**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 in the Lesson 9 folder. 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?



The direction of the association is negative i.e. as time is increasing the solution concentration is decreasing. The relationship appears to be curvilinear.

The three points at time = 1 have more spread than the remaining data points at a given time. There appears to be a violation of the equal variance assumption. However the data size is too small to state concretely without further analysis - we can say more about this after analyzing the residuals-fits plot.

* 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 minitab output is:

**Regression Analysis: Y versus Time**

Analysis of Variance

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

Regression 1 12.5971 12.5971 55.99 0.000

Time 1 12.5971 12.5971 55.99 0.000

Error 13 2.9247 0.2250

Lack-of-Fit 3 2.7673 0.9224 58.60 0.000

Pure Error 10 0.1574 0.0157

Total 14 15.5218

Model Summary

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

0.474314 81.16% 79.71% 73.45%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 2.575 0.249 10.35 0.000

Time -0.3240 0.0433 -7.48 0.000 1.00

Regression Equation

Y = 2.575 - 0.3240 Time

|  |  |  |
| --- | --- | --- |
|  |  |  |

The above plot indicates that the normality assumption needs further validation. So we do Ryan Joiner test:



The following points are noteworthy:

* The residual plot confirms that the "equal variance" assumption is violated: the residuals vs. fits plot exhibits fanning and therefore provides yet more evidence that the variance of the error terms might not be equal.
* The RJ test gives p-value = 0.098 > alpha = 0.05 (assumed). We fail to reject the null hypothesis that data is normally distributed. There isn’t enough evidence to doubt the normality assumption.
* From the residuals – fits plot it is further clear that the relationship between the variables is curvilinear.
  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.

We see that there are some issues with the error terms – they don’t have equal variance. Going by the heuristic that transforming the y values corrects problems with the error terms (unequal variances and/or non-normality) and may also help with non-linearity. Therefore it will be prudent to try transforming the y-variable to start with.

* 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.

The plot is as follows:

 The relationship still appears to be curvilinear although to a lesser extent. The three points at time = 1 have more spread than the remaining data points at a given time. There appears to be a violation of the equal variance assumption. However the data size is too small to state concretely without further analysis - we can say more about this after analyzing the residuals-fits plot.

The regression model:

Analysis of Variance

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

Regression 1 3.56592 3.56592 198.50 0.000

Time 1 3.56592 3.56592 198.50 0.000

Error 13 0.23354 0.01796

Lack-of-Fit 3 0.21411 0.07137 36.74 0.000

Pure Error 10 0.01943 0.00194

Total 14 3.79946

Model Summary

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

0.134032 93.85% 93.38% 91.29%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 1.6998 0.0703 24.18 0.000

Time -0.1724 0.0122 -14.09 0.000 1.00

Regression Equation

SqrtY = 1.6998 - 0.1724 Time

|  |  |
| --- | --- |
|  |  |



The following points are noteworthy:

* The residual plot confirms that the "equal variance" assumption is violated: the residuals vs. fits plot exhibits fanning and therefore provides yet more evidence that the variance of the error terms might not be equal.
* From the residuals – fits plot it is further clear that the relationship between the variables is curvilinear.
* The R-square has increased from 79.71 to 93.38%
* The RJ test gives p-value = 0.056 > alpha = 0.05 (assumed). The p-value is at the border line.

In summary while the y-transformation has increased the R-squared, it has not improved the situation substantially as far as the characteristics of the residuals are concerned.

* 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.

 The relationship appears to be linear. The situation with equal variance has also improved from the previous models.

The regression model:

Analysis of Variance

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

Regression 1 24.2946 24.2946 1840.86 0.000

Time 1 24.2946 24.2946 1840.86 0.000

Error 13 0.1716 0.0132

Lack-of-Fit 3 0.0590 0.0197 1.75 0.220

Pure Error 10 0.1125 0.0113

Total 14 24.4662

Model Summary

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

0.114880 99.30% 99.24% 99.08%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 1.5080 0.0602 25.03 0.000

Time -0.4500 0.0105 -42.91 0.000 1.00

Regression Equation

LnY = 1.5080 - 0.4500 Time

|  |  |
| --- | --- |
|  |  |



The following points are noteworthy:

* The residual plot confirms that the "equal variance" assumption is **valid**: the residuals vs. fits plot exhibits equal variance
* The RJ test gives p-value > 0.100 > alpha = 0.05 (assumed). We fail to reject the null hypothesis that data is normally distributed. There isn’t enough evidence to doubt the normality assumption
* The R-square has increased to 99.24%
* From the residuals – fits plot it is further clear that the relationship between the variables is linear since there is no discernible pattern in the residuals – fits plot.

In summary the y-transformation has improved the situation on multiple accounts: it has increased the R-squared, it has improved the situation substantially as far as the characteristics of the residuals are concerned: they now exhibit equal variance.

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

The model in part e (ln(y) as the response variable and time as the x-variable) exhibits the best characteristics:

* The residual plot confirms that the "equal variance" assumption is **valid**: the residuals vs. fits plot exhibits equal variance
* The RJ test gives p-value > 0.100 > alpha = 0.05 (assumed). We fail to reject the null hypothesis that data is normally distributed. There isn’t enough evidence to doubt the normality assumption
* The R-square has increased to 99.24%
* From the residuals – fits plot it is further clear that the relationship between the variables is linear since there is no discernible pattern in the residuals – fits plot.

In summary the y-transformation has improved the situation on multiple accounts: it has increased the R-squared, it has improved the situation substantially as far as the characteristics of the residuals are concerned: they now exhibit equal variance.

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.
   1. 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?

Calories = 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?

β3 has a p-value = 0.034 < alpha = 0.05 and so we can reject the null hypothesis that β3 = 0. Therefore we can conclude that β3 in the above model is significant at the 5% significance level

Interpretation: In the presence of β3 (coefficient for interaction bet carb and fat), the meaning of the regression coefficients β1 and β2 changes from the earlier definition: The regression coefficients β1 and β2 no longer indicate the change in the mean response with a unit increase of the predictor variable, with the other predictor variable held constant at any given level. It can be shown that the change in the mean response with a unit increase in X I when X2 is held constant is:

β1 + β3 X2

Similarly, the change in the mean response with a unit increase in X2 when XI is held constant is:

Β2 + β3 X1

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

1. 20 carb units
2. 30 carb units
3. 40 carb units
4. 50 carb units

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

1. Calories = 33.6 + 6.28 \* 20 + (10.81 - 0.2053 \* 20) \* fat = 159.2 + 6.704 \* fat
2. Calories = 33.6 + 6.28 \* 30 + (10.81 - 0.2053 \* 30) \* fat = 222 + 4.651 \* fat
3. Calories = 33.6 + 6.28 \* 40 + (10.81 - 0.2053 \* 40) \* fat = 284.8 + 2.598 \* fat
4. Calories = 33.6 + 6.28 \* 50 + (10.81 - 0.2053 \* 50) \* fat = 347.6 + 0.545 \* fat
   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 slope of the prediction equation at carb level 20, 30, 40 and 50 is 6.704, 4.651, 2.598 and 0.545 respectively. The slope is different in each case and this is in alignment with the fact that a significant interaction effect implies that the slopes are different.

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.

1. Y, ln(x)

The scatterplot for this is:



The regression output is:

|  |  |  |  |
| --- | --- | --- | --- |
| Y1, ln(X1) | Y2, ln(X2) | Y3, ln(X3) | Y4, ln(X4) |
|  |  |  |  |

**Selected Model: 4** because: linear relationship, equal variance and normality of residuals, assume indepencence. The other 3 models show a curvilinear relationship in the scatterplots and the residual vs fits plot.

1. ln(y), x

The scatterplot for this is:



The regression output is:

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 2 | 3 | 4 |
|  |  |  |  |

**Selected Model:** 3, because: linear relationship, equal variance and normality of residuals, assume indepencence. The other 3 models show a curvilinear relationship in the scatterplots and the residual vs fits plot. The residuals plot for the other 3 models also show that equal variance assumption is violated.

1. ln(y), ln(x)

The scatterplot for this is:



The regression output is:

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 2 | 3 | 4 |
|  |  |  |  |

**Selected Model:** 1, because: linear relationship, equal variance and normality of residuals, assume indepencence. Models 2 and 4 show a curvilinear relationship in the scatterplots and the residual vs fits plot. Model 3 has lesser extent of curvilinear but the equal variance assumption is also looking doubtful. Model 1 has better characteristics as compared to model 3.

1. y with x and x2

y1 with x1 and x12



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

2.03235 68.73% 67.40% 60.34%

y2 with x2 and x22



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

1.83209 91.51% 91.15% 90.64%

y3 with x3 and x32



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

1.90713 76.94% 75.96% 74.04%

y4 with x4 and x42



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

0.552775 65.76% 64.30% 1.28%

**Selected Model: 2**, because:model1 and 3 show fanning in the residuals-fits plot. Model 4 residuals indicate a pattern. Model 2 presents good characteristics.

Model 2 minitab output is:

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

X2Sq 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

X2Sq 1.294 0.229 5.65 0.000 16.67

The output shows that while X2Sq is significant, S2 itself isn’t. We keep both variables in accordance with the hierarchical principle. Some points about other models:

Model 1: X1Sq is not significant

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 0.440 0.614 0.72 0.477

X1 2.070 0.646 3.21 0.002 8.64

X1Sq 0.030 0.115 0.27 0.792 8.64

Model 4: Though the coef are significant, the inherent pattern in data is apparent from the lower R-squared

Analysis of Variance

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

Regression 2 27.576 13.7879 45.12 0.000

X4 1 16.170 16.1705 52.92 0.000

X4Sq 1 6.361 6.3611 20.82 0.000

Error 47 14.361 0.3056

Total 49 41.937

Model Summary

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

0.552775 65.76% 64.30% 1.28%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 1.202 0.159 7.56 0.000

X4 1.232 0.169 7.27 0.000 7.60

X4Sq -0.1298 0.0284 -4.56 0.000 7.60

Regression Equation

Y4 = 1.202 + 1.232 X4 - 0.1298 X4Sq

1. **(6X4= 24 points)** Open the “Chemical Reaction Data” dataset in the Lesson 9 folder. 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)*.
   1. Fit a quadratic regression model and report the estimated regression equation. [Remember to use the Model button in the Regression dialog in Minitab v17.]

Analysis of Variance

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

Regression 2 6.2792 3.1396 16.59 0.012

Concentration 1 3.7122 3.7122 19.61 0.011

Concentration\*Concentration 1 2.1043 2.1043 11.12 0.029

Error 4 0.7571 0.1893

Total 6 7.0363

Model Summary

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

0.435052 89.24% 83.86% 0.00%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 1.288 0.275 4.68 0.009

Concentration 25.65 5.79 4.43 0.011 15.16

Concentration\*Concentration -56.5 16.9 -3.33 0.029 15.16

Regression Equation

Velocity = 1.288 + 25.65 Concentration - 56.5 Concentration\*Concentration

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

To do so, we test the hypotheses: H0 : β2 = 0 and HA : β2 ≠ 0

We can do this in 2 ways: Partial F test and the T test

**Partial F test**

The full model: yi=(β0+β1xi1+β2xi2)+ϵi

The reduced model: yi=(β0+β1xi1)+ϵi

The test: general linear statistic: 

F∗ = SSR(x2|x1) / 1 ÷ MSE = 2.1043 / 0.1893 = 11.116

The P-value is the probability — if the null hypothesis were true — that we would get an F-statistic larger than 11.116. Comparing our partial F-statistic to an F-distribution with 1 numerator degree of freedom and 4 denominator degrees of freedom, Minitab tells us that the probability is close to 0.971004 that we would observe an F-statistic smaller than 11.116

F distribution with 1 DF in numerator and 4 DF in denominator

x P( X ≤ x )

11.116 0.971004

Therefore, the probability that we would get an F-statistic larger than 11.116 is: 0.028996

That is, the P-value = 0.029 < alpha (assumed 0.05) and therefore there is sufficient evidence to conclude that the quadratic term is related to Velocity

**T test**

The test and the p-value for this are directly in the regression output. So we have

T\* = Coef / SECoef = -56.5/16.9= -3.33

The p-value for the T test = 0.029 < alpha (assumed 0.05) and therefore there is sufficient evidence to conclude that the quadratic term is related to Velocity

**From both the tests we conclude that there is sufficient evidence to conclude that the quadratic term of concentration of substrate is related to the rate of velocity of the chemical reaction.**

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



RJ test:



The following observations can be made:

* There is a discernible pattern in the residuals vs fits plot pointing to the non linearity of the relationship.
* The residuals seems to be normally distributed as confirmed by the RJ test p-value > 0.1
* The equal variance assumptions looks like it is violated. However the data size is too small to say that concretely.
  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*



Analysis of Variance

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

Regression 2 6.98473 3.49237 270.93 0.000

lnConc 1 0.00105 0.00105 0.08 0.790

lnConc\*lnConc 1 0.21011 0.21011 16.30 0.016

Error 4 0.05156 0.01289

Total 6 7.03629

Model Summary

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

0.113536 99.27% 98.90% 94.83%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 3.920 0.237 16.57 0.000

lnConc 0.048 0.167 0.29 0.790 29.06

lnConc\*lnConc -0.1041 0.0258 -4.04 0.016 29.06

Regression Equation

Velocity = 3.920 + 0.048 lnConc - 0.1041 lnConc\*lnConc



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

To do so, we test the hypotheses: H0 : β2 = 0 and HA : β2 ≠ 0

We can do this in 2 ways: Partial F test and the T test

**Partial F test**

The full model: yi=(β0+β1xi1+β2xi2)+ϵi

The reduced model: yi=(β0+β1xi1)+ϵi

The test: general linear statistic:

F∗ = SSR(x2|x1) / 1 ÷ MSE = 0.21011 / 0.01289 = 16.30

The P-value is the probability — if the null hypothesis were true — that we would get an F-statistic larger than 16.30. Comparing our partial F-statistic to an F-distribution with 1 numerator degree of freedom and 4 denominator degrees of freedom, Minitab tells us that the probability is close to 0.971004 that we would observe an F-statistic smaller than 16.30

F distribution with 1 DF in numerator and 4 DF in denominator

x P( X ≤ x )

16.3 0.984361

Therefore, the probability that we would get an F-statistic larger than 16.30 is: 0.015639

That is, the P-value = 0.016 < alpha (assumed 0.05) and therefore there is sufficient evidence to conclude that the quadratic term is related to Velocity

**T test**

The test and the p-value for this are directly in the regression output. So we have

T\* = Coef / SECoef = -0.1041/ 0.0258 = -4.04

The p-value for the T test = 0.016 < alpha (assumed 0.05) and therefore there is sufficient evidence to conclude that the quadratic term is related to Velocity

**From both the tests we conclude that there is sufficient evidence to conclude that the quadratic term of ln of concentration of substrate is related to the rate of velocity of the chemical reaction.**

* 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.

First of all we must confirm that the predictors are in the range where model was made. e^-3 = 0.0498 which is in the range of the predictor.

We know that ln(e^-3) = -3. So the prediction has to be with the lnConc = -3

We have minitab output as:

**Prediction for Velocity**

Regression Equation

Velocity = 3.920 + 0.048 lnConc - 0.1041 lnConc\*lnConc

Variable Setting

lnConc -3

Fit SE Fit 95% CI 95% PI

2.84091 0.0651959 (2.65990, 3.02192) (2.47741, 3.20441)

The Minitab output tells us that we can be 95% confident that the rate of velocity of a chemical reaction for a randomly selected substrate of concentration e^-3 will be between 2.47741 and 3.20441