**Andrew Hillard**

**Data Analysis 2**

**HW2**

Problem 1

1. Based on the grid search, the potential starting values that minimize the SSE are the following: , , .
2. The logarithmic transformation gives the following model.

The above model can be thought of as a multiple linear regression model where , , , , , and . Using SAS, we solve for the parameters in this model and get , , .

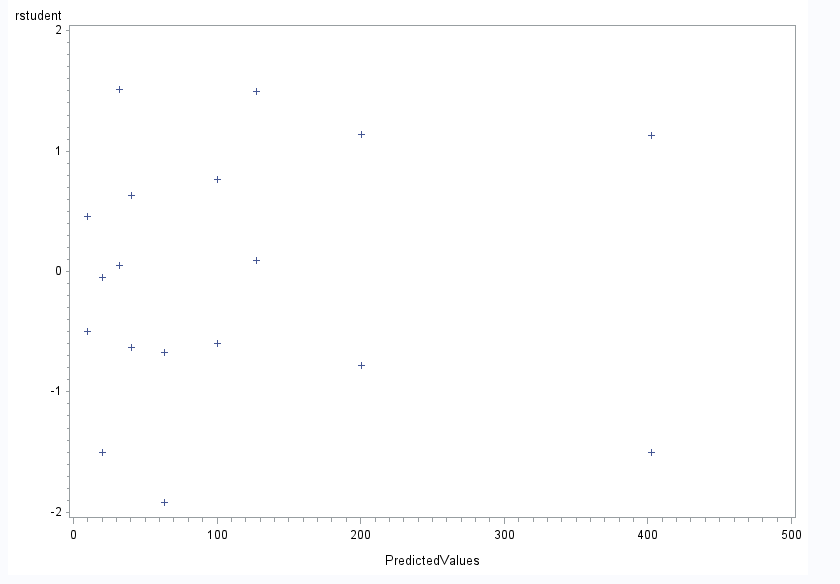
Therefore, the potential starting values are , , and .

1. The Newton-Raphson and Gauss-Newton algorithms produced the same estimates of the parameters: , , and . However, the Newton-Raphson algorithm took four iterations to converge while the Gauss-Newton only took three. The steepest descent algorithm never converged and therefore did not provide estimates for the parameters.
2. The parameter estimates are “close-to-linear” since the Hougaard measure of skewness is less than 0.1 for each parameter. The Hougaard measure is 0.07, 0.0587, 0.0325 for , , and respectively.
3. The residual plots, shown at the end of this problem, appear to be scattered in an independent and random manner. The normal probability plot of the residuals also looks good, and gives evidence to support that the residuals are normally distributed.
4. The reduced model for this test sets . The full model does not have constraints on and . The error sum of squares is calculated for the full and reduced and used in the following formula to find the F test statistic.

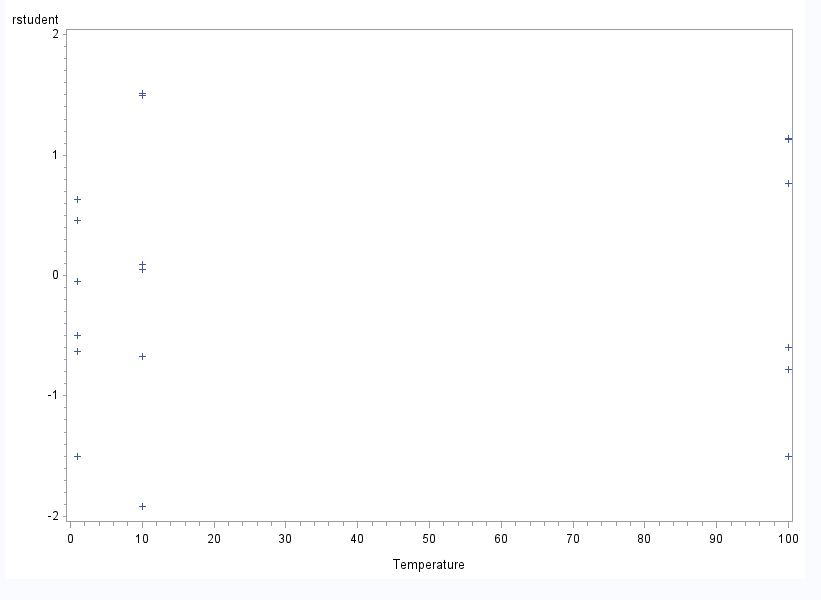
There is sufficient evidence at the 5% level to conclude that

1. The 95% family confidence interval for and using the Bonferroni procedure is , and .
2. and are significantly different values.

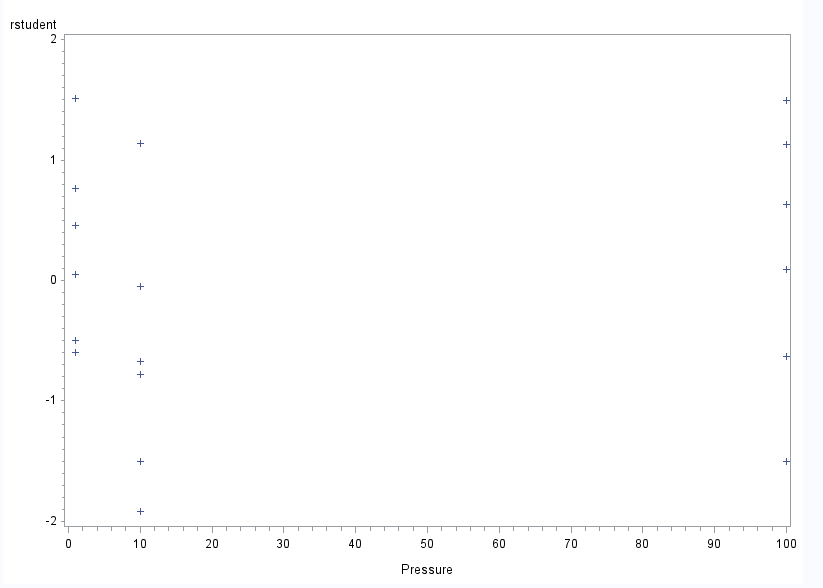
**Figure 1: Plot of Residuals against Predicted Values**



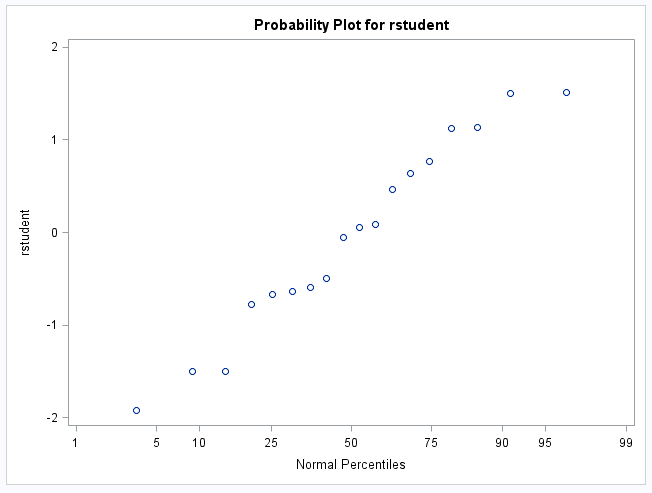
**Figure 2: Plot of Residuals against Temperature**



**Figure 3: Plot of Residuals against Pressure**



**Figure 4: Normal Probability Plot of Residuals**



Problem 2

The two models that will be fit to the data are as follows.

The Linear Model:

The Quadratic Model:

The starting values for the linear model were found by finding the parameter estimates of the linear regression , which resulted in and . The starting values for the quadratic model were found by finding the parameter estimates of the linear regression , which resulted in and , and . These starting values were then used in conjunction with the Gauss-Newton Algorithm to find estimates for the parameters in the above models. The results were the following estimated models.

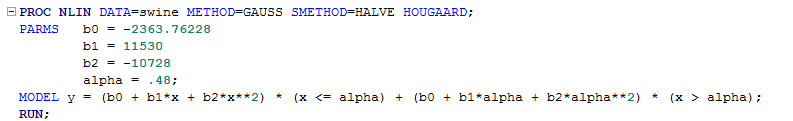
The Linear Model:

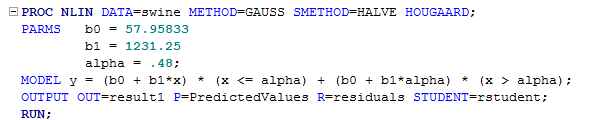
The Quadratic Model:

To test whether the quadratic model is more appropriate than the linear model, we test against . If the data supports the null hypothesis, than the quadratic parameter is not significant and the linear model is a more appropriate model. The test uses the F test statistic with the reduced model being the linear model, and the full model being the quadratic model.

The p-value is quite high and the conclusion is that , indicating that the linear model is a more appropriate fit for the data. The plot of residuals against predicted and x values supports the appropriateness of the linear model, showing no concerning patterns.

Please find the code that was used to create the quadratic and linear model respectively.





Problem 3

The following model assumes the test and reference vaccines have different responses.

We will use this model as the full model when we test hypotheses. is plotted against the responses for the test and reference vaccines. The test and reference vaccine plots show very little difference in shape. As such, the same starting values will be used for both , , and and , , and . The starting value for a is set to 0.5, which is the approximate upper limit of the data. We will do a grid search for b using the values 0.5, 1.0, and 1.5. These values were chosen since the rate of change changes significantly as b goes from greater than 1 to less than 1. The starting value for will be 4.5, the median of the .

Using the above starting values, , , and the Gauss-Newton method is applied and the following model is estimated.

The above model is the full model and has an and degrees of freedom equal to 26. The null hypothesis, that the response for the test vaccine is different than the reference vaccine, is tested using the reduced model given below.

Reduced Model ->

The reduced model has an and degrees of freedom equal to 29. The F test statistic for the hypothesis that the two vaccines have different responses is as follows.

The conclusion is therefore that the two vaccines do have a different response at a 5% significance level.

We then test how the two responses are different by testing the null hypotheses that , , and . The SSE of the reduced model that corresponds to the each null hypothesis is found and the F test statistic is calculated using the SSE of the full model, which was given above. The F-statistics and p-values are given below for each test.

F-test for

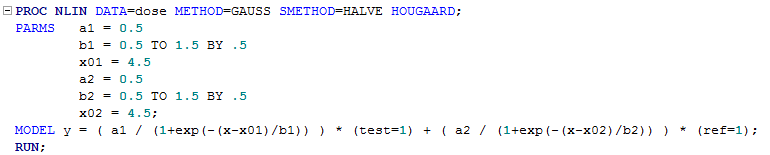
F-test for

F-test for

The conclusion of all the F-tests is that there is no difference between like parameters in the two response models. These conclusions are also reached if the t-test is applied to the difference between like parameters of the two models.

The responses between the test vaccine and reference vaccine are statistically different, however the like parameters between the two responses are not statistically different. This paradoxical conclusion is likely because there are high dependencies between the parameters, though the F-test takes into consideration dependencies between parameters. Evidence for this explanation is given by the high correlations between parameters in the approximate correlation matrix.

Please find the snippet of SAS code below to show how the original full model was fit.



Problem 4

The Gompertz growth model can be used to model the House Finch population since the species birth rates should first increase and then slow as resource limits are reached, an attribute of the Gompertz model. The Gompertz model is given as follows.

Variable a is the asymptote, b sets the displacement along the x-axis, and c sets the growth rate. The variable t sets a year as time zero. The starting value for a can be approximated based on the graph. The data appears to be leveling out at around 1120. Therefore, the starting value for is 11250. If we set t equal to 1966, we can plug in the data point (1966, 8) and solve for b, which will give us If we plug in another data point (1980, 614), we can solve for c, which will approximately be .05. These will act as are starting values.

Using the Gauss-Newton Method, the estimated Gompertz model is as follows.

The residuals for this model appear to show some evidence of dependency, which is concerning. The normal probability plot also appears to show evidence of non-normalcy. The Shapiro-Wilk test gives a p-value of 0.0050, which supports the alternative hypothesis of non-normality. The Hougaard measures for the parameters all indicate that the parameter estimates are close to linear, or reasonably close to linear.

When I attempt to re-parameterize the response variable as log(y), the residuals and normality plots do not improve. There is something that is effecting linearity, but I am unsure as to the solution.

Visually, the model does well at predicting the data. However, the model has an asymptote of 20906.3. The data seems to start to level around 11250. I would prefer if the model could be constrained so that it has an asymptote that matches the data. Otherwise, I’m happy with the model.