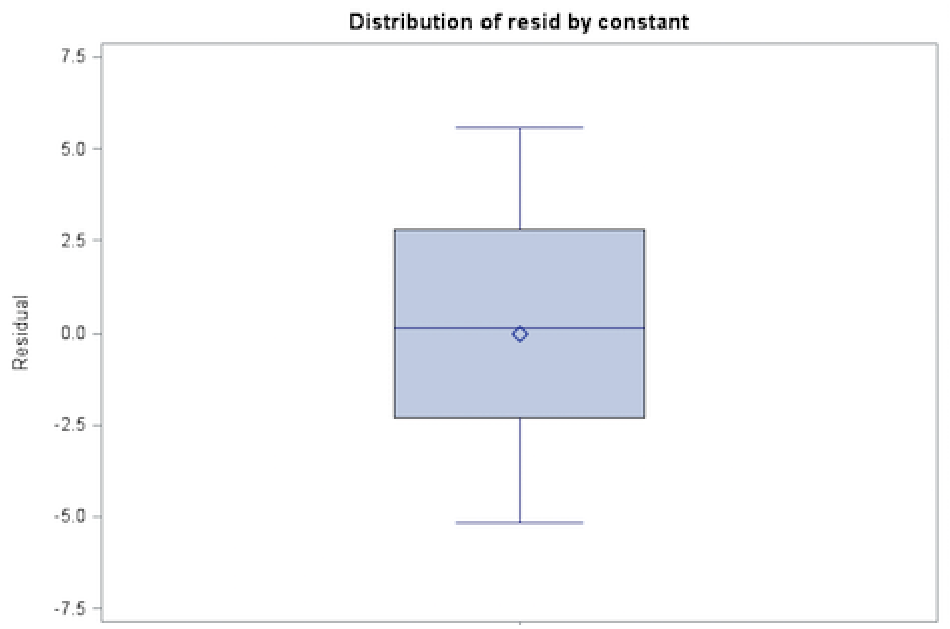
Assignment 4

Ty Darnell

3.6

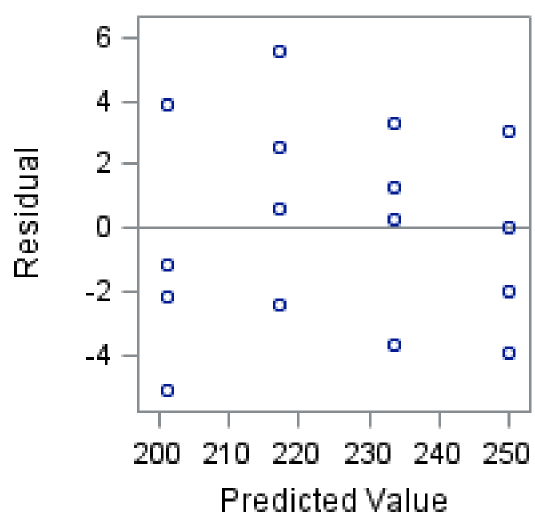
a)

Obs	hardness	hours	pred	resid
1	199	16	201.150	-2.150
2	205	16	201.150	3.850
3	196	16	201.150	-5.150
4	200	16	201.150	-1.150
5	218	24	217.425	0.575
6	220	24	217.425	2.575
7	215	24	217.425	-2.425
8	223	24	217.425	5.575
9	237	32	233.700	3.300
10	234	32	233.700	0.300
11	235	32	233.700	1.300
12	230	32	233.700	-3.700
13	250	40	249.975	0.025
14	248	40	249.975	-1.975
15	253	40	249.975	3.025
16	246	40	249.975	-3.975



The box plot of the residuals looks symmetric since the median is in the middle of the box and is very close to the mean. This suggest normality. The mean is 0 as expected since the mean of the residuals is always 0.

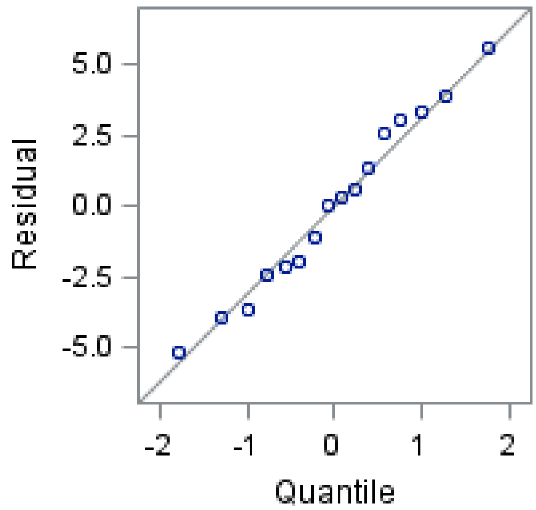
b)



The there is no pattern to the residuals as they appear arranged randomly around the line y=0. This supports the regression model being appropriate.

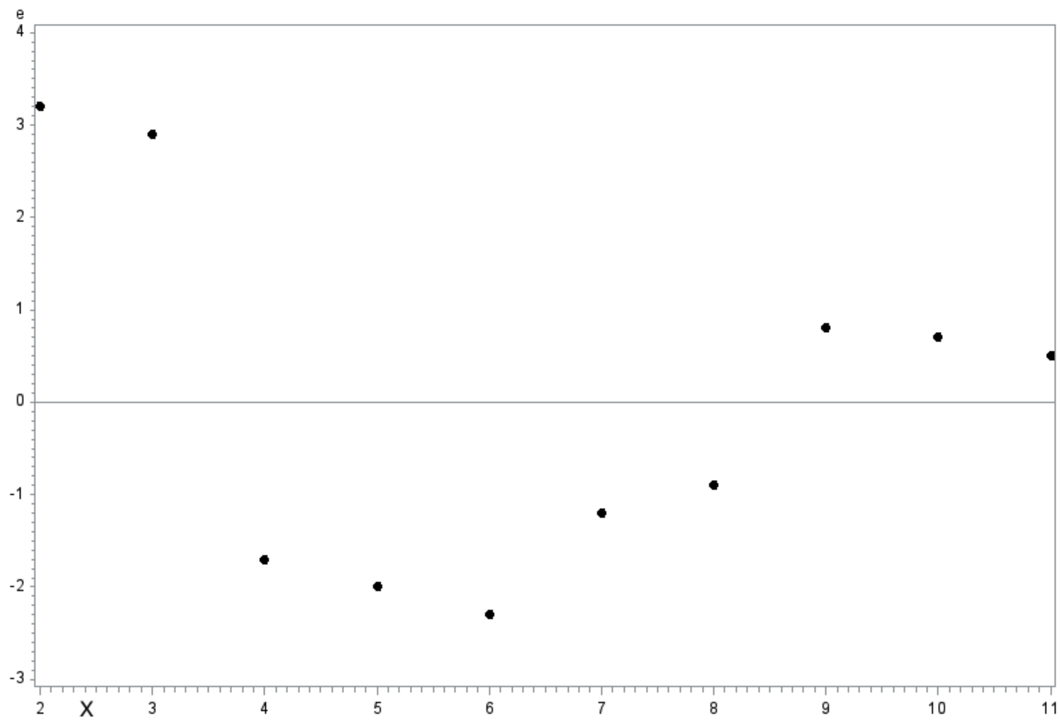
No departures from the regression model are evident.

c)



The residuals appear normally distributed because the QQ plot is approximately a straight line.
The coefficient of correlation between ordered residuals and expected values under normality = .99167
This is greater than the corresponding critical value .941 so this supports the error terms being normally distributed.
Therefore the normality assumption appears to be reasonable.

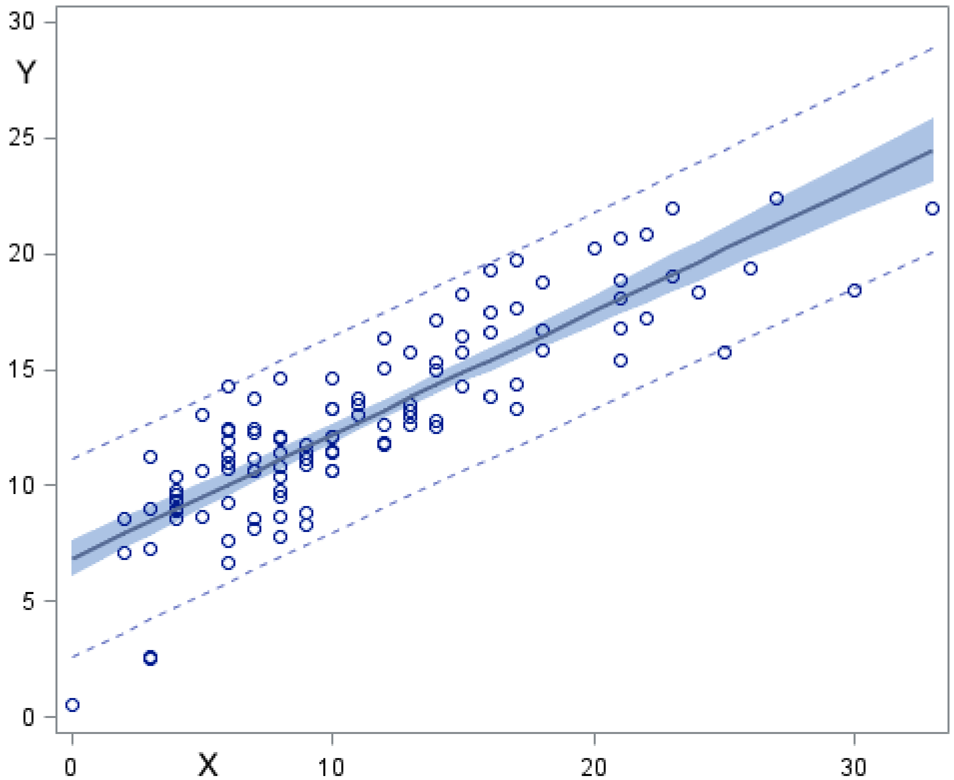
3.9

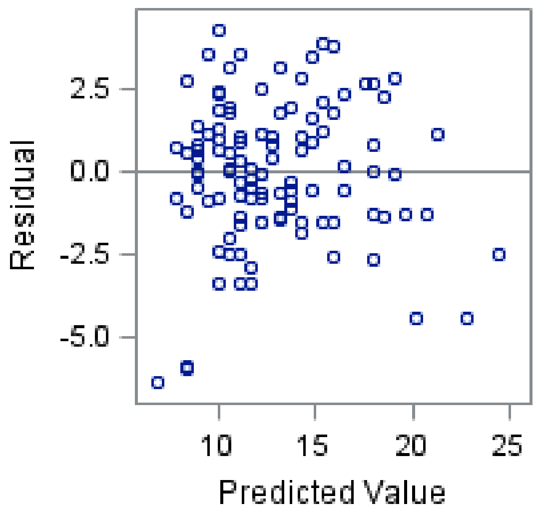


The residuals suggest the relationship is not linear. Also there does not appear to be constant variation in the error terms. A box cox transformation may be appropriate to linearize the relationship since both x and y will need to be transformed.

3.18

a)

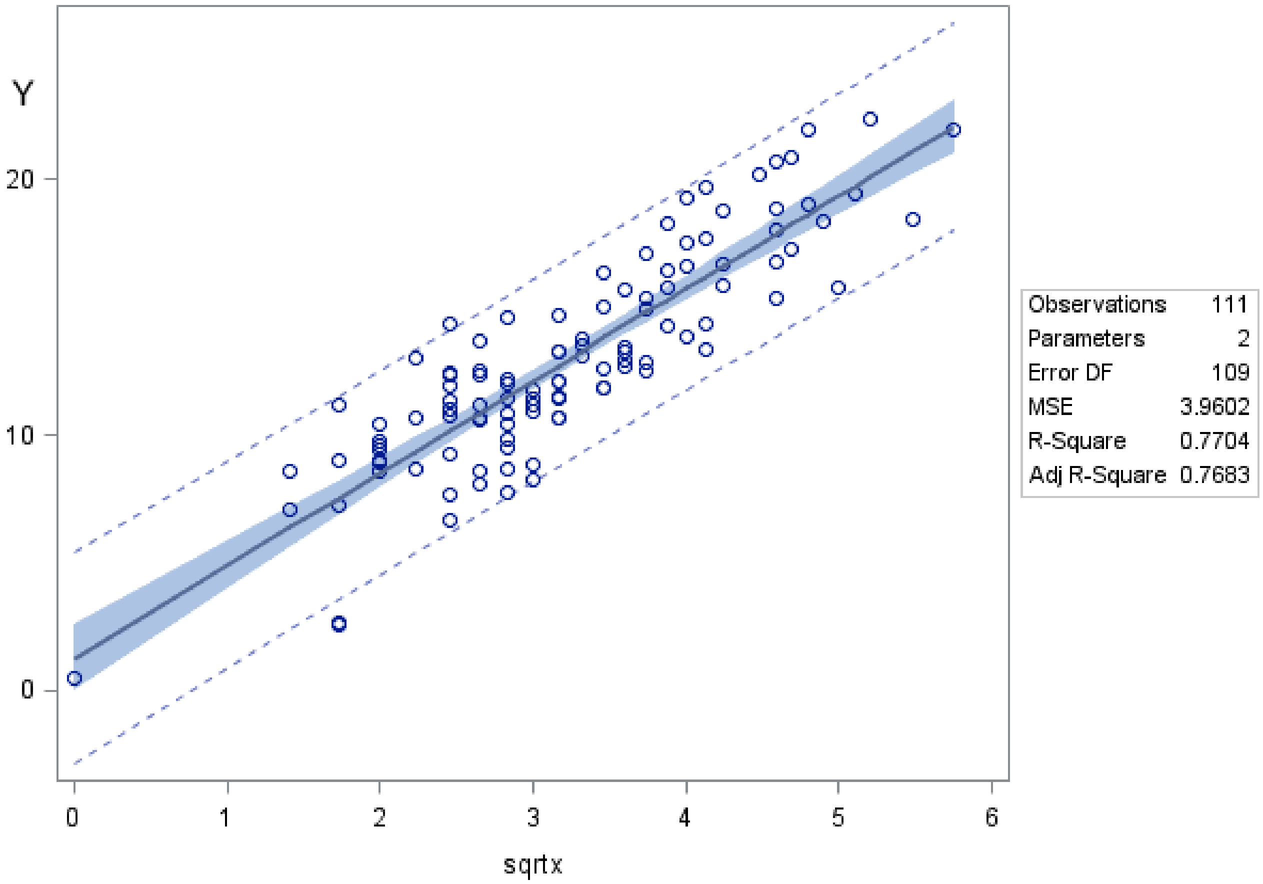




Looking at the scatterplot a linear relation does not appear adequate. A transformation of X would be appropriate to linearize the data, since the error terms appear to have constant variation, looking at the residuals plotted against the predicted values.

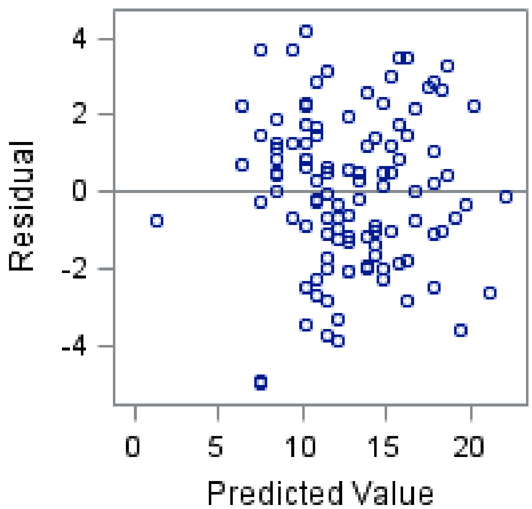
b) $\hat{Y} = 3.62352X' + 1.2547$

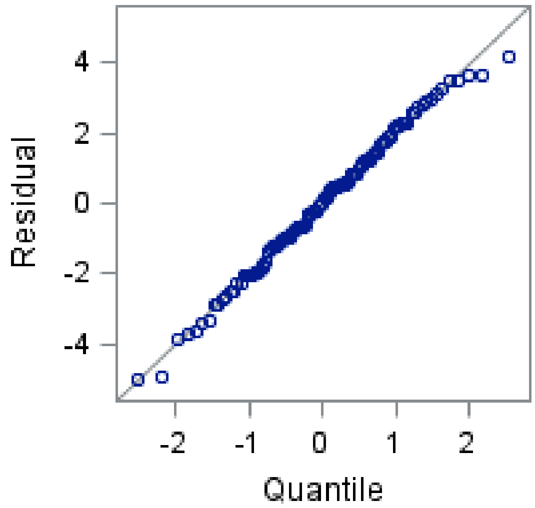
c)



The regression line appears to be a good fit for the transformed data. It is a better fit compared to the original regression line.

d)





The residuals appear randomly arranged around y=0. This suggests a reasonably linear relationship. Also the error terms appear to have constant variance. The normal probability plot approximates a straight line. This suggests that the error distribution is normal.

e) $\hat{Y} = 3.62352\sqrt{X} + 1.2547$

5.2

For all of the matrix problems (5.2,5.5,5.13,Results) see attached work on scratch paper at the end

1) $\begin{bmatrix} 5 & 9 \\ 11 & 11 \\ 10 & 8 \\ 6 & 12 \end{bmatrix}$

2) $\begin{bmatrix} -1 & -7 \\ -5 & -1 \\ 0 & 6 \\ 2 & 4 \end{bmatrix}$

2) $\begin{bmatrix} 58 & 86 \end{bmatrix}$

4) $\begin{bmatrix} 14 & 22 & 11 & 6 \\ 49 & 54 & 20 & 16 \\ 71 & 82 & 32 & 24 \\ 76 & 80 & 28 & 24 \end{bmatrix}$

5) $\begin{bmatrix} 65 & 94 \\ 55 & 77 \end{bmatrix}$

5.5

1) $\begin{bmatrix} 1259 \end{bmatrix}$
$\begin{bmatrix} \sum Y^2 \end{bmatrix}$

2) $\begin{bmatrix} 6 & 17 \\ 17 & 55 \end{bmatrix}$
$\begin{bmatrix} n & \sum X \\ \sum X & \sum X^2 \end{bmatrix}$

3) $\begin{bmatrix} 81 \\ 261 \end{bmatrix}$
$\begin{bmatrix} \sum Y \\ \sum XY \end{bmatrix}$

5.13

$$\begin{bmatrix} 55/41 & -17/41 \\ -17/41 & 6/41 \end{bmatrix}$$

Results from 5.5 and 5.13

1) $\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \begin{bmatrix} .43902 \\ 4.60976 \end{bmatrix}$

2) $\begin{bmatrix} 18.878 \\ 5.0488 \\ 9.6585 \\ 14.268 \\ 14.268 \\ 18.878 \end{bmatrix}$

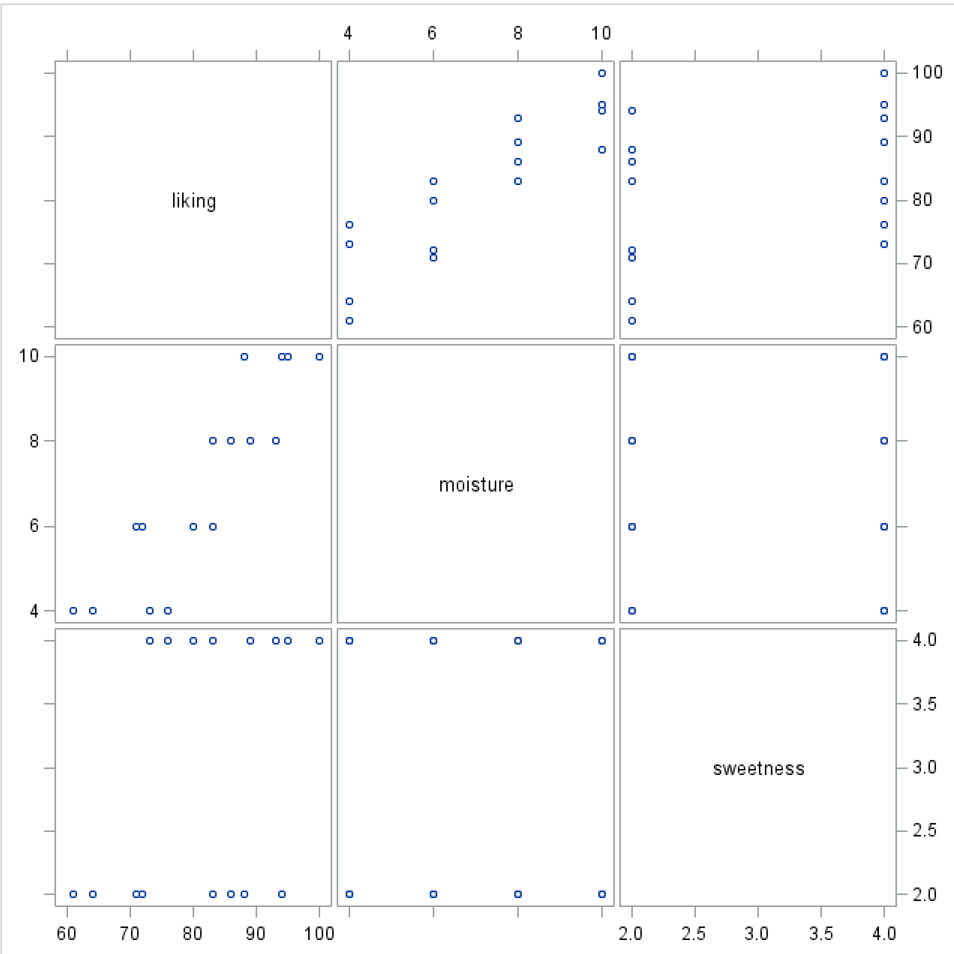
3) $\begin{bmatrix} -.31707 & .17073 \\ .92683 & -.26829 \\ .51220 & -.12195 \\ .09756 & .02439 \\ .09756 & .02439 \\ -.31707 & .17073 \end{bmatrix}$

4) $\begin{bmatrix} -2.878 \\ -.0488 \\ .34146 \\ .7317 \\ -1.268 \\ 3.1219 \end{bmatrix}$

Ty Darnell
Assignment 5

6.5

a)



Pearson Correlation Coefficients, N = 16 Prob > |r| under H0: Rho=0			
	liking	moisture	sweetness
liking	1.00000	0.89239 <.0001	0.39458 0.1304
moisture	0.89239 <.0001	1.00000	0.00000 1.0000
sweetness	0.39458 0.1304	0.00000 1.0000	1.00000

The correlation matrix tells you the correlation coefficient between each of the variables. Liking and moisture have r= .89239. Liking and sweetness have r = .39458. Sweetness and moisture have r=0, this means that there is not collinearity between the two predictor variables. The p-values tell you if the correlation between two variables is significant (if $p < \alpha$)
The scatter plot matrix plots each of the variables against each other. We are most interested in the plots of the response variable (liking) against each of the predictor variables (sweetness and moisture).

b)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1872.70000	936.35000	129.08	<.0001
Error	13	94.30000	7.25385		
Corrected Total	15	1967.00000			

Root MSE	2.69330	R-Square	0.9521
Dependent Mean	81.75000	Adj R-Sq	0.9447
Coeff Var	3.29455		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	37.65000	2.99610	12.57	<.0001
moisture	1	4.42500	0.30112	14.70	<.0001
sweetness	1	4.37500	0.67332	6.50	<.0001

$liking = 37.65 + 4.425 moisture + 4.375 sweetness$

b_1 is estimated as the solution of the ordinary least squares normal equation for the slope of the moisture variable. It is interpreted as the change in the mean response, $E(liking)$, when X_1 (moisture) increases by one unit while all the other independent variables remain constant.

6.6

a)

*see table from 6.5 b
$H_0 : \beta_1 = \beta_2 = 0$
H_1 : Not all $\beta_k = 0, k = 1, 2$
$\alpha = .01$
$F^* = 129.08$
p-value is close to 0

*decision rule: if $p - value < 0$ reject H_0 and conclude H_1
if $p - value \geq 0$ fail to reject H_0 and conclude H_0 (there is no significant statistical relation)*

Since $p - value < .01$ reject H_0 and conclude H_1 (not all coefficients = 0)
So there is a significant statistical relation at the .01 level.
The test implies at least one of β_1 and β_2 does not equal 0.

b)

$p - value = 2.6587$ x 10^{-9} (close to 0)

6.7

a)

*see table from 6.5 b
Coefficient of multiple determination
R^2(MLR)= $SSR/SST0 = 1872.7/1967 = .9521$
Interpretation: Approximately 95% of the variation in Y(liking) is explained by the regression model.

6.8

a)

Predicted Value	Std Error Mean Predict	99% CL Mean	
77.2750	1.1267	73.8811	80.6689

$E\{Y_h\} = \hat{Y}_h = 77.275$
$s\{\hat{Y}_h\} = 1.1267$
$\alpha = .01$
$t(1 - \alpha/2; n - p) = t(.995; 13) = 3.012$
Lower bound = 77.275 - (3.012)(1.1267) = 73.8811
Upper bound = 77.275 + (3.012)(1.1267) = 80.6689
The 99% CI is [73.8811, 80.6689]. Over repeated sampling, 99 out of 100 confidence intervals will contain $E\{Y_h\}$. We are 99% confident that this interval contains $E\{Y_h\}$.

b)

moisture	sweetness	liking	pred	lower	upper	stdi
5	4	.	77.275	68.4808	86.069	2.91946

$E\{Y_h\} = \hat{Y}_h = 77.275$
$s\{\hat{Y}_h\} = 2.91946$
$\alpha = .01$
$t(1 - \alpha/2; n - p) = t(.995; 13) = 3.012$
Lower bound = 77.275 - (3.012)(2.91946) = 68.4808
Upper bound = 77.275 + (3.012)(2.91946) = 86.069
The 99% CI is [68.4808,86.4809]. Over repeated sampling, 99 out of 100 confidence intervals will contain $Y_{h(new)}$. We are 99% confident that this interval contains $Y_{h(new)}$.

6.22

a)

This is not a general linear model, but it can be transformed into one.

b)

It is not a GLM. It can be transformed into a GLM by taking the natural log of both sides resulting in:
$\ln Y = \ln \epsilon + \beta_0 + \beta_1 X_1 + \beta_2 X^2$
let $\ln Y = Y'$ and $\ln \epsilon = \epsilon'$
$Y' = \beta_0 + \beta_1 X_1 + \beta_2 X^2 + \epsilon'$
Y' is a GLM

c)

This is not a general linear model and cannot be transformed into a general linear model.

d)

This is not a general linear model and cannot be transformed into a general linear model.

e)

This is a can be transformed into a general linear model by letting $Y' = 1/Y$ then taking the natural log of both sides.
let $\ln Y' = Y''$
we now have a general linear model.

Ty Darnell
Assignment 6

a)

$\hat{symptoms} = 76.87179 + .09641 hassles + -0.09848 support$

The intercept indicates the predicted value of symptoms when $X_1 = X_2 = 0$ ($hassles = support = 0$)
b_1 indicates the change in the predicted y value when hassles increases by one unit while support remains constant.
b_2 indicates the change in the predicted y value when support increases by one unit while hassles remains constant.

b)

Figure 1:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	89.58494	2.29150	39.09	<.0001
hassles	1	0.08594	0.01921	4.47	<.0001
support	1	0.14636	0.30524	0.48	0.6336
sup_has	1	-0.00506	0.00236	-2.14	0.0368

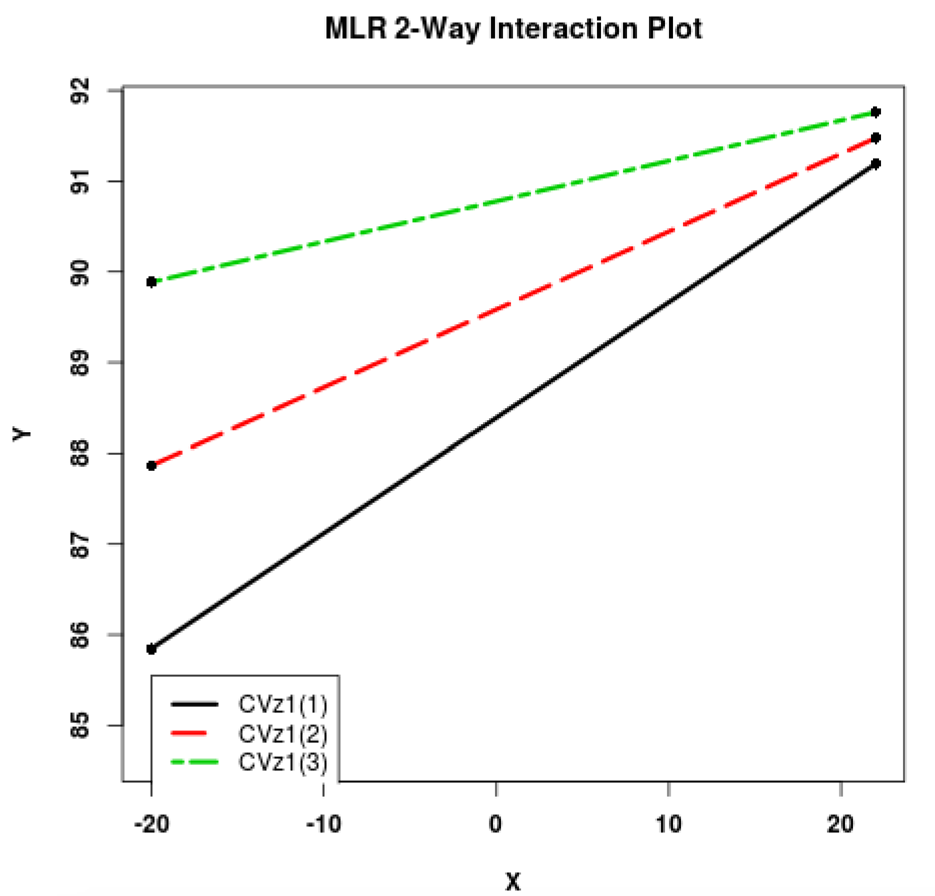
Covariance of Estimates				
Variable	Intercept	hassles	support	sup_has
Intercept	5.2509868268	0.0019227324	-0.044954195	0.0009299469
hassles	0.0019227324	0.000369137	0.0003209676	0.0000115435
support	-0.044954195	0.0003209676	0.0931737691	-0.000269891
sup_has	0.0009299469	0.0000115435	-0.000269891	5.5831066E-6

$\hat{symptoms} = 89.58494 + .08594(hassles) + .14636(support) + -.00506(hassles * support)$

The intercept indicates the predicted value of symptoms when $X_1 = X_2 = 0$ $(hassles = support = 0)$
b_1 indicates the change in the predicted y value when hassles increases by one unit while support is held at 0.
b_2 indicates the change in the predicted y value when support increases by one unit while is hassles is held at 0.
In the interaction model the effect of both hassles and support depends on the level of the other variable.

c)

Figure 2:



1 sd below mean(1):
$\hat{symptoms} = 88.3904 + .1272 hassles$
at mean(2):
$\hat{symptoms} = 89.5849 + .0859 hassles$
1 sd above mean(3):
$\hat{symptoms} = 90.7795 + .0446 hassles$

d)

Figure 3: Output from MLR calculator

```
Region of Significance
==============================================================
 Z at lower bound of region = 6.255
 Z at upper bound of region = 302.0227
 (simple slopes are significant *outside* this region.)

Simple Intercepts and Slopes at Conditional Values of Z
==============================================================
 At Z = cv1...
   simple intercept = 88.3904(3.4917), t=25.3146, p=0
   simple slope     = 0.1272(0.0235), t=5.4126, p=0
 At Z = cv2...
   simple intercept = 89.5849(2.2915), t=39.0944, p=0
   simple slope     = 0.0859(0.0192), t=4.473, p=0
 At Z = cv3...
   simple intercept = 90.7795(3.2748), t=27.7209, p=0
   simple slope     = 0.0446(0.0305), t=1.4642, p=0.1492

Simple Intercepts and Slopes at Region Boundaries
==============================================================
 Lower Bound...
   simple intercept = 90.5004(2.8869), t=31.3489, p=0
   simple slope     = 0.0543(0.0271), t=2.0066, p=0.05
 Upper Bound...
   simple intercept = 133.789(92.0717), t=1.4531, p=0.1522
   simple slope     = -1.4423(0.7188), t=-2.0066, p=0.05
```

*p is close to 0 but not equal to 0

simple slopes and results of significance tests for $\alpha = .05$:

1 sd below mean: simple slope = .1272 the relationship between hassles and symptoms is significant since $p < .05$

at mean: simple slope = .0859 the relationship between hassles and symptoms is significant since $p < .05$

1 sd above mean: simple slope = .0446 the relationship between hassles and symptoms is not significant since $p > .05$

e)

Since the simple slopes decrease when the response function against hassles is considered for higher levels of support, there is an interference interaction effect between the two variables. This is shown in figure 2, the interaction plot. This is also evident from the interaction model shown in part b: $\hat{symptoms} = 89.58494 + .08594(hassles) + .14636(support) + -.00506(hassles * support)$, since the coefficient for the interaction term, b_3 is negative.

f)

*see figure 3

Region of significance for simple slope

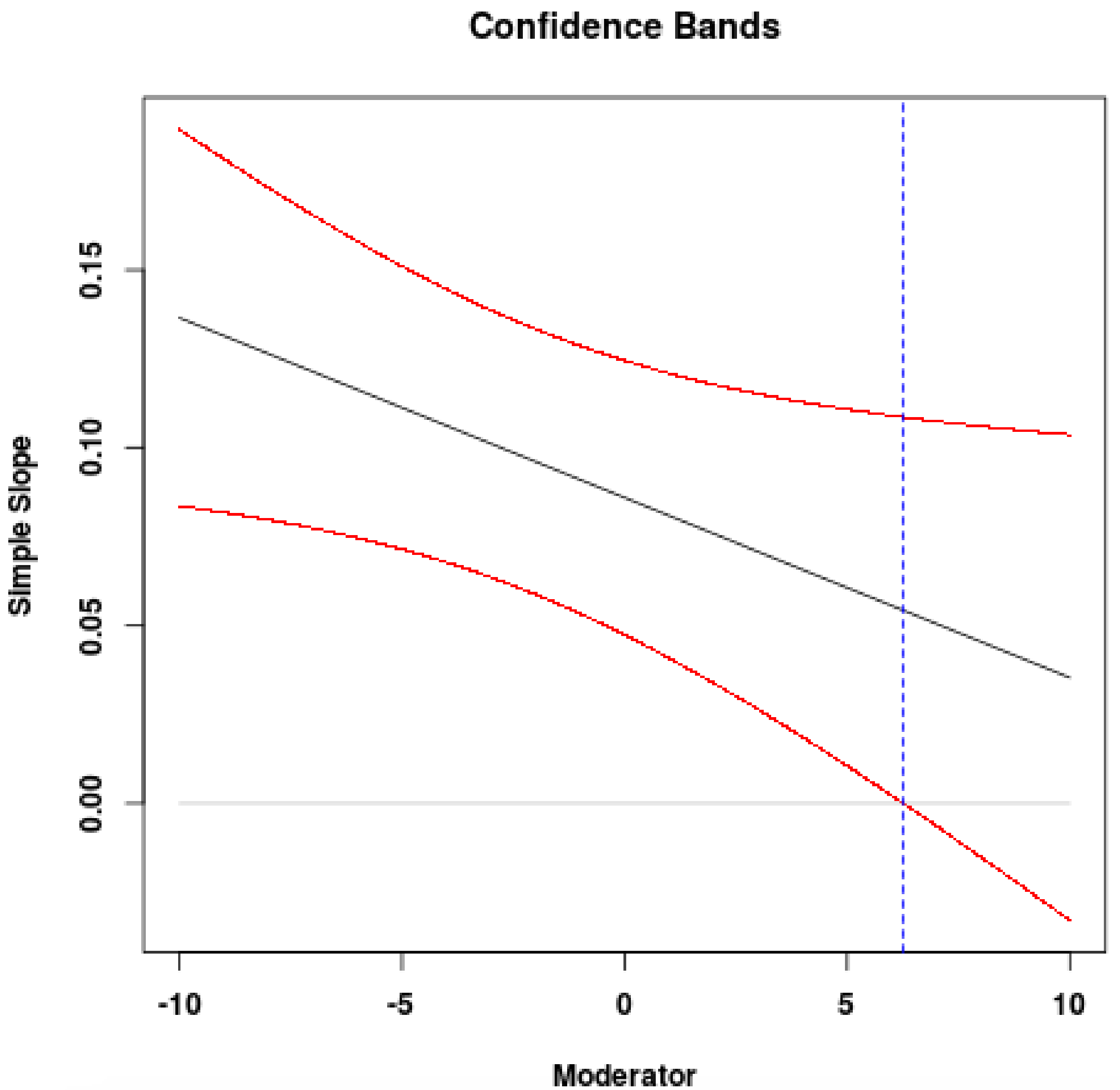
$[6.255, 302.0227]$

simple slopes are significant outside of this region

So when support is less than 6.255 or more than 302.0277, the relationship between hassles and symptoms is significant.

g)

Figure 4:



At values of support to the right of the vertical dotted line, the relationship between hassles and and symptoms is significant.

Ty Darnell
Assignment 7

a)

Figure 1:

```
Dummy Coding Table:
D1 - 1 if Quit Smoking
D2 - 1 if Current Smoker
-------------------
Category    D1   D2
Current     0    1
Quit        1    0
Never       0    0
-------------------
```

Figure 2:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	9.95398	0.40609	24.51	<.0001
age	1	0.10107	0.03900	2.59	0.0160
D1	1	0.56128	0.55990	1.00	0.3261
D2	1	1.93175	0.58934	3.28	0.0032

Age is Mean Centered
D1 = quit
D2 = current

Response function for regression model:
$\hat{runtime} = 9.95398 + .10107 age + .56128 quit + .193175 current$

the parameter estimate for the intercept is the runtime controlling for smoking, and holding age at the mean.

the estimate for the slope of age is the increase in runtime corresponding to a 1 unit increase in age controlling for smoking.

the estimate for the slope of D1 (quit) indicates how much higher runtime is for quitters than never smokers controlling for age.

the estimate for the slope of D2 (current smokers) indicates how much higher runtime is for current smokers than for never smokers controlling for age.

b)

Figure 3: Interaction

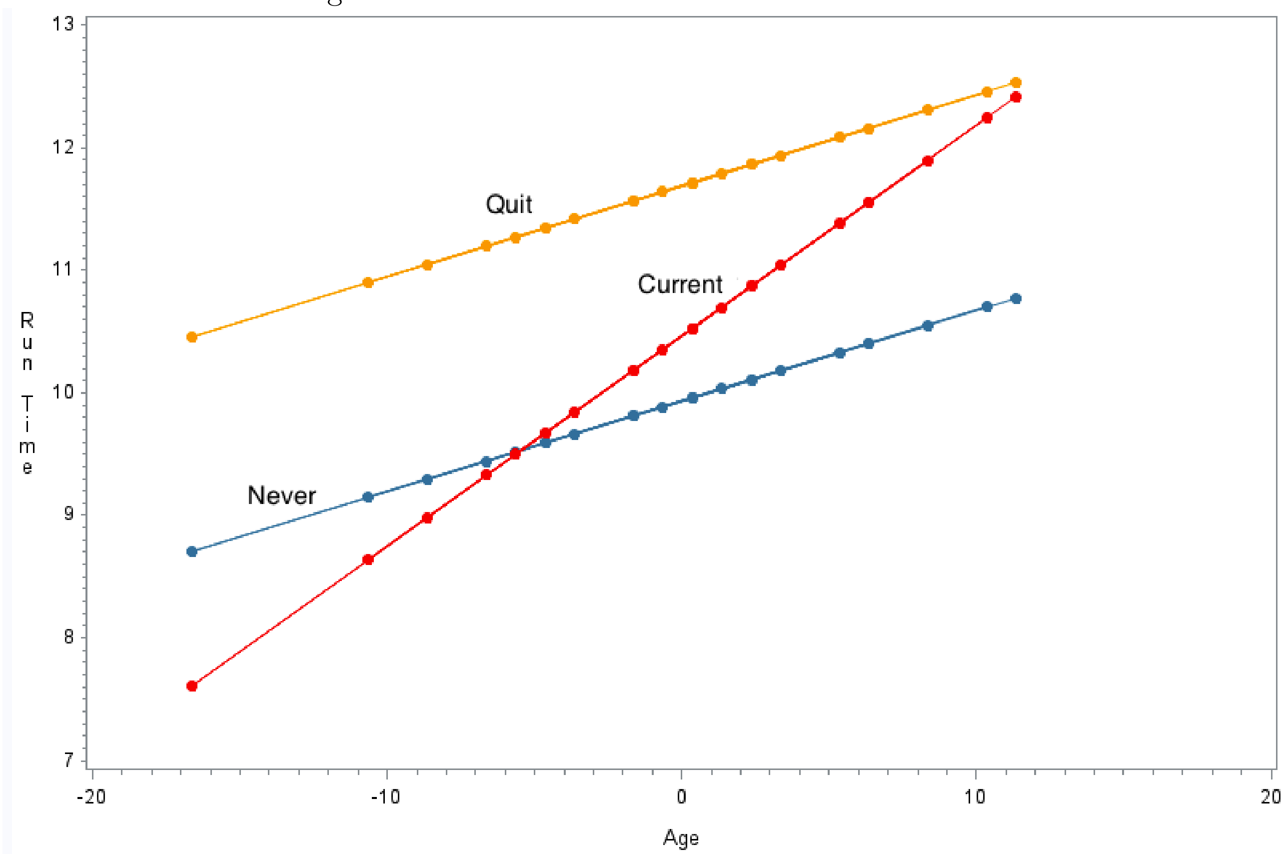
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	9.93363	0.41694	23.82	<.0001
age	1	0.07408	0.08631	0.86	0.4000
D1	1	0.53153	0.58047	0.92	0.3698
D2	1	1.75508	0.62273	2.82	0.0100
D3	1	-0.00019189	0.10262	-0.00	0.9985
D4	1	0.09732	0.11445	0.85	0.4043

D3 = age*D1
D4 = age*D2

Response function for interaction model:
$\hat{runtime} = 9.93363 + .07408age + .53153quit + 1.75508smoker + -.00019(age * quit) + .09732(age * smoker)$

c)

Figure 4: Interaction Probe Plot



There is an interaction effect since the lines have different slopes. The slope for never smokers and quitters is very similar, quitters have a higher intercept. Smokers have the lowest intercept but their run time increases the most rapidly as age increases. It makes sense that over time never smokers would have the shortest run time for 1.5 miles since smoking has been known to decrease cardiovascular performance.

d)

Regression model using restpul as the dependent variable, weight and smoke as the independent variables. Weight is mean centered. See Figure 1 for smoke dummy coding table.

Figure 5:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	57.38275	2.33617	24.56	<.0001
weight	1	-0.53629	0.15928	-3.37	0.0026
D1	1	-4.97079	3.10141	-1.60	0.1221
D2	1	-5.44546	3.55864	-1.53	0.1390

D1 = quit
D2 = current

Response function for regression model:
$\hat{restpul} = 57.38275 + -.53629 weight + -4.97079 quit + -5.44546 current$

Figure 6: Interaction using smoke as the moderator

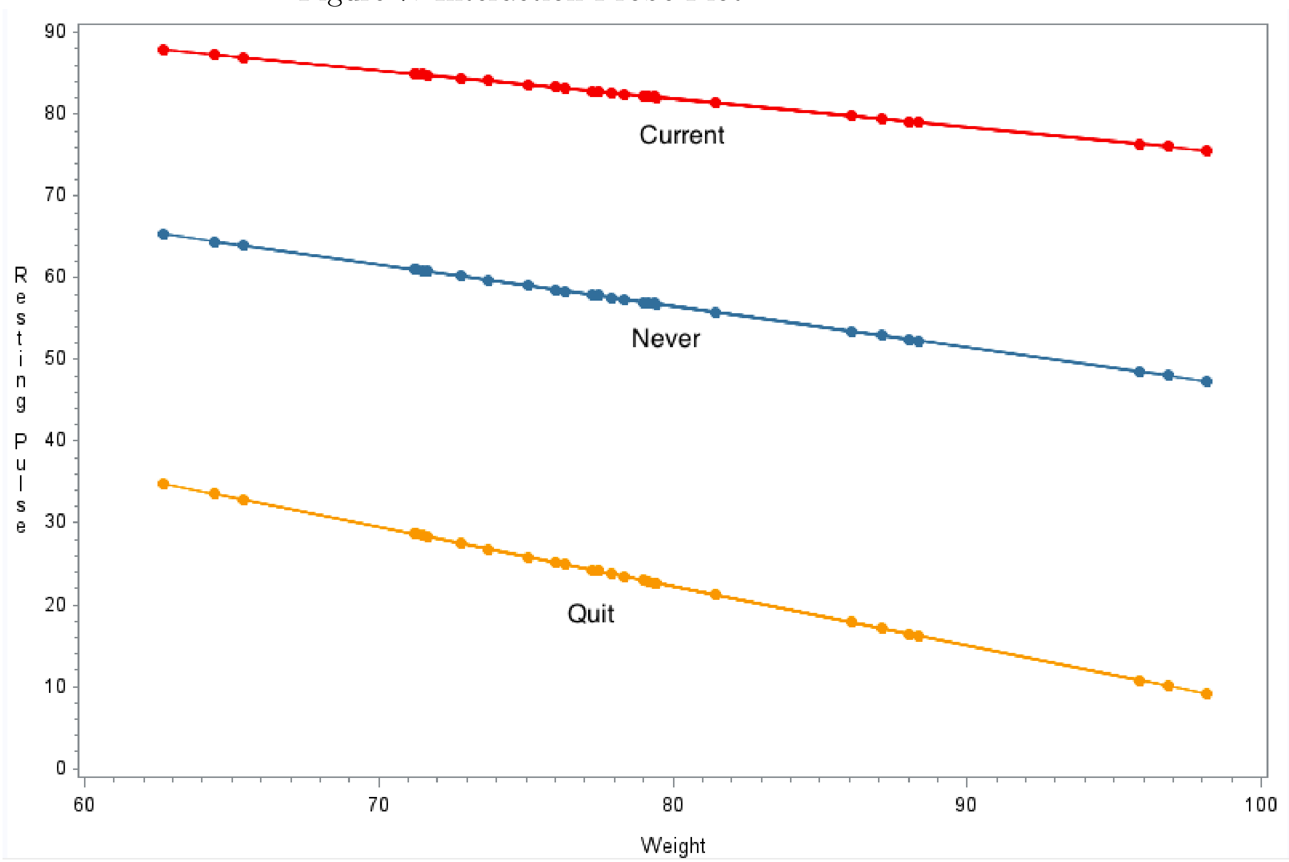
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	96.92977	22.50426	4.31	0.0003
weight	1	-0.50489	0.26819	-1.88	0.0731
D1	1	12.66186	31.57950	0.40	0.6923
D2	1	-16.65909	31.77179	-0.52	0.6053
D3	1	-0.22002	0.38561	-0.57	0.5741
D4	1	0.15887	0.40755	0.39	0.7004

D3 = weight*D1
D4 = weight*D2

Response function for interaction model:

$\hat{restpul} = 96.92977 + -.50489 weight + 12.66186 quit + -16.65909 current + -.22002(weight * quit) + .15887(weight * current)$

Figure 7: Interaction Probe Plot



Current and never smokers appear to have a very similar slope whereas quit is appears slightly different. There does not appear to be much of an interaction effect between smoke and weight, since there does not not appear to be much difference in the slopes of the 3 regression lines.

Ty Darnell
Assignment 8

10.5

a)

Figure 1: Fitted Added Variable Plot for Moisture

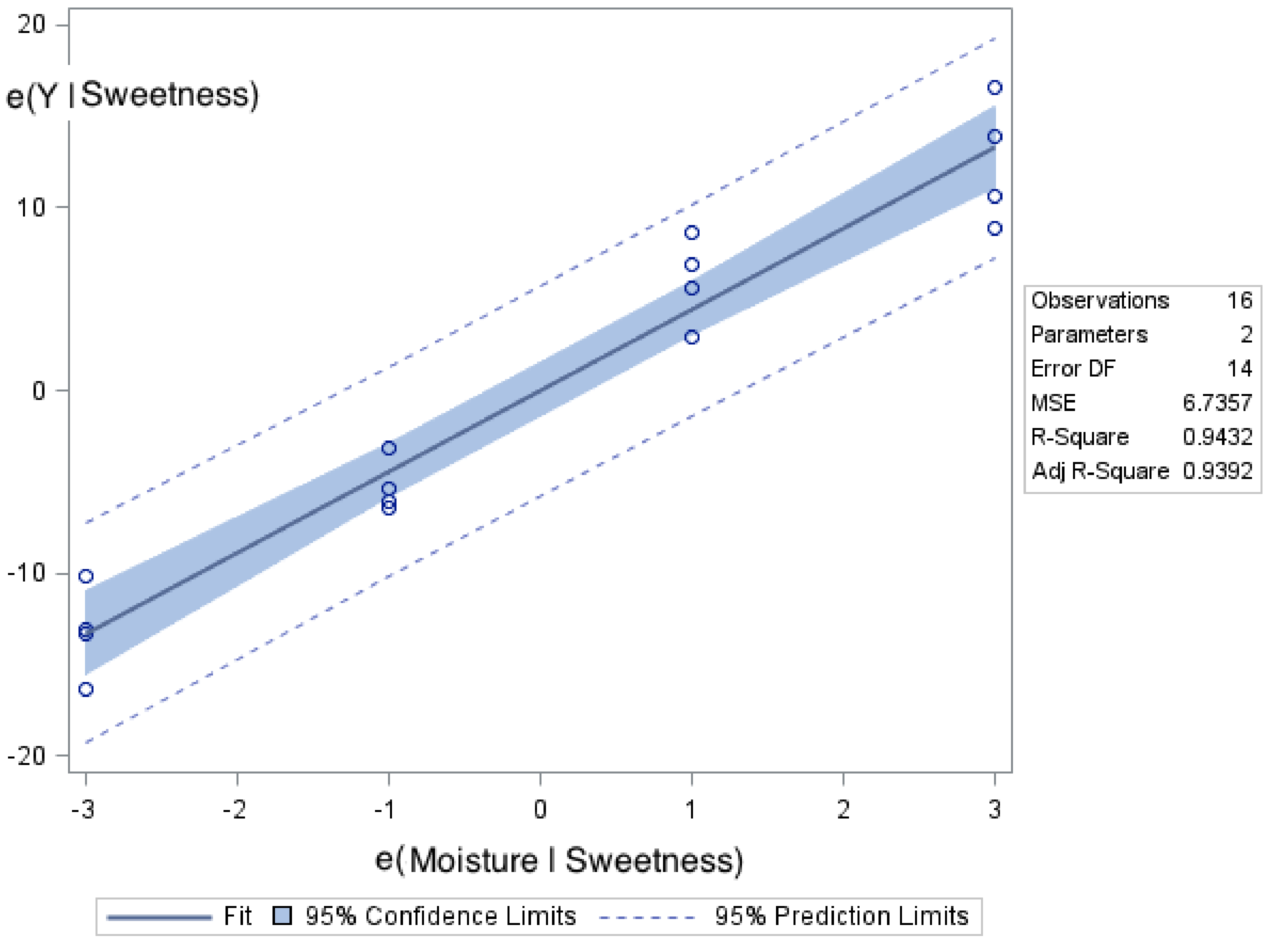
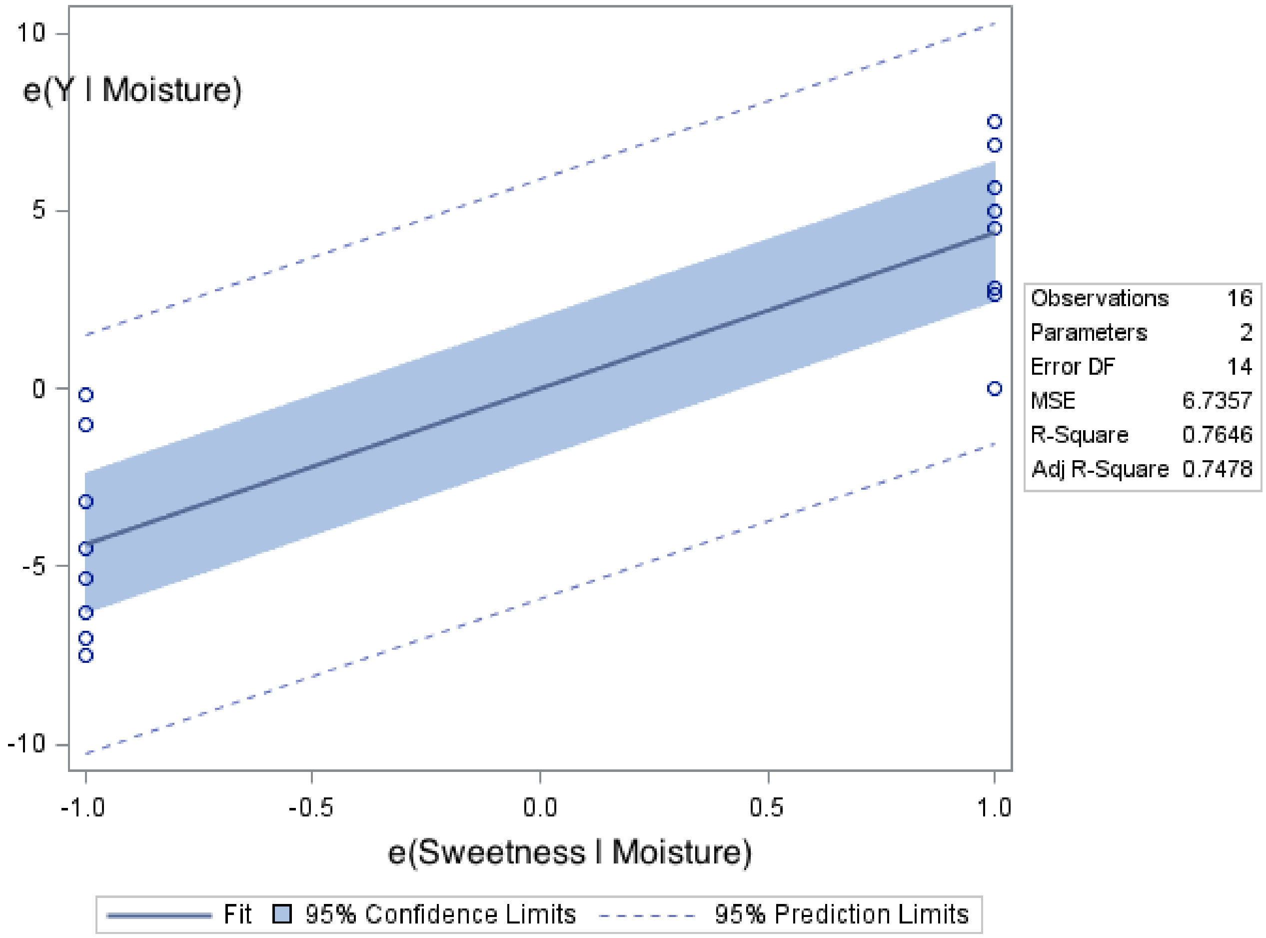


Figure 2: Fitted Added Variable Plot for Sweetness



b)

Regression function from 6.5b:

$$\hat{liking} = 37.65 + 4.425 moisture + 4.375 sweetness \tag{1}$$

Looking at the added residual plot for sweetness (Figure 2), there is a pattern of the residuals occurring only at two extremes. This suggests the regression relationship for sweetness is inappropriate and that sweetness does not add to the explanatory power of the model, when moisture is already included.

Looking at the added residual plot for moisture (Figure 1), the pattern suggests that a linear relationship between y and moisture exists, when sweetness is already present in the model. This suggests moisture should be kept in the model.

leverage)

Figure 3: Leverage Values

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

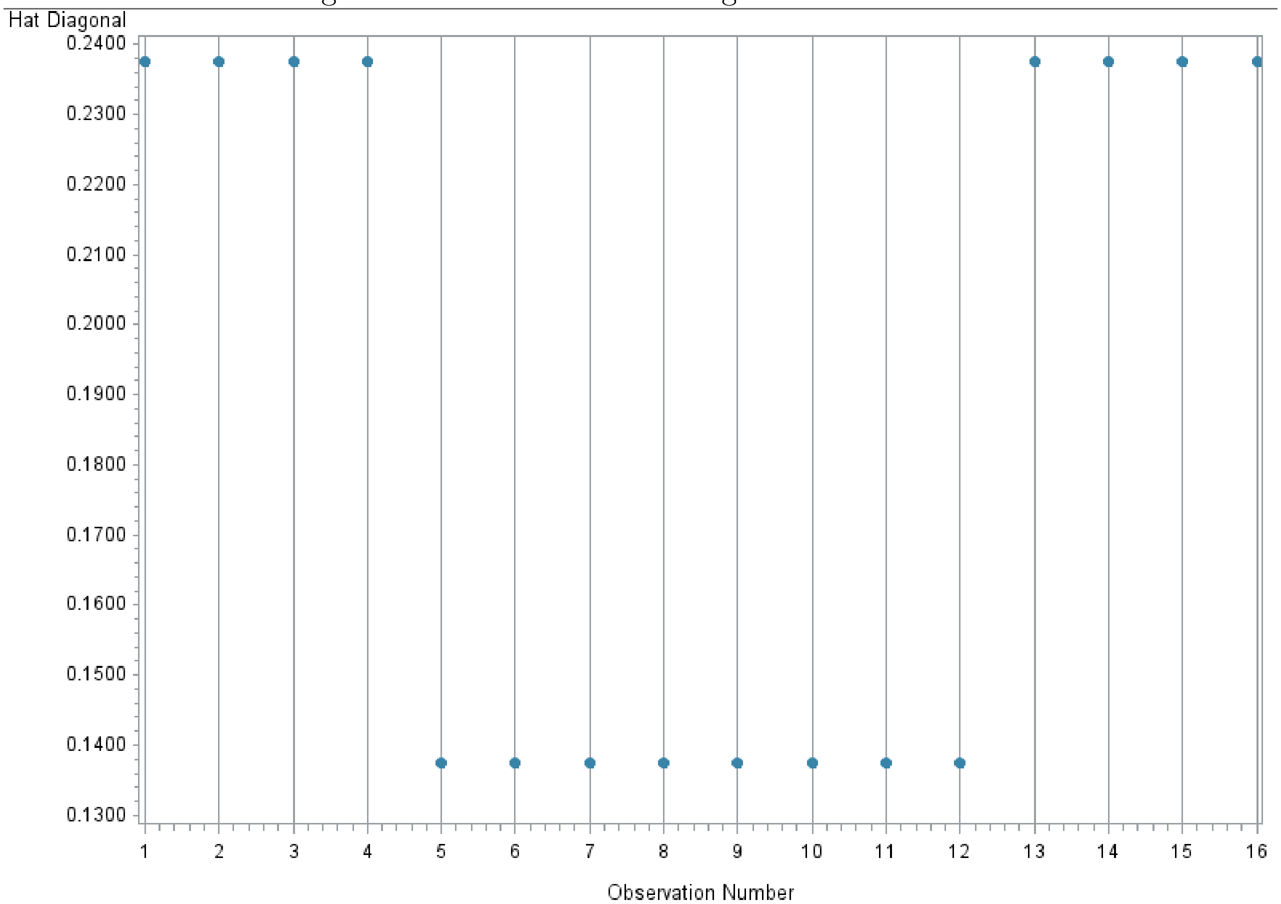
Cutoff value:
$2p/n = 2(3)/16 = .375$
Using .375 for a cutoff value returns no results since no values over .375 exist.

Figure 4: Moderate High Leverage Values

Obs	HatDiagonal
1	0.2375
2	0.2375
3	0.2375
4	0.2375
13	0.2375
14	0.2375
15	0.2375
16	0.2375

Using values between .2 and .5 to find moderate high leverage

Figure 5: Index Plot for Leverage



The index plot shows you the leverage values. Taking $h_{ii} > 0.5$ to indicate very high leverage we see that no very high leverage values exist. Taking $0.2 < h_{ii} < 0.5$ to indicate moderate high leverage we see 8 observations, all at .2375 which are moderate high leverage. This means that half of the observations are moderately high outliers in the X-dimensions.

10.9

a)

Figure 6: Studentized Deleted Residuals

Obs	Residual	RStudent
1	-0.1000	-0.0409
2	0.1500	0.0613
3	-3.1000	-1.3606
4	3.1500	1.3860
5	-0.9500	-0.3669
6	-1.7000	-0.6649
7	-1.9500	-0.7672
8	1.3000	0.5046
9	1.2000	0.4651
10	-1.5500	-0.6044
11	4.2000	1.8230
12	2.4500	0.9778
13	-2.6500	-1.1397
14	-4.4000	-2.1027
15	3.3500	1.4897
16	0.6000	0.2457

Bonferroni-corrected critical value = 3.30778
$\alpha = .1$
$df = 12$
Bonferroni outlier test procedure:
We want to test the largest absolute value of the studentized deleted residuals to see if it is a y-outlier.
largest $|t_i| = |-2.1027|$
Decision Rule:
If $|t_i| > 3.30778$, case i is flagged as a y-outlier.
If $|t_i| \leq 3.30778$, case i is not flagged as a y-outlier.
Conclusion:
$|-2.1027| < 3.30778$
Since the largest absolute value of the studentized deleted residuals is less than the Bonferroni-corrected critical value, there are no cases to flag as y-outliers.

e)

Figure 7: DFFITS and DFBETAS for case 14

Observation	Residual	RStudent	HatDiagonal	CovRatio	DFFITS	DFB_Intercept	DFB_moisture	DFB_sweetness
14	-4.4000	-2.1027	0.2375	0.6507	-1.1735	0.8388	-0.8077	-0.6020

Cook's distance for case 14 = 0.36341
DFFITS:
Since this is a small data set (16 observations) we will use the guideline $|DFFITS| > 1$ to identify influential cases.
Since $|-1.1735| > 1$ observation 14 is an influential case by DFFITS. This suggests observation 14 is influential on its own fitted value.
Cook's Distance:
Using 4/n for the cutoff value for Cook's Distance.
$4/n = 4/16 = .25$
Since $.36341 > .25$ this suggests case 14 is influential on all fitted values.
DFBETAS:
We will use guideline $DFBETAS > 1$ since we have a small data set.
SInce all of the $DFBETAS < 1$ this suggests observation 14 is not influential

on the regression coefficients.

Conclusion:

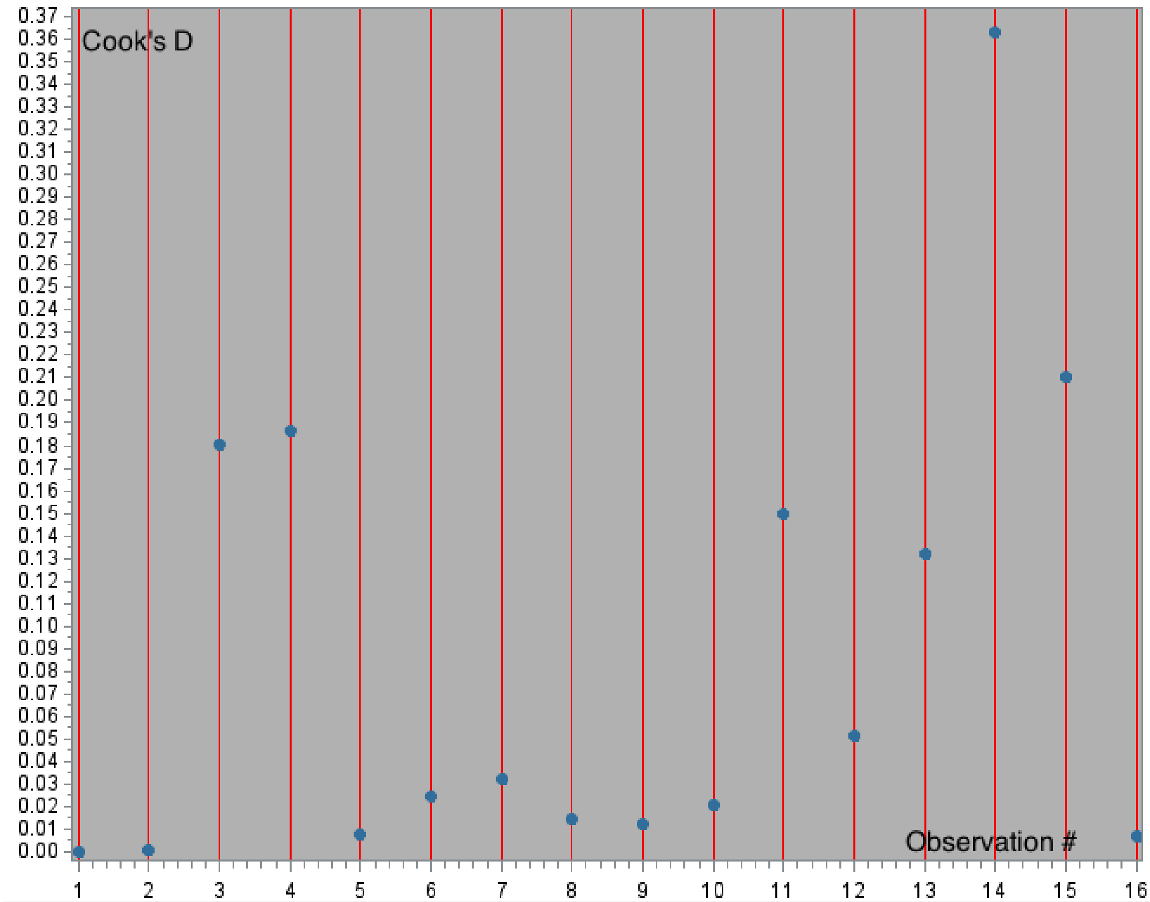
The results of comparing the values of Cook's D, DFFITS, and DFBETAS with their respective cut off values suggests observation 14 has influence on its own fitted value as well as all of the fitted values of y but is not influential on the regression coefficients. Conclude observation 14 is likely an influential case based on DFFITS and Cook's D.

g)

Figure 8: Cook's Distance

Obs	liking	moisture	sweetness	cookd
1	64	4	2	0.00019
2	73	4	4	0.00042
3	61	4	2	0.18039
4	76	4	4	0.18626
5	72	6	2	0.00767
6	80	6	4	0.02455
7	71	6	2	0.03230
8	83	6	4	0.01435
9	83	8	2	0.01223
10	89	8	4	0.02041
11	86	8	2	0.14983
12	93	8	4	0.05098
13	88	10	2	0.13182
14	95	10	4	0.36341
15	94	10	2	0.21066
16	100	10	4	0.00676

Figure 9: Index Plot for Cook's Distance



Using $4/n = .25$ as cut-off value for Cook's D.
Only observation 14 should be flagged as influential by this measure
since all other observations $< .25$
Looking at the index plot (Figure 9) observation 14 is noticeably higher than the other values, supporting observation 14 being an influential case.

Ty Darnell
Assignment 9

1

a)

Figure 1: All possible regressions

Number in Model	R-Square	Adjusted R-Square	AIC	SBC	Variables in Model
1	0.6429	0.6314	195.9065	198.89953	concentration
1	0.4461	0.4282	210.3953	213.38831	age
1	0.1197	0.0913	225.6823	228.67535	weight
2	0.7527	0.7362	185.7858	190.27537	concentration age
2	0.7189	0.7001	190.0107	194.50023	concentration weight
2	0.6002	0.5735	201.6366	206.12609	age weight
3	0.8548	0.8398	170.2055	176.19157	concentration age weight

Using the R^2 and adjusted R^2 criteria, the model with concentration, age and weight is the best model since it has the highest R^2 and adjusted R^2 (.8548 and .8398 respectively). This model also has the lowest AIC and SBC values, which indicates it is also the best model by these criteria.

b)

Figure 2: Stepwise Forward Selection $\alpha = .1$

Variable concentration Entered: R-Square = 0.6429 and C(p) = 42.3306

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	19927	19927	55.81	<.0001
Error	31	11068	357.04768		
Corrected Total	32	30996			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	154.66173	9.86110	87830	245.99	<.0001
concentration	-55.55969	7.43706	19927	55.81	<.0001

Variable age Entered: R-Square = 0.7527 and C(p) = 22.4041

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	23329	11665	45.65	<.0001
Error	30	7666.10196	255.53673		
Corrected Total	32	30996			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	176.24154	10.22598	75903	297.03	<.0001
concentration	-43.41076	7.11830	9503.75418	37.19	<.0001
age	-0.65689	0.18002	3402.37605	13.31	0.0010

Variable weight Entered: R-Square = 0.8548 and C(p) = 4.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	26496	8831.84717	56.92	<.0001
Error	29	4499.97363	155.17150		
Corrected Total	32	30996			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	120.04728	14.77370	10246	66.03	<.0001
concentration	-39.93933	5.59995	7893.04452	50.87	<.0001
age	-0.73677	0.14139	4213.18431	27.15	<.0001
weight	0.77642	0.17188	3166.12833	20.40	<.0001

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	concentration	1	0.6429	0.6429	42.3306	55.81	<.0001
2	age	2	0.1098	0.7527	22.4041	13.31	0.0010
3	weight	3	0.1021	0.8548	4.0000	20.40	<.0001

Figure 3: Stepwise Backward Elimination $\alpha = .15$

All Variables Entered: R-Square = 0.8548 and C(p) = 4.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	26496	8831.84717	56.92	<.0001
Error	29	4499.97363	155.17150		
Corrected Total	32	30996			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	120.04728	14.77370	10246	66.03	<.0001
concentration	-39.93933	5.59995	7893.04452	50.87	<.0001
age	-0.73677	0.14139	4213.18431	27.15	<.0001
weight	0.77642	0.17188	3166.12833	20.40	<.0001

All variables left in the model are significant at the 0.1500 level.

The best model selected by both stepwise forward selection and backward elimination is the model with all three independent variables.

$$clear\hat{a}nce = 120.047 - .39.939concentration - .737age + .776weight$$

This has the highest R^2 and lowest C(p). Lower values of Mallow's C(p) indicate that the model is relatively precise.

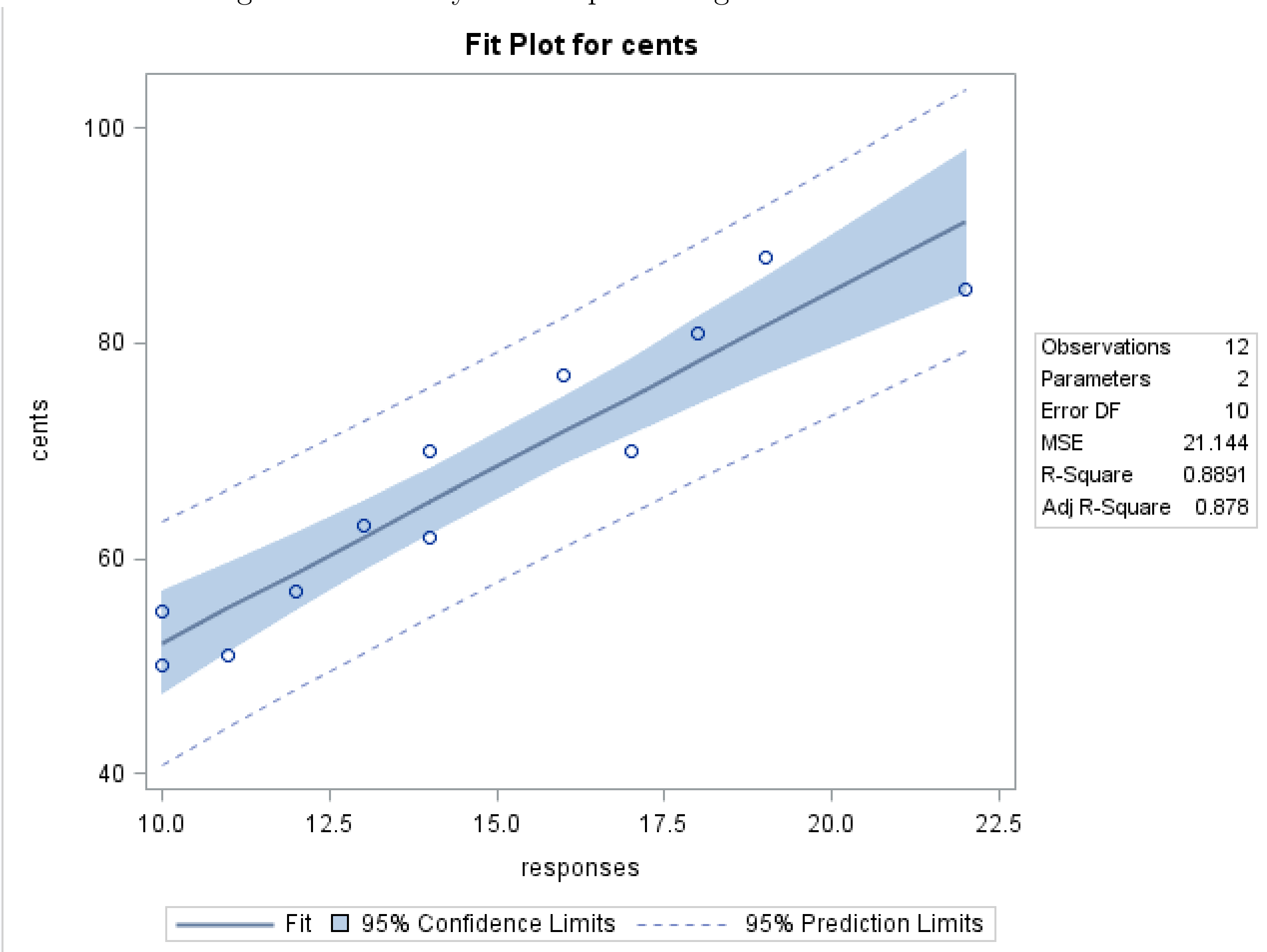
c)

The all possible regressions technique and stepwise regression both selected the same model with all three independent variables. Therefore the results from part a and part b support each other.

2

a)

Figure 4: Ordinary Least Squares Regression

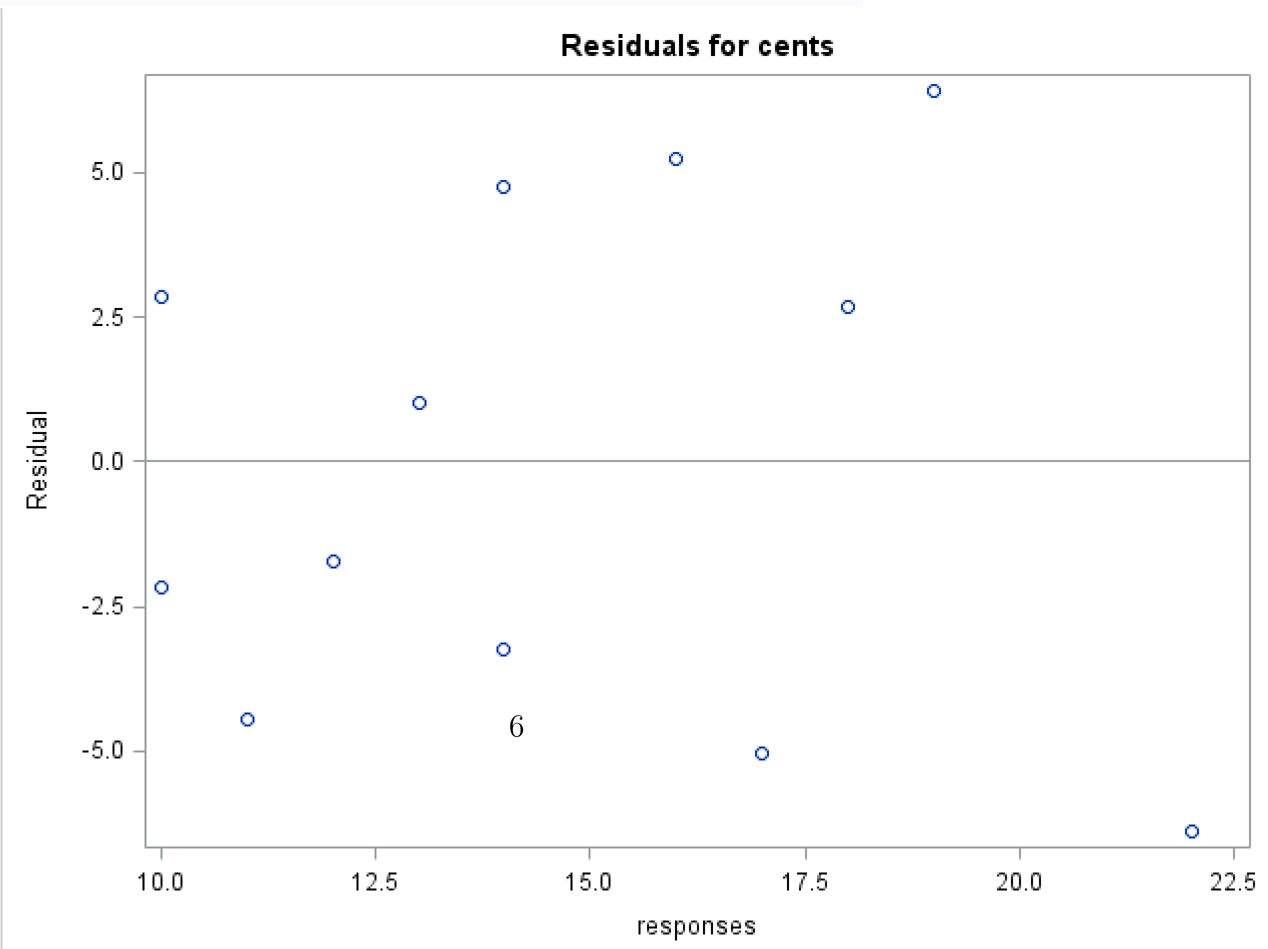


$\hat{cents} = 19.473 + 3.269 responses$

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1695.47339	1695.47339	80.19	<.0001
Error	10	211.44328	21.14433		
Corrected Total	11	1906.91667			

Root MSE	4.59830	R-Square	0.8891
Dependent Mean	67.41667	Adj R-Sq	0.8780
Coeff Var	6.82071		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	19.47269	5.51618	3.53	0.0054
responses	1	3.26891	0.36505	8.95	<.0001



The residual plot suggests heteroscedasticity, since there is a fanning pattern of the residuals.

b)

Figure 5: Residuals split into two groups by size of fitted value

Obs	cents	responses	pred	resid	responses1
1	50	10	52.1618	-2.16176	1
2	55	10	52.1618	2.83824	1
3	51	11	55.4307	-4.43067	1
4	57	12	58.6996	-1.69958	1
5	63	13	61.9685	1.03151	1
6	70	14	65.2374	4.76261	1
7	62	14	65.2374	-3.23739	0
8	77	16	71.7752	5.22479	0
9	70	17	75.0441	-5.04412	0
10	81	18	78.3130	2.68697	0
11	88	19	81.5819	6.41807	0
12	85	22	91.3887	-6.38866	0

Figure 6: Brown-Forsythe test $\alpha = .05$

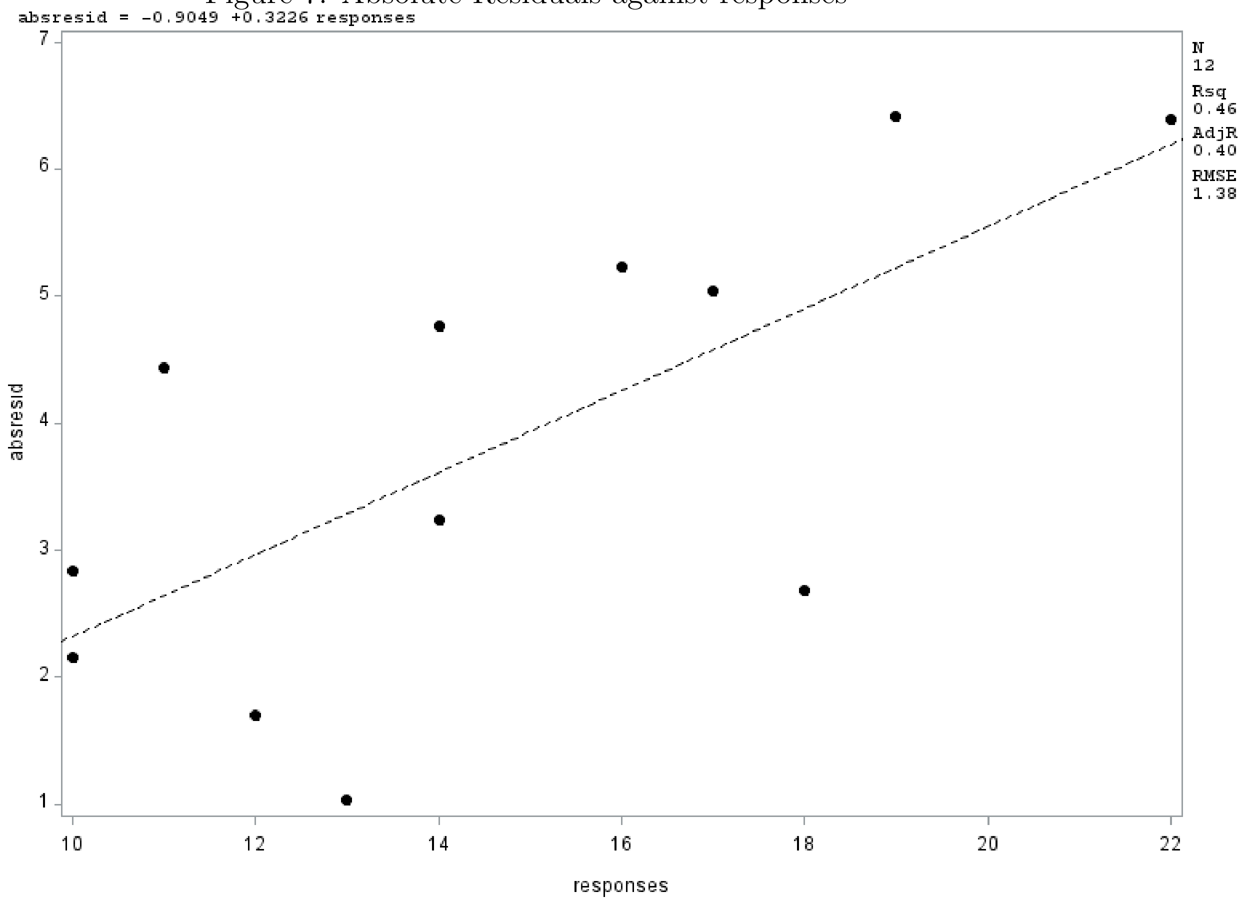
Brown and Forsythe's Test for Homogeneity of cents Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
responses1	1	10.0833	10.0833	0.37	0.5568
Error	10	272.8	27.2833		

Level of responses1	N	cents	
		Mean	Std Dev
0	6	77.1666667	9.74508423
1	6	57.6666667	7.63326055

Decision Rule:
If $p - value \leq \alpha$, conclude the error variance is not constant.
If $p - value > \alpha$ conclude the error variance is significantly different from constant.
Conclusion:
Since p-value= .5568 > .05, we conclude that the error variance is not significantly different constant and does not vary significantly with the level of the predicted values.

c)

Figure 7: Absolute Residuals against responses



The linear regression line is not a good fit for the absolute residual plot. This suggests that the standard deviation of the error term varies with the level of responses.

d)

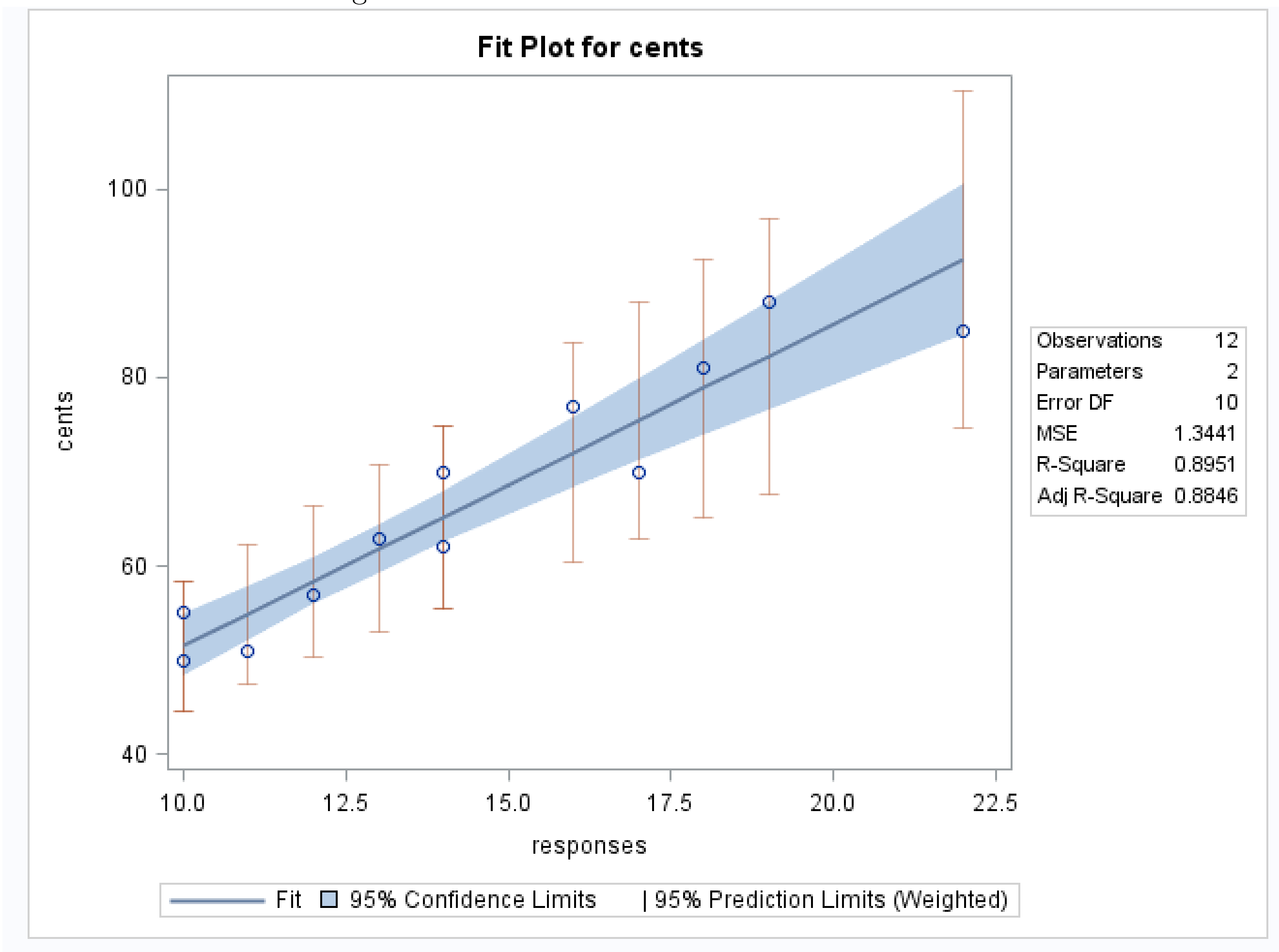
Figure 8: Weighted Least Squares Regression

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	114.71892	114.71892	85.35	<.0001
Error	10	13.44104	1.34410		
Corrected Total	11	128.15996			

Root MSE	1.15935	R-Square	0.8951
Dependent Mean	60.69528	Adj R-Sq	0.8846
Coeff Var	1.91012		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	17.30064	4.82774	3.58	0.0050
responses	1	3.42111	0.37031	9.24	<.0001

Figure 9: Fit Plot for WLS



$\hat{cents} = 17.301 + 3.421 responses$

e

Regression function from OLS:
$\hat{cents} = 19.473 + 3.269 responses$
Regression function from WLS:
$\hat{cents} = 17.301 + 3.421 responses$
The R^2 from the WLS model is .8951 compared to .8891 from the OLS model. Comparing adjusted R^2 WLS model is .8946 compared to .878 from the OLS model. The WLS model is a little bit better fit based on it's slightly higher

adjusted R^2.