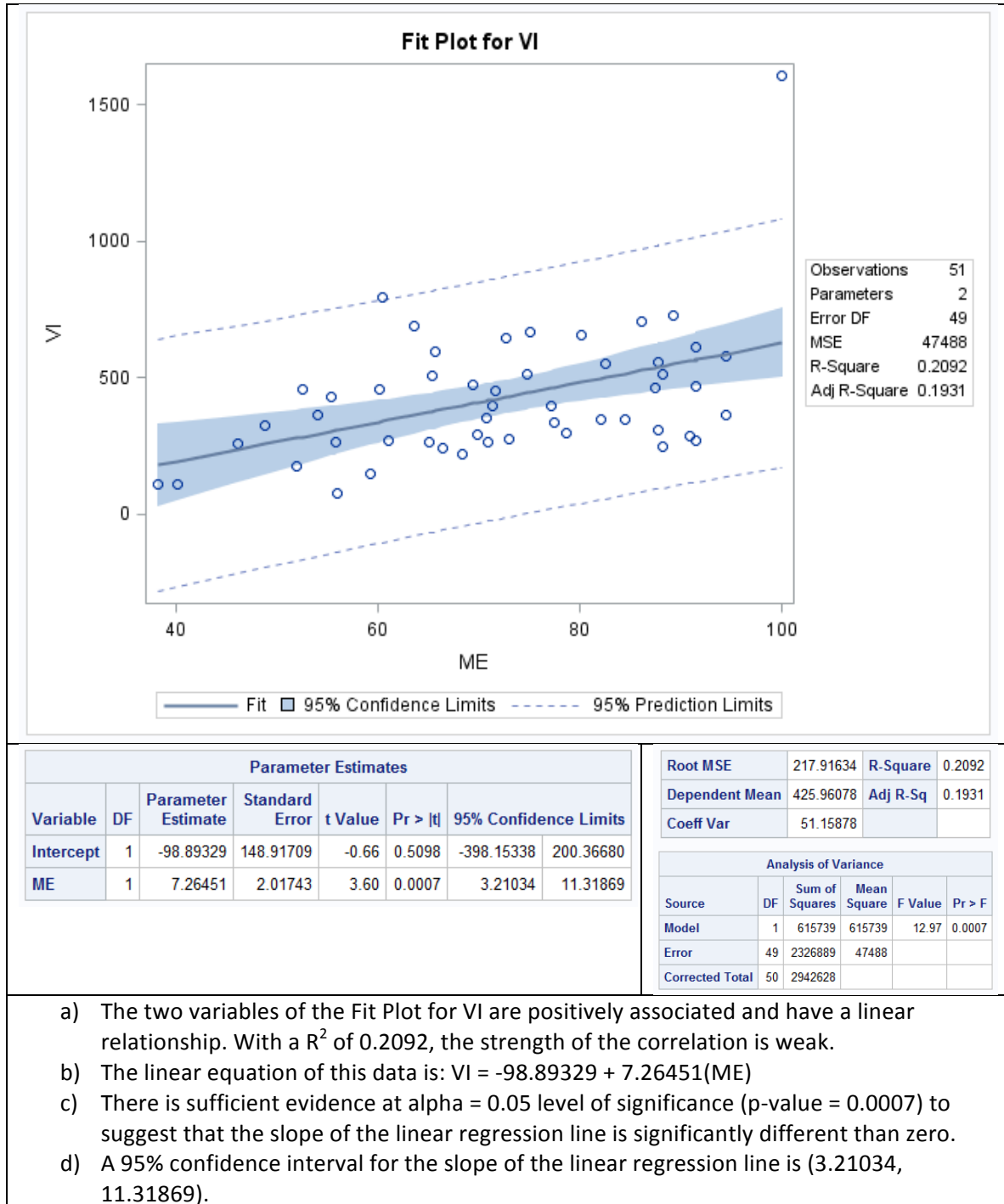


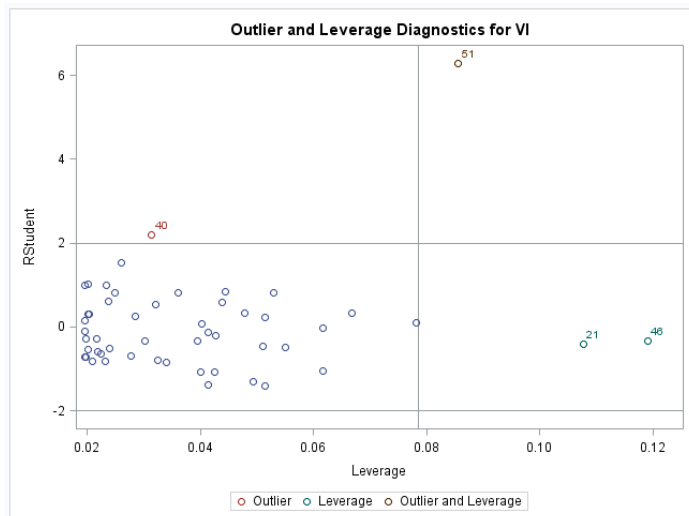
Name: James Tsai
Section: MSDS6371-401
Date: 11/14/15

Question 1.



Question 2.

- a) The assumptions of the linear regression model are:
- Linearity
 - Normality
 - Constant Variance
 - Independence.
- b) The scatter plot of the residual versus the predicted values should have the following:
- The residuals bounce “randomly” around the line, which suggests that the assumption that the relationship is linear is well founded.
 - The residuals form a horizontal band around the line, which suggests that variances of error are relatively equal.
- c) The QQplot for the residuals should look like a straight line from bottom left to upper right if the residuals are normally distributed.
- d) The scatter plot of the residual versus the predicted values should have the following:
- The residuals bounce “randomly” around the line, which suggests that the assumption that the relationship is linear is well founded.
 - The residuals form a relatively straight band around the line, which suggests that variances of error are relatively equal.
 - No one residual stands out, which suggest there are no outliers.
- e) If the constant variance assumption were violated, we would not see a consistent band of points around the data. For example, the plot would look like a cone if variances increased as the explanatory variable increased.
- f) Yes there are 3 high leverage values in the data set. Observations 51, 46, and 21 all meet the criteria of high leverage based on the output from SAS. These three observations correspond to the states DC, VT, and ME respectively. The high leverage values correspond to values leverage greater than 0.078 on the explanatory axis as calculated by $2p/n = 2*(2/51) = 0.078$.



To confirm the leverage values, we use the following equation and data:

Leverage Equation	$h_i = \frac{1}{(n-1)} \left[\frac{X_i - \bar{X}}{s_X} \right]^2 + \frac{1}{n}$	
X-bar = 72.25	n = 51	s _x = 15.276
State	ME	Leverage Value
DC	100.0	0.086
VT	38.2	0.1194
ME	40.2	0.108

Question 3.

REGRESSION ANALYSIS

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	2.57533	0.24873	10.35	<.0001
time	1	-0.32400	0.04330	-7.48	<.0001

Analysis of Variance

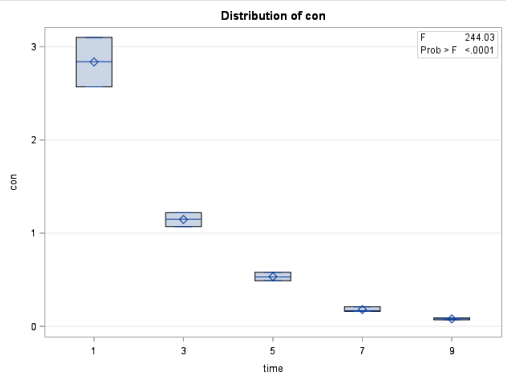
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	12.59712	12.59712	55.99	<.0001
Error	13	2.92465	0.22497		
Corrected Total	14	15.52177			

a) The estimate slope for the regression line is -0.32400. This means for every increase in unit of time, the concentration decreases by 0.32400.

ANOVA ANALYSIS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	15.36437333	3.84109333	244.03	<.0001
Error	10	0.15740000	0.01574000		
Corrected Total	14	15.52177333			

Distribution of con



b) 1. $H_0: \mu_1 = \mu_3 = \mu_5 = \mu_7 = \mu_9$, H_A : At least one pair of means is different

2. Skip

3. F-Value = 244.03

4. P-Value < 0.0001

5. Reject H_0

6. There is sufficient evidence to suggest that at least one pair of means is different.

c) The sum of squares total are the same because the corrected total is the always based on the grand mean model; that is to say our model (variance between group) and error (variance within group) sum of squares must account for the total sum of squares regardless of which model we are using. A model that fits perfectly would have a R^2 of 1 since the model would account for all sum of squares (since $R^2 = \text{Model}/\text{Corrected Total}$)

d) The sum of squares error for the regression line is greater than the sum of squares error in the ANOVA analysis since the regression line has to account not only for variance within the group, but also a “lack of fit” error. In this case, the regression line is a poor

fit, as it does not pass thru the mean of each group. The plot suggests a curved relation and therefore a “lack of fit” error is added. In the ANOVA analysis, it uses the separate means model and only has to account for error within the groups.

e) See answer d)

LACK OF FIT ANALYSIS

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	12.59712	12.59712	55.99	<.0001
Error	13	2.92465	0.22497		
Lack of Fit	3	2.76725	0.92242	58.60	<.0001
Pure Error	10	0.15740	0.01574		
Corrected Total	14	15.52177			

	DF	SS	MS	F-Value	P-Value
Regression	13	2.92465	0.22497		
ANOVA	10	0.1574	0.01574		
Lack of Fit	3	2.76725	0.92242	58.60	<.0001

f) 1. H_0 : The linear regression model fits the data, H_A : The linear regression model does not fit the data
 2. Skip
 3. F-Value = 58.60
 4. P-Value < 0.0001
 5. Reject H_0
 6. There is sufficient evidence to suggest that the linear regression model does not fit the data.