distance values:

```
## [1] "DFFITS for Case 14:  -1.17353122614161"
```

```
## [1] "DFBETAS for Case 14:  2.23606557377049"
## [2] "DFBETAS for Case 14:  -0.216393442622951"
## [3] "DFBETAS for Case 14:  -0.360655737704918"
```

```
## [1] "Cook's Distance for Case 14:  0.363412344731862"
```

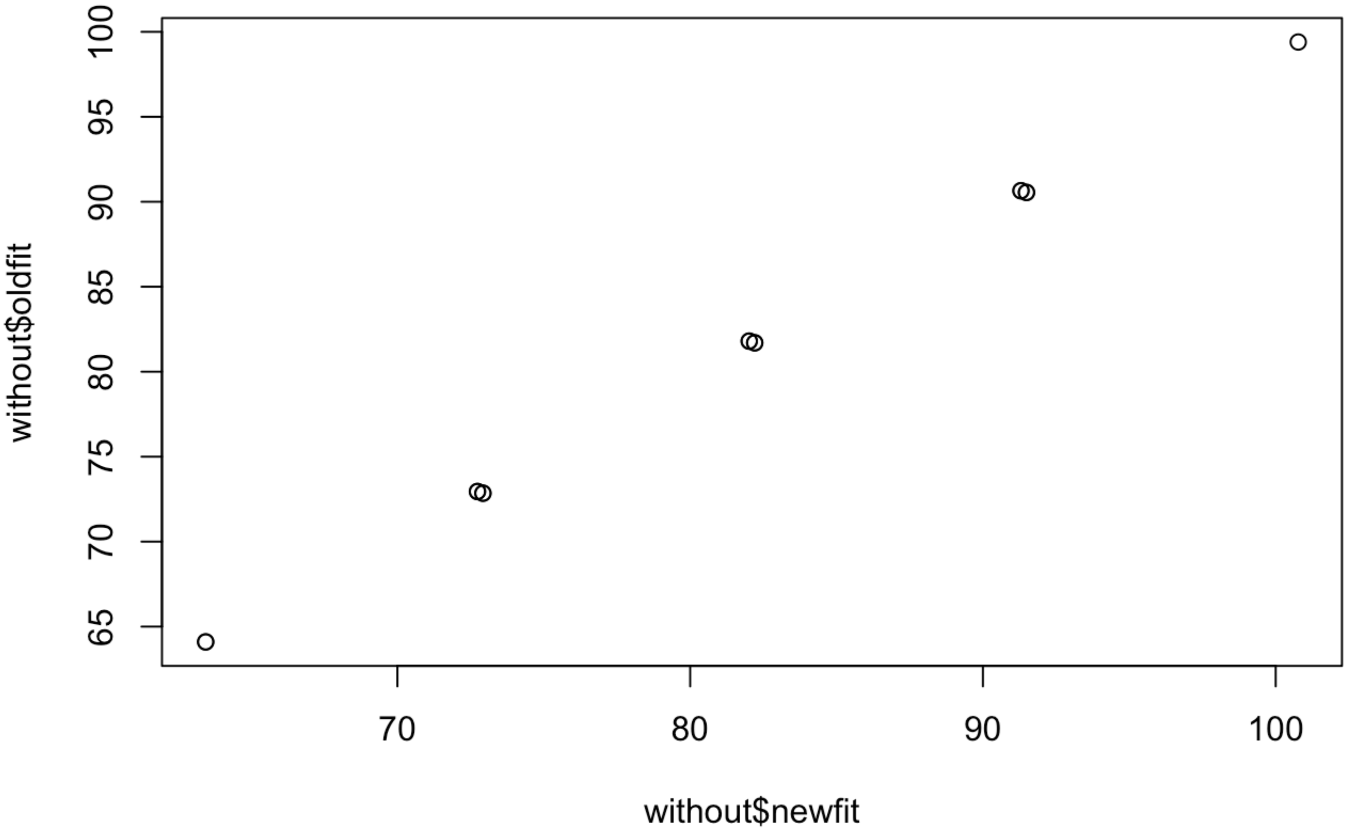
Looking at these measures, the case can be made that observation #14 is an outlier observation of Y. The DFBETAS value for Y is over 2, which indicates an influential point. The DFFITS threshold is given by:

$$2 * \frac{\sqrt{(p+1)}}{(n-p-1)}$$

If the absolute value of the DFFITS number is greater than this, it is an outlier. In this model, p = the number of predictors (including intercept), and n = 16. Since we only have main effects, p = 3, and the DFFITS threshold is 0.3333. The absolute value for the 14th observation's DFFITS number is greater than that, so it is an influential point. As for the Cook's Distance, it may be larger than that of the other points, so it would also support that the point is an outlier.?

Part D

The average absolute percent difference in the fitted values both with and without case 14.



```
## [1] "Average absolute percent difference:  0.63093992641063"
```

The average absolute percent difference in fitted values for the model with and without the 14th observation is less than one percent. While that observation did hold influence, it is not influential enough to drastically change the model when removed.

Part E

Cook's distance, D, for each of the cases in the data.

```
##            1            2            3            4            5 
## 0.0001877130 0.0004223542 0.1803921815 0.1862582123 0.0076655286 
##            6            7            8            9           10 
## 0.0245466787 0.0322971439 0.0143542862 0.0122308711 0.0204060192 
##           11           12           13           14           15 
## 0.1498281704 0.0509831969 0.1318214458 0.3634123447 0.2106609008 
##           16 
## 0.0067576676
```

Observation # 14 is influential as seen earlier, but there may also be some influence from the 15th, 4th, and 3rd observations. To a lesser extent the 11th and 13th observations have some influence as well.