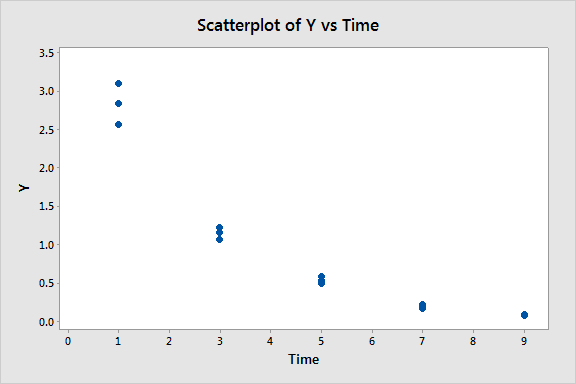
**STAT 501 – Homework 8 – Covering Lesson 9 – Due Date Oct 25**

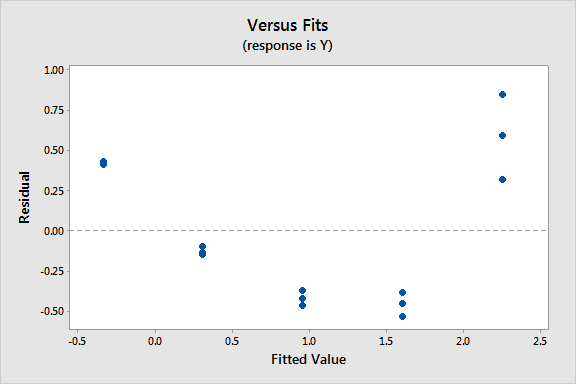
**Instructions**: Use Word to type your answers within this document. Then, submit your answers in the appropriate dropbox in ANGEL by the due date. The point distribution is located next to each question. If there are multiple parts, then the points are divided equally over the subparts.

1. **(6x4 = 24 points)** Use the “Solution Concentrations” dataset. Let *y* = concentration of a chemical solution and *x* = time (hours) since preparation of the solution. There are *n* = 15 observations with 3 observations at each of five different times (unique times are 1, 3, 5, 7, and 9 hours).
   1. The first two columns of the dataset give *y* and *x*. Graph *y* versus *x*. What are the noteworthy features of this plot? For instance, what is the direction of the association, are the data linear or curvilinear, is there any non-constant variance issue?



**There is a negative, curvilinear association. The variance decreases as time increases (or it could be said that it is larger when mean concentration is higher).**

* 1. Use statistical software to fit a straight-line model to the data using *y* = concentration and *x* = time. As part of doing the regression, request a graph of residuals. [In Minitab, use Stat > Regression > Regression; in the dialog box for setting up the regression, click Graphs to request the residual plot.] Write a brief interpretation of the “residuals versus fits” plot. Describe the model difficulties a residual plot with the pattern that you see indicates.

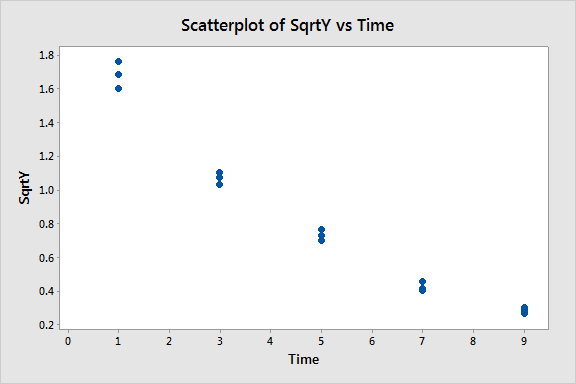


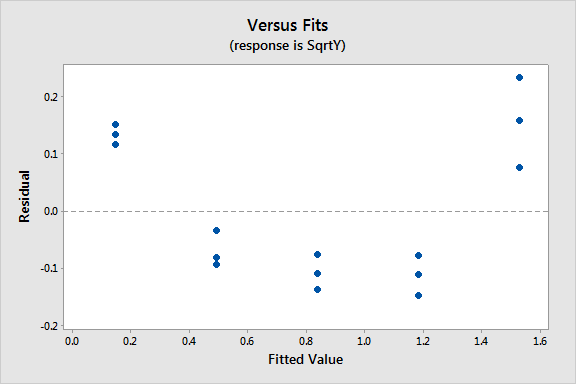
**The pattern is curved and also shows nonconstant variance. Thus the straight-line equation is not suitable and the assumption of constant variance is not valid.**

* 1. Refer back to the graphs for the previous two parts. Explain why it would be best to first try transforming the *y*-variable in this case, rather than the *x*-variable.

**Transforming *y* will affect the variance, which is something we need here because the variance is not constant.**

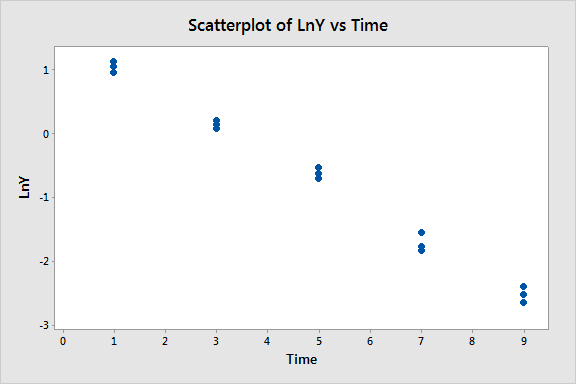
* 1. The third column in the dataset gives the square root of *y* (the square root of the concentration). Using the square root of *y* as the response variable and time as the *x*-variable, plot the square root of *y* versus *x*, fit a regression model, and examine the plot of “residuals versus fits.” Describe your findings.

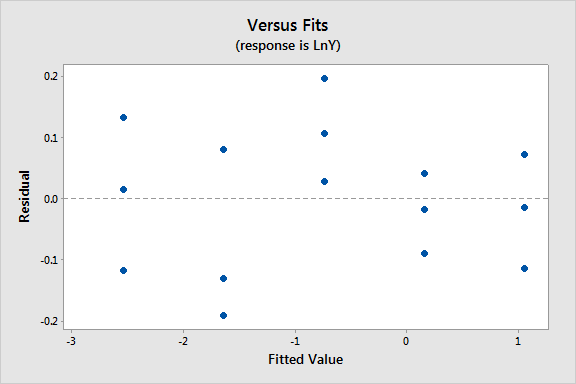




**From both plots we learn that the pattern is still curved and also shows non-constant variance. Thus, the model with the square root of *y* is not suitable and the assumption of constant variance is not valid.**

* 1. The fourth column in the dataset gives ln(*y*) (the natural logarithm of y). Using ln(*y*) as the response variable and time as the *x*-variable, plot ln(*y*) versus *x*, fit a regression model, and examine the plot of “residuals versus fits.” Describe your findings.





**From both plots, we learn that although the appearance is still not perfectly ideal, it is better than for the untransformed *y* and the square root of *y*. The variance looks to be constant.**

* 1. Of the models examined in the previous parts of this problem, which is the “best?” Briefly explain.

**The model with ln(*y*) as the response looks to be the best. The plot of ln(*y*) versus *x* appears close to linear. In the residual plot, the variance looks to be constant and there is no clear curvature.**

1. **(5x4 = 20 points)** The “Diet” dataset is a simulated dataset containing *y* = calories, *x1* = carbohydrate (carb) and *x2*= fat intake for eleven meals. It is believed that both carb and fat have a linear effect on calorie amount generated but that there is a possible interaction effect also.
2. Use Minitab to estimate the regression equation

E(y) = β0 + β1 x1 + β2 x2 + β3 x1 x2

[The command sequence (v17) is Stat 🡪 Regression 🡪 Regression 🡪 Fit regression model 🡪 select “Model” tab 🡪 Select carb and fat under the “Predictors” box and click “Add” for “Interactions through order 2.” You should be able to see “carb\*fat” added to the list of “Terms in the model.”]

What is the fitted regression equation to predict the calorie amount using this model?

**= 33.6 + 6.28 carb + 10.81 fat - 0.2053 carb\*fat**

1. Test whether β3 in the above model is significant at the 5% significance level. What does β3 signify in the context of this problem?

**The p-value is 0.034, which is less than 0.05, so β3 in the above model is significant at the 5% significance level. β3 signifies the interaction effect between carb and fat intake amounts on the calories generated.**

1. Derive prediction equations to compute the calorie amount based on the fat intake for a meal consisting of:
2. 20 carb units

**= 33.6 + 6.28(20) + 10.81 fat - 0.2053(20)\*fat = 159.2 + 6.704 fat**

1. 30 carb units

**= 33.6 + 6.28(30) + 10.81 fat - 0.2053(30)\*fat = 222.0 + 4.651 fat**

1. 40 carb units

**= 33.6 + 6.28(40) + 10.81 fat - 0.2053(40)\*fat = 284.8 + 2.598 fat**

1. 50 carb units

**= 33.6 + 6.28(50) + 10.81 fat - 0.2053(50)\*fat = 347.6 + 0.545 fat**

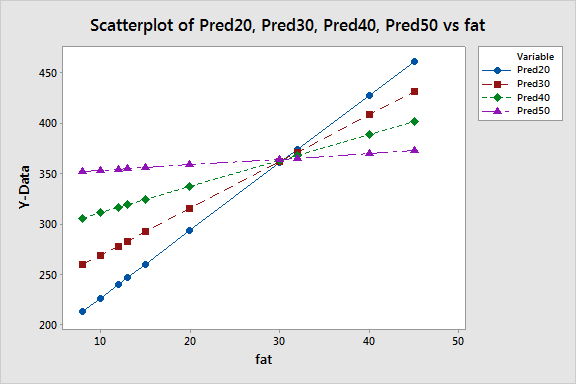
*[Hint: Use the equation from part (a), plug in each value of carb, and simplify.]*

1. Overlay the four regression equations in part (c) on the same graph.

[One way to do this in Minitab is to use Calc > Calculator to calculate four new variables representing the prediction equations for 20, 30, 40, and 50 carb units. Then select Graph > Scatterplot (with Connect Line) and select four pairs of Y and X variables:

1. Y = prediction equation for 20 carb units and X = fat
2. Y = prediction equation for 30 carb units and X = fat
3. Y = prediction equation for 40 carb units and X = fat
4. Y = prediction equation for 50 carb units and X = fat

Finally click Multiple Graphs and select “Overlaid on the same graph.”]



1. Use your answers to the previous parts to comment on the interaction effect between carb and fat intake on the amount of calories generated.

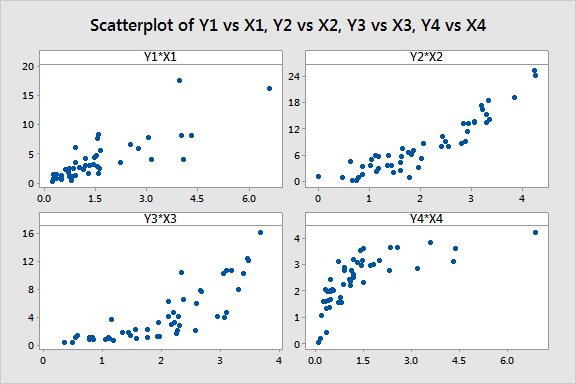
*[Hint: Compare the slopes of the prediction equations in part (c).]*

**The interaction effect is statistically significant and is estimated to be -0.2053. This means that the calorie amount generated by fat depends on the carb intake value. The fact that the estimated value is negative means that, as the carb value increases, the slopes of the calories vs fat lines decrease (see the graphs above and the line equations in part c). This essentially means that the calorie amount increase with respect to fat decreases as the amount of carb increases in the meal.**

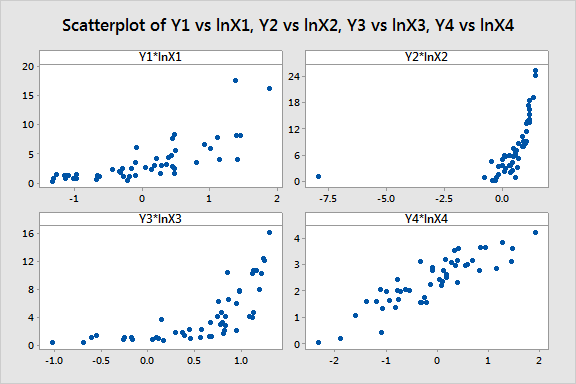
1. **(4x8 = 32 points)** The “Transformations” dataset is a simulated dataset containing four pairs of variables (*y1*, *x1*), (*y2*, *x2*), (*y3*, *x3*), and (*y4*, *x4*). Each pair of variables is best modeled using linear regression with one of the following four models:
2. Response variable *y*, predictor variable ln(*x*) [natural logarithm]
3. Response variable ln(*y*), predictor variable *x*
4. Response variable ln(*y*), predictor variable ln(*x*)
5. Response variable *y*, predictor variables *x* and *x2*

Your task is to determine which model goes with each pair of variables. You should use each model exactly once. Include Minitab plots and output to support your conclusions.

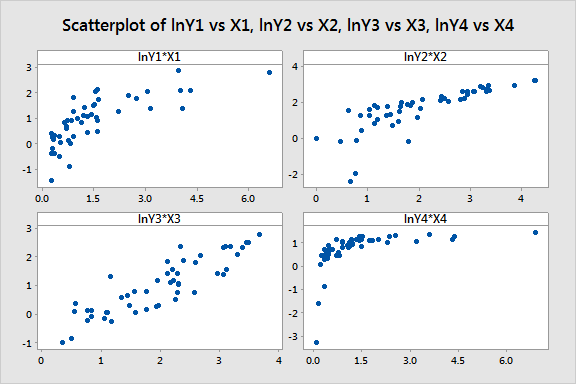
**Scatterplots of each pair of variables indicate the need for transformations since none satisfy the LINE conditions:**



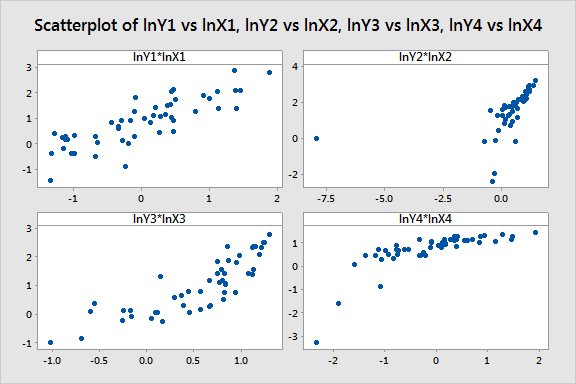
**Scatterplots of y versus ln(x) for each pair of variables indicate (*y4*, *x4*) satisfy the LINE conditions best:**



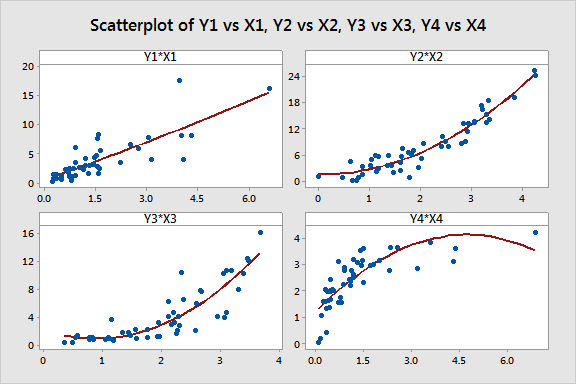
**Scatterplots of ln(y) versus x for each pair of variables indicate (*y3*, *x3*) satisfy the LINE conditions best:**



**Scatterplots of ln(y) versus ln(x) for each pair of variables indicate (*y1*, *x1*) satisfy the LINE conditions best:**



**Scatterplots of y versus x for each pair of variables with fitted quadratic regression lines added indicate a quadratic model fits the (*y2*, *x2*) data best:**



**So, (*y1*, *x1*) is best modeled with response variable ln(*y*), predictor variable ln(*x*):**

**Regression Analysis: lnY1 versus lnX1**

**Analysis of Variance**

**Source DF Adj SS Adj MS F-Value P-Value**

**Regression 1 29.15 29.1489 109.53 0.000**

**lnX1 1 29.15 29.1489 109.53 0.000**

**Error 48 12.77 0.2661**

**Total 49 41.92**

**Model Summary**

**S R-sq R-sq(adj) R-sq(pred)**

**0.515875 69.53% 68.89% 67.04%**

**Coefficients**

**Term Coef SE Coef T-Value P-Value VIF**

**Constant 0.8446 0.0730 11.57 0.000**

**lnX1 0.9137 0.0873 10.47 0.000 1.00**

**So, (*y2*, *x2*) is best modeled with response variable *y*, predictor variables *x* and *x2*:**

**Regression Analysis: Y2 versus X2**

**Analysis of Variance**

**Source DF Adj SS Adj MS F-Value P-Value**

**Regression 2 1699.63 849.816 253.18 0.000**

**X2 1 0.07 0.066 0.02 0.889**

**X2\*X2 1 107.07 107.073 31.90 0.000**

**Error 47 157.76 3.357**

**Total 49 1857.39**

**Model Summary**

**S R-sq R-sq(adj) R-sq(pred)**

**1.83209 91.51% 91.15% 90.64%**

**Coefficients**

**Term Coef SE Coef T-Value P-Value VIF**

**Constant 1.55 1.01 1.53 0.132**

**X2 -0.14 1.03 -0.14 0.889 16.67**

**X2\*X2 1.294 0.229 5.65 0.000 16.67**

**So, (*y3*, *x3*) is best modeled with response variable ln(*y)*, predictor variable *x*:**

**Regression Analysis: lnY3 versus X3**

**Analysis of Variance**

**Source DF Adj SS Adj MS F-Value P-Value**

**Regression 1 35.267 35.2666 173.11 0.000**

**X3 1 35.267 35.2666 173.11 0.000**

**Error 48 9.779 0.2037**

**Total 49 45.045**

**Model Summary**

**S R-sq R-sq(adj) R-sq(pred)**

**0.451359 78.29% 77.84% 76.65%**

**Coefficients**

**Term Coef SE Coef T-Value P-Value VIF**

**Constant -0.865 0.157 -5.51 0.000**

**X3 0.9334 0.0709 13.16 0.000 1.00**

**So, (*y4*, *x4*) is best modeled with response variable *y*, predictor variable ln(*x)*:**

**Regression Analysis: Y4 versus lnX4**

**Analysis of Variance**

**Source DF Adj SS Adj MS F-Value P-Value**

**Regression 1 32.591 32.5906 167.37 0.000**

**lnX4 1 32.591 32.5906 167.37 0.000**

**Error 48 9.346 0.1947**

**Total 49 41.937**

**Model Summary**

**S R-sq R-sq(adj) R-sq(pred)**

**0.441269 77.71% 77.25% 75.86%**

**Coefficients**

**Term Coef SE Coef T-Value P-Value VIF**

**Constant 2.4817 0.0628 39.50 0.000**

**lnX4 0.9112 0.0704 12.94 0.000 1.00**

**Residual plots confirm our findings:**

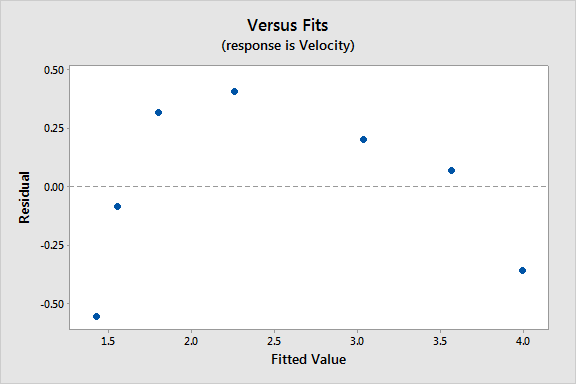
1. **(6x4 = 24 points)** Open the “Chemical Reaction” dataset. Measurements concerning the rate of velocity of a chemical reaction for different concentrations of a substrate were taken. We are interested in characterizing the relationship between concentration *(x)* and velocity *(y)*.
2. Fit a quadratic regression model and report the estimated regression equation. [Remember to use the Model button in the Regression dialog in Minitab v17.]

**The estimated regression equation is .**

1. Perform a statistical test of the significance of the quadratic term. Specifically, report the test statistic, degrees of freedom, *p*-value, and your conclusion in the context of the problem.

**The test statistic is –3.33 with 4 degrees of freedom, the *p*-value is 0.029, so we reject the null hypothesis and conclude that the quadratic term is a significant predictor of velocity.**

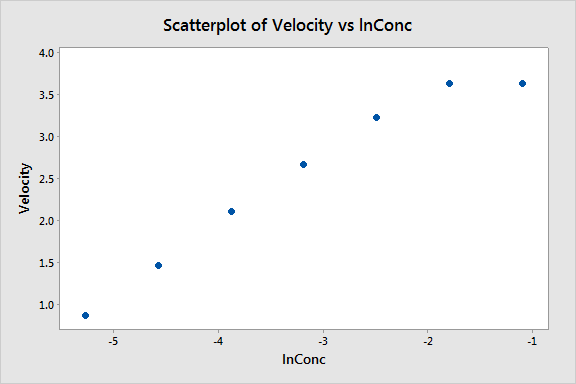
1. Visually assess a plot of the residuals versus the fitted values and comment on any violations to the regression assumptions.



**Given the strong nonlinear pattern there appears to be an overall lack of fit with respect to the linearity condition.**

1. Create a new variable equal to the natural logarithm of concentration. [In Minitab, use Calc > Calculator, enter a new variable name (such as lnConc) in the Store result in variable box, and in the Expression box enter ln(Concentration).] Provide a plot of velocity versus the natural logarithm of concentration. Also, fit and report the estimated equation for the following regression model:

*Veli*= β0 + β1 ln(*Conci*) + β2 (ln(*Conci*))2 + ε*i*



**The estimated regression model is:**

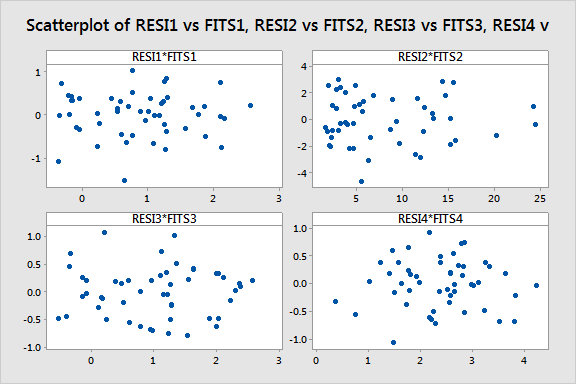
**.**

1. Perform a statistical test of the significance of the quadratic term in the previous model. Again, report the test statistic, degrees of freedom, p-value, and your conclusion in the context of the problem.

**The test statistic is –4.04 with 4 degrees of freedom, the *p*-value is 0.016, and so we reject the null hypothesis and conclude that the quadratic term is a significant predictor of velocity. Note that even though the linear term is not statistically significant, it is still retained due to the hierarchy principle.**

1. Using the model in part (d), what is a 95% prediction interval for velocity at a concentration level of *e*–3? Interpret this interval in the context of the problem.

**Since ln(*e*−3) = −3, we are just plugging –3 into our predicted model above. The resulting interval is (2.47741, 3.20441). This means that with 95% confidence, we expect the velocity to be between 2.47741 and 3.20441 for a future concentration level of *e*−3.**

****