UNIT 11 HW

1. From Problem 26, Chapter 8:

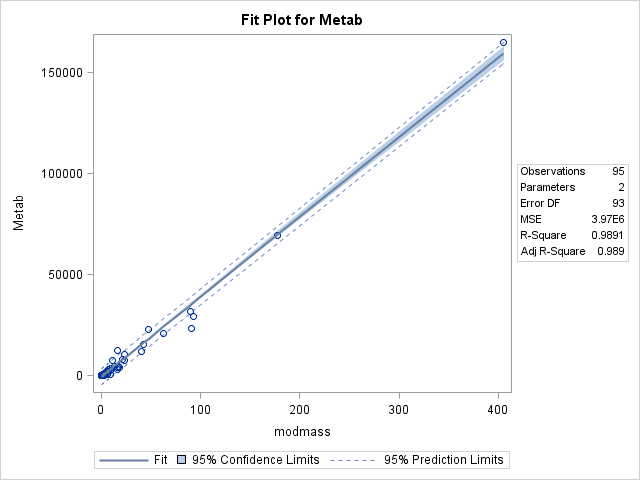
The Metabolic data set has the average mass, metabolic rate, and average lifespan of 95 different species of mammals. Kleiber’s Law states that the metabolic rate of an animal species, on average, is proportional to its mass raised to the power ¾. Judge the adequacy of this theory with these data. Ultimately, for this problem, we want to find the best model. (At this point, you will limit the analysis to the two variables under study, though the data set has more variables.) In the current data set, assume that mass has not yet been raised to the power ¾.

* Use alpha = 0.05.
* Use **SAS** for this problem.
* Include **relevant** code and output. Make sure you directly answer the questions. Do NOT assume the answer is obvious from the output.

Specifically, provide/answer the following:

* + 1. Judging by a scatterplot alone, does it seem reasonable that the metabolic rate of an animal species, on average, is proportional to its mass raised to the power of ¾? (Recall that if some variable y is proportional to the variable x, then (with nonzero m) is a well-fitting model.) In other words, does the data (metabolic rate, mass3/4) reasonably fall along a straight (nonhorizontal) line and nearly pass through the origin?

Based on the fit model of the mass^¾ and metabolism, it appears to show an R-squared of 0.9891, which would appear to be an excellent fit. In other words, the data would indicate that the model accounts for the 98% of the variance in the metabolism and mass, which is an indication that there is a linear relationship between mass and metabolism.



* + 1. We want to find the “best” model to predict metabolic rate from mass3/4 **and** make appropriate statistical inferences. Therefore, address all the assumptions prior to the analysis (using mass3/4). If the assumptions are not met, handle the data appropriately. If a transformation is used to satisfy the assumptions, address the assumptions again to ensure that the transformation is logical, and carry out your analysis on your newly transformed data. For example, you should include a scatter plot for the original data AND transformed data, etc. (Hint: if a transformation is necessary, try one of the transformations discussed in class first.) Either way, keep the “mass3/4” in the model; do not go back to regular “mass,” although mass3/4 may be transformed if it makes sense for the assumptions. At minimum, provide and interpret the following elements to address assumptions FOR THE ORIGINAL DATA AND ANY TRANSFORMED DATA (IF you use a transformation). You may include more graphs if you find them useful.

1. A scatterplot with the following included on the graph: regression line, confidence intervals of the regression line, and prediction intervals of the regression line.

ii. A scatterplot of residuals.

iii. A histogram of residuals with the normal distribution superimposed.

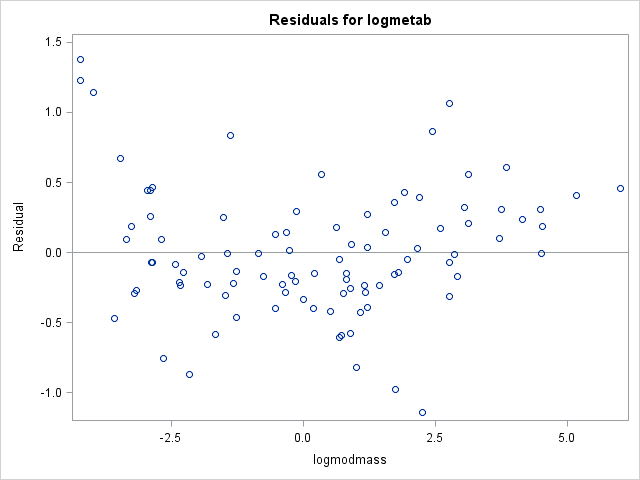
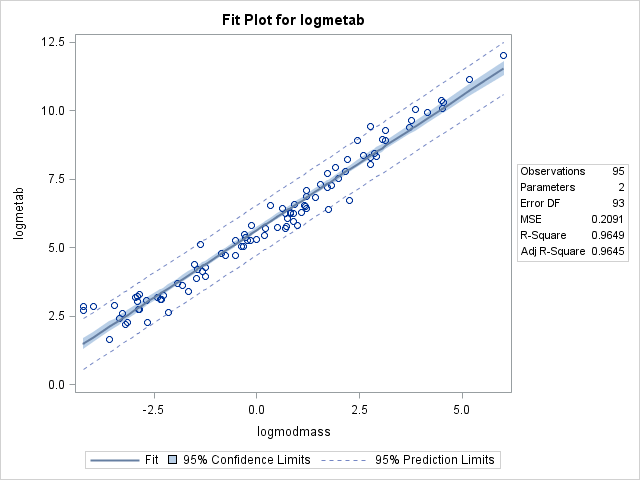
iv. A discussion supporting the use of the model you chose (support that the assumptions are met).

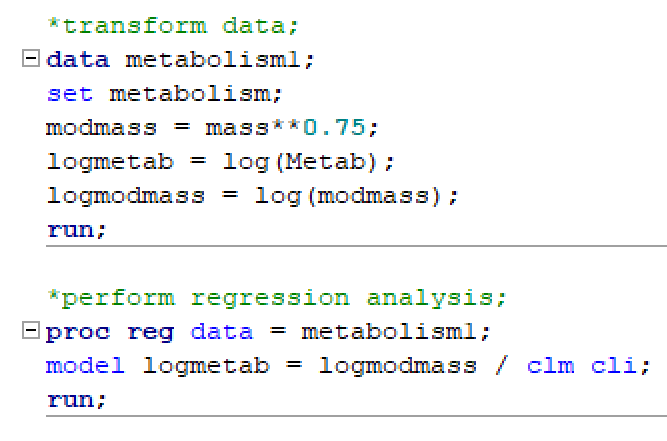
Linearity – based on the scatterplot, it there is insufficient evidence to suggest the relationship is not linear between mass3/4 and metabolism.

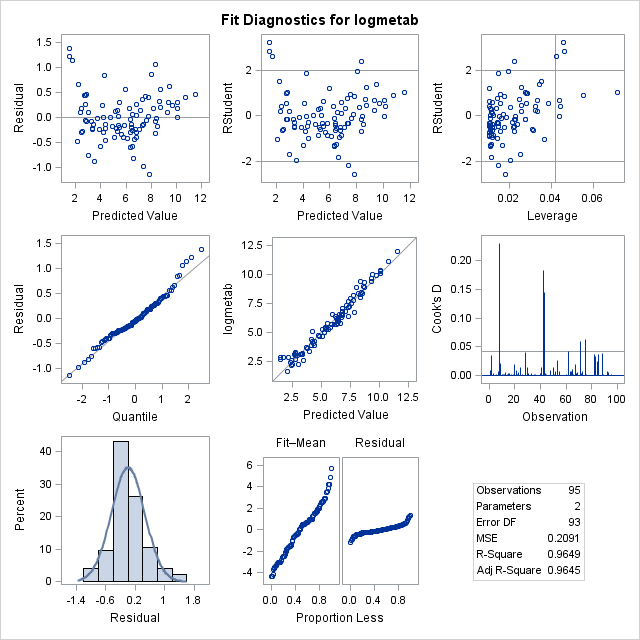
Equal-Spread – Based on the residuals plot, we can see the data does not have equal variance. In fact, as the mass increases, so does the variance in the data.

Normality – Furthermore, the residuals indicate the data is left skewed and this could probably be attributed to what appear to be outliers in the data set.

Independence – We do not know much about the sampling methods of the study itself, so we will assume the data is in fact independent.

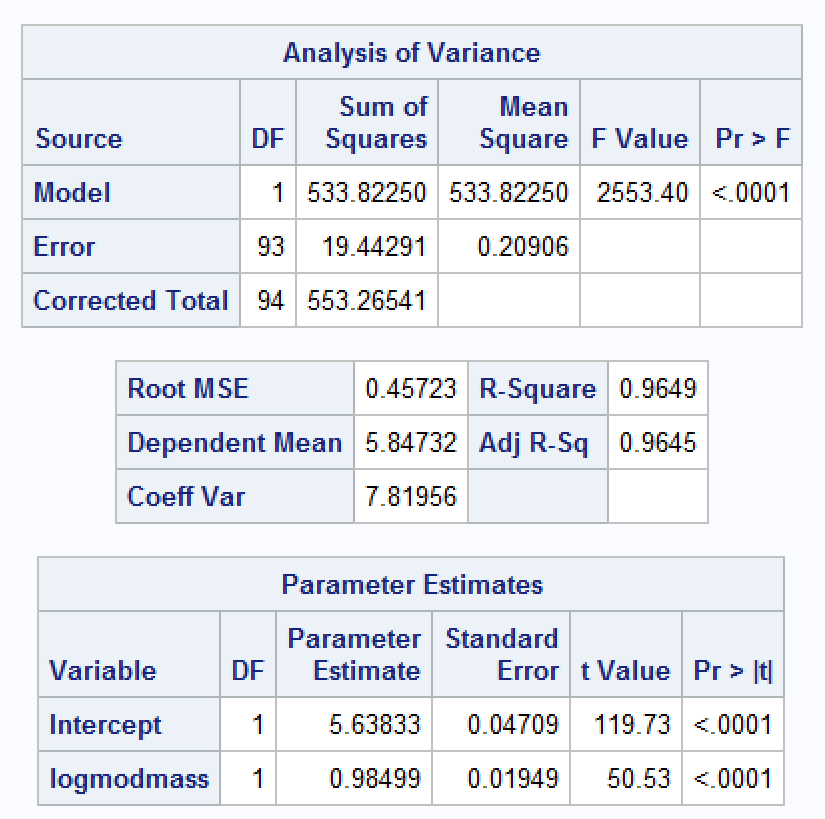






The log-log model was preferred over other models because the model handled the lack of normality and outliers exceptionally well. A closer look into the residuals plot clearly shows the model following the linear trajectory, with slight variance. Thanks to this model, the assumptions are now met.

* + 1. Once a reasonable model is found (possibly using a transformation), provide a table showing the t-statistics and p-values for the significance of the regression parameters .



* + 1. The estimated regression equation. Make sure the dependent variable is noted as the predicted value or predicted mean value, not just the dependent variable.
    2. Interpretation of the model, paying special attention if you used a transformation (hint!). That is, interpret the slope as well as the **confidence interval**.

A doubling of Mass3/4 equates to the multiplicative change of 20.98488 = 1.97 in the median of the distribution of the metabolism rate for the given mass3/4. In other words, a doubling of Mass3/4 increases the estimated median of the metabolism rate by 97.9%. Therefore, a 95% confidence interval for the multiplicative increase in median of metabolism rate after doubling the mass3/4 is between 93.7%, 103.3%.

* + 1. A measure of the proportion of variation in the response that is accounted for by the explanatory variable. Interpret this measure clearly.

It is estimated that the Mass3/4 estimates about 96.5% of the variation in the Metabolic Rate.

1. From Problem 29, Chapter 8:

The autism data show the prevalence of autism per 10,000 ten-year-old children in the United States in each of five years. Analyze the data to describe the change in the distribution of autism prevalence per year during this time period.

* Use alpha = 0.05.
* Use **R** for this problem.
* Include **relevant** code and output. Make sure you directly answer the questions. Do NOT assume the answer is obvious in the output.

Specifically, provide/answer the following:

* + 1. Address all the assumptions for a linear regression model prior to the analysis. If the assumptions are not met, handle the data appropriately. If a transformation is used, address the assumptions again with the transformed data to ensure that the transformation is logical. The questions below should reflect this. For example, you should include a scatter plot for the original data AND transformed data, etc. (Hint: if a transformation is necessary, try one of the transformations discussed in class first.) At minimum, provide and interpret the following elements to address assumptions FOR THE ORIGINAL DATA AND ANY TRANSFORMED DATA (IF you use a transformation). You may include more graphs if you find them useful.

1. A scatterplot with the following included on the graph: regression line, confidence intervals of the regression line, and prediction intervals of the regression line.

ii. A scatterplot of residuals.

iii. A histogram of residuals with the normal distribution superimposed.

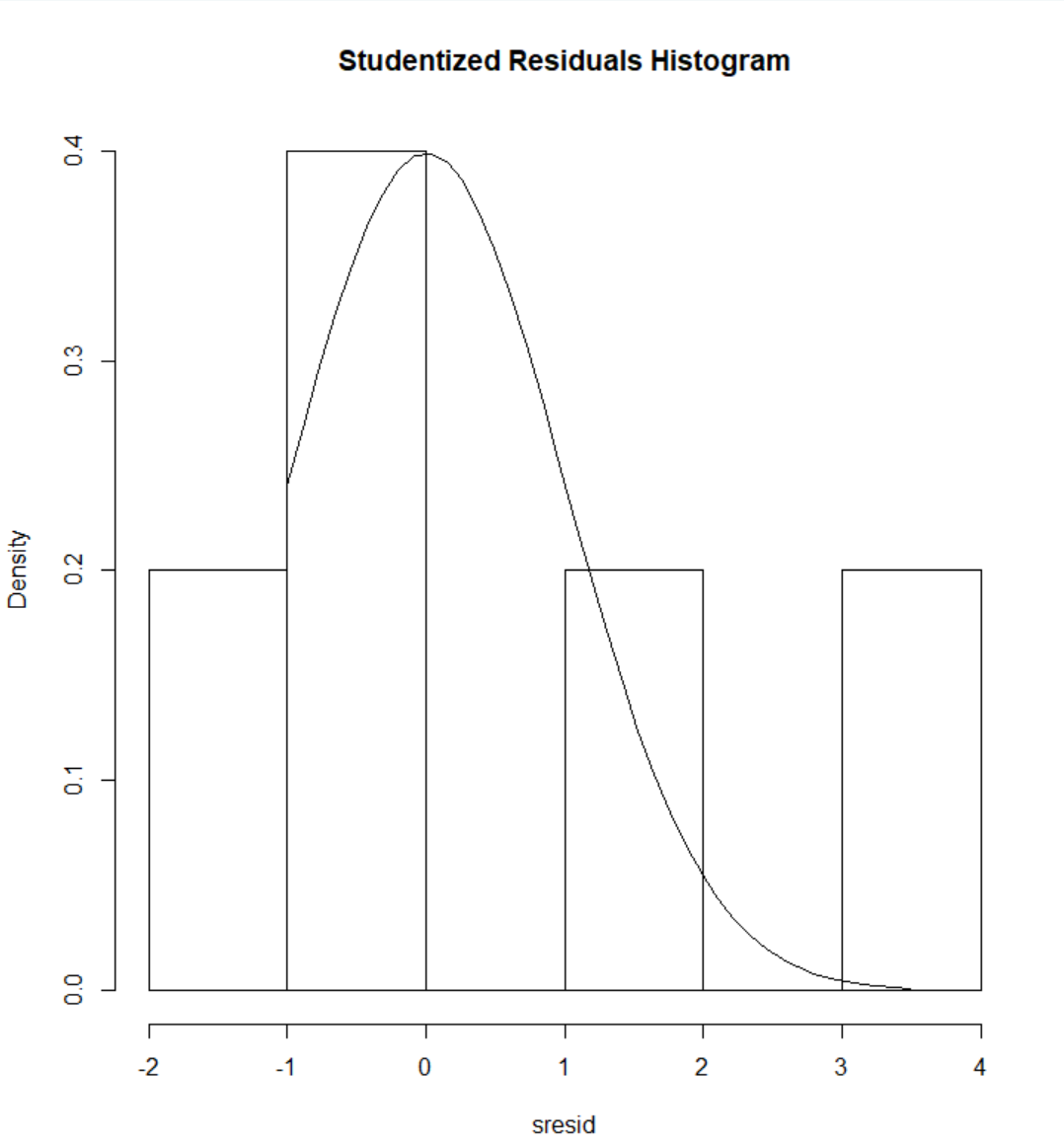
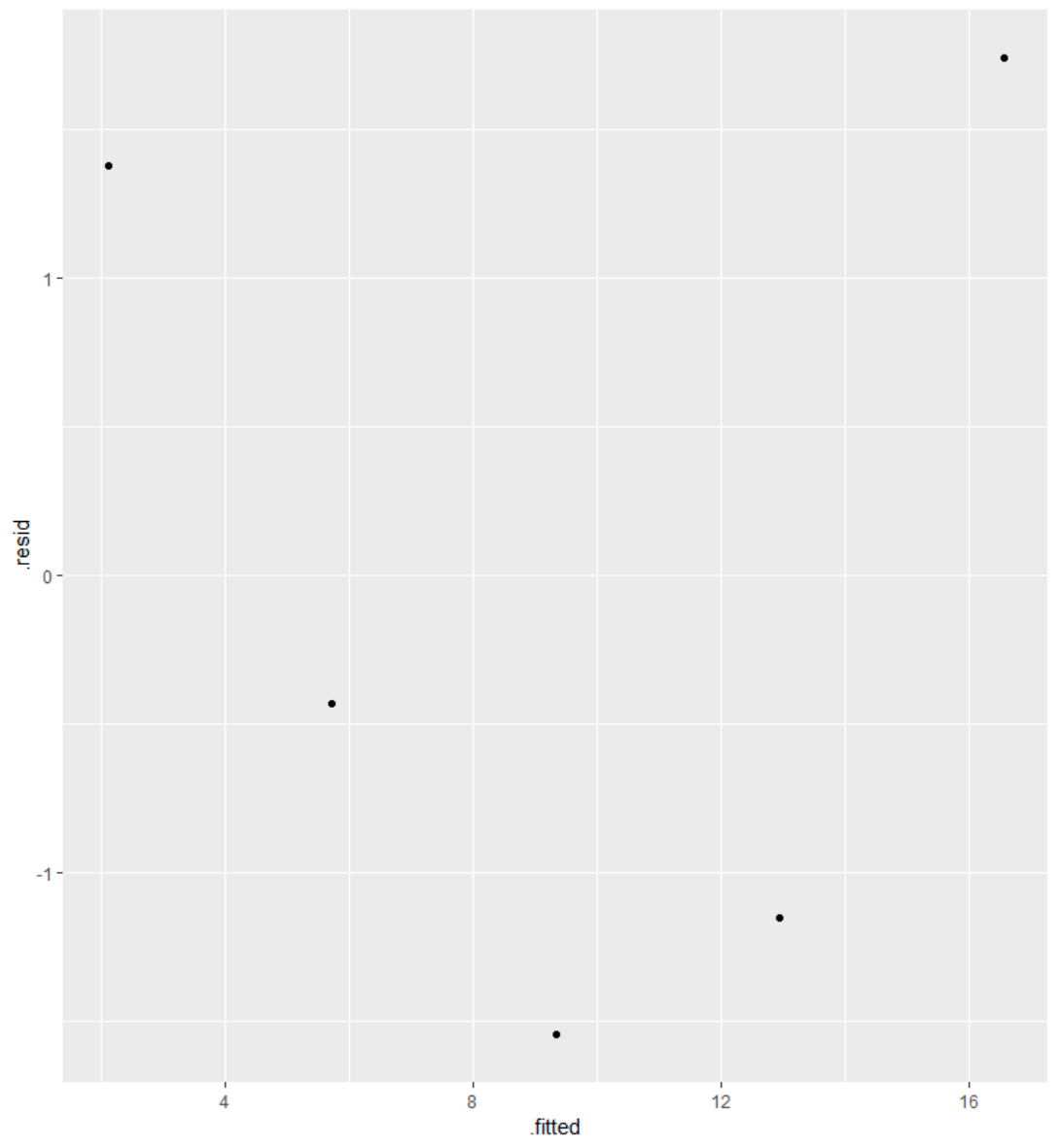
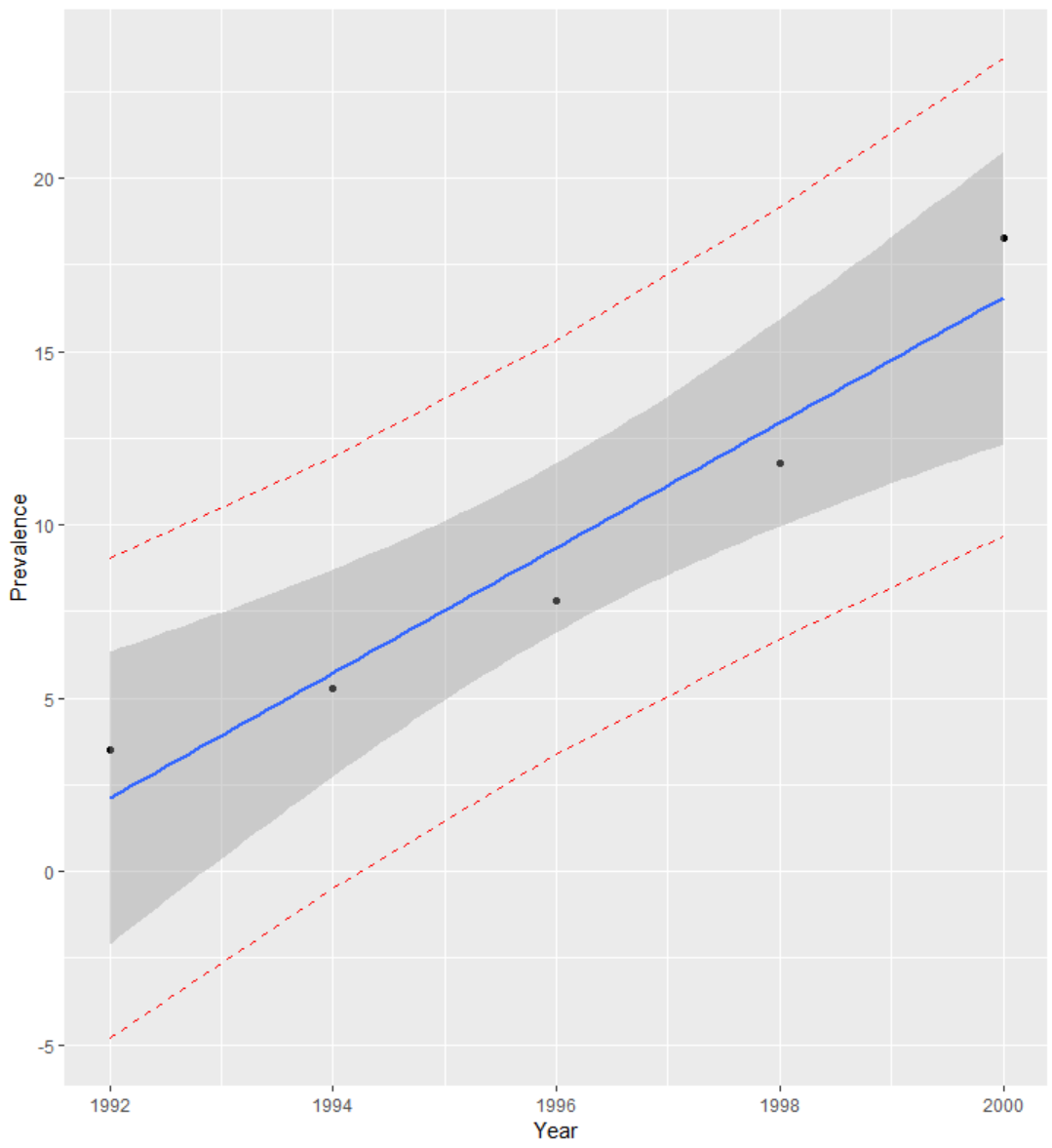
iv. A discussion supporting the use of the model you chose (support that the assumptions are met).

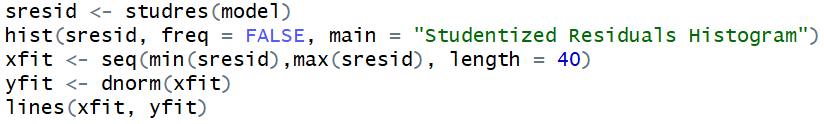
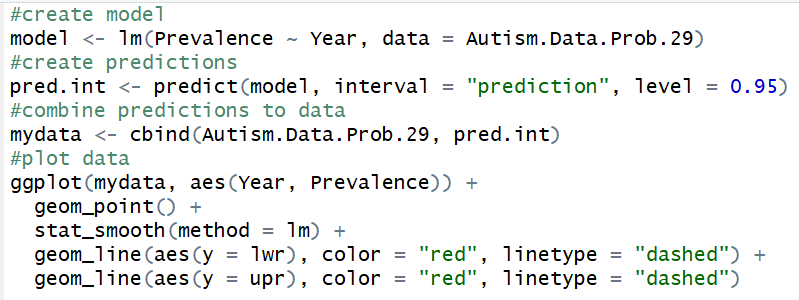
Linearity – The model suggests there is evidence of a non-linear relationship in the data. The curve in the data set doesn’t match up to the linear model as much as it would to a non-linear model.

Equal-Spread – The residuals scatterplot magnifies the lack of evidence in favor of a linear relationship. In fact, it appears to indicate the relationship may be more parabolic in nature.

Normality – The residuals histogram clearly demonstrates the data is right-skewed.

Independence – Without any additional knowledge about the sampling methods, we are unable to truly determine if the data is independent. There is concern for the year element in the data and whether the samples are repeats of prior years, etc. However, we will proceed as if data is independent.



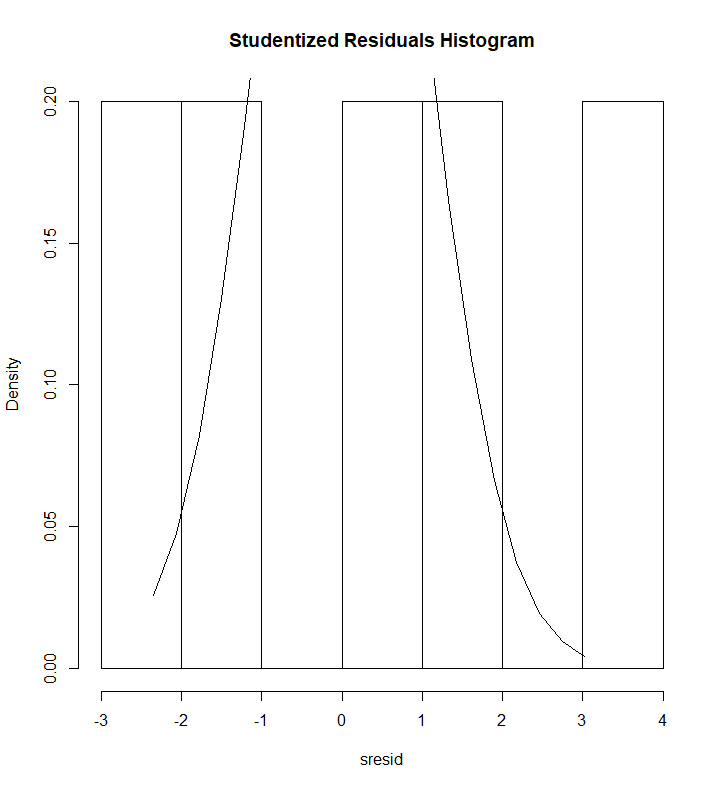
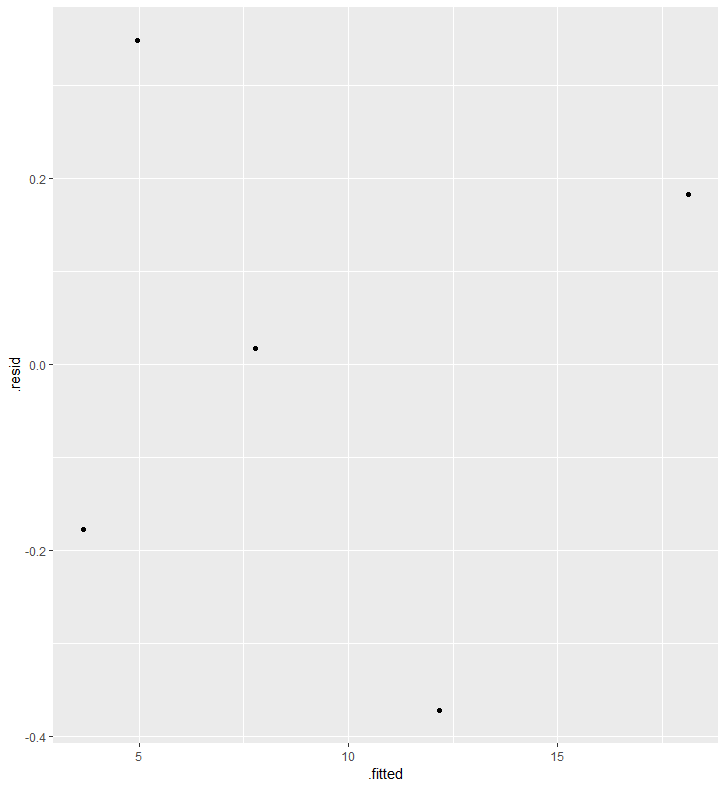
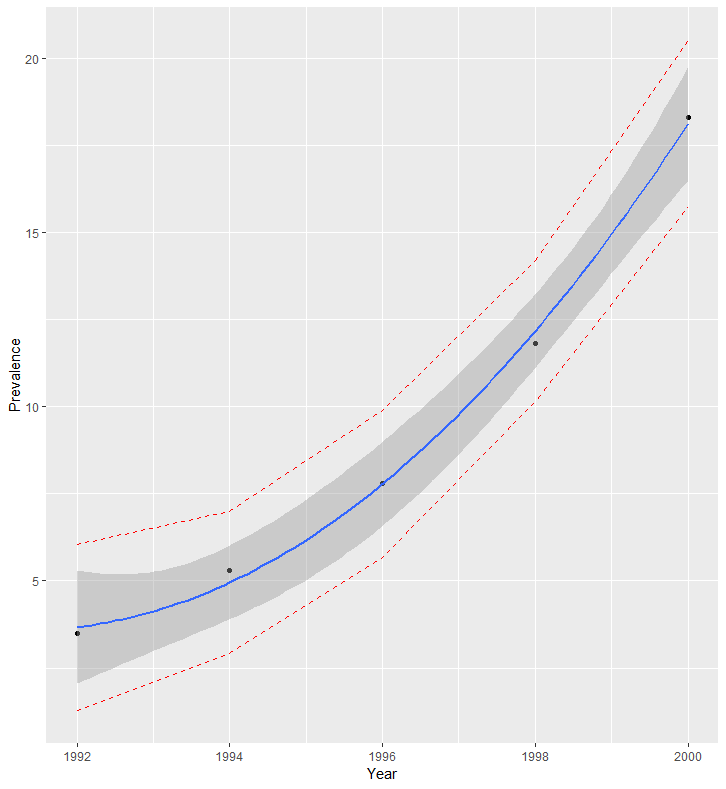


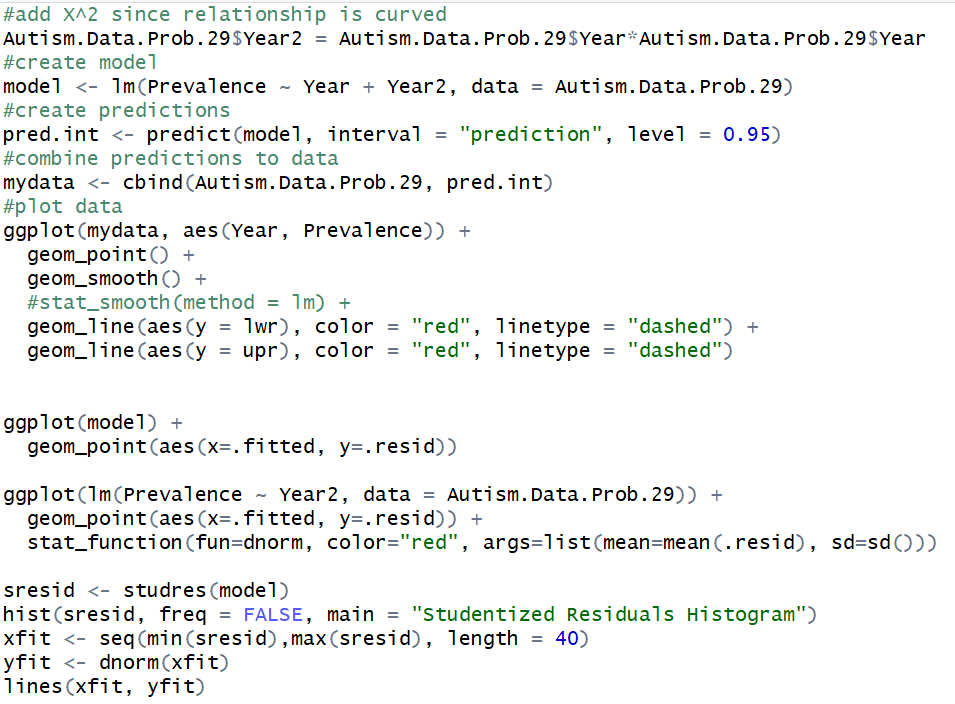
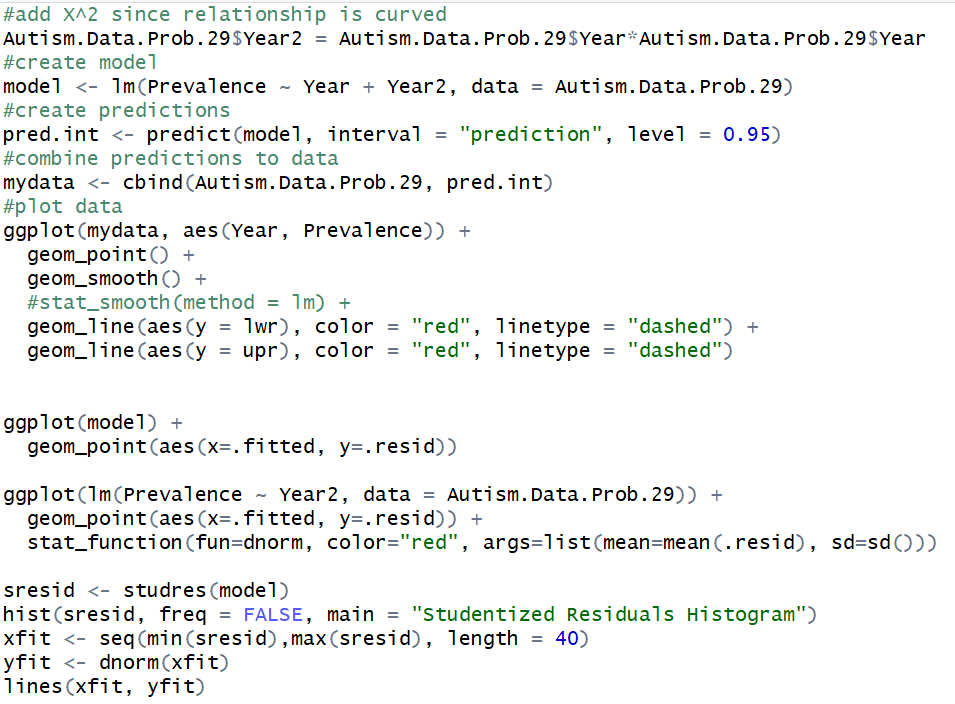
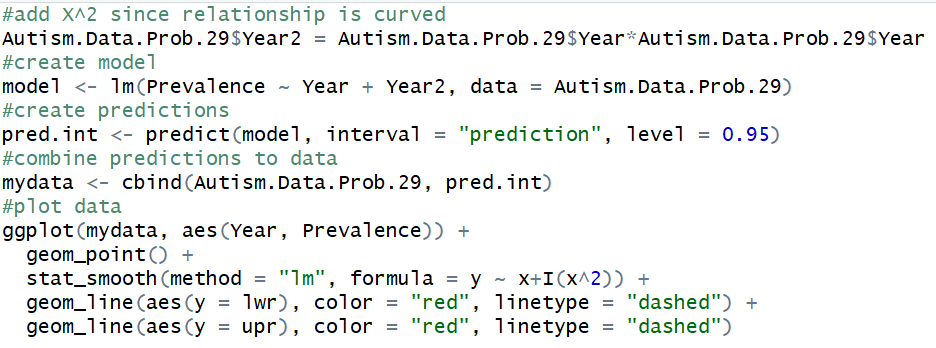
Linearity – The model fits into a quadratic model, which as illustrated in the plot below, fits much better than the standard linear model.

Equal-Spread – The residuals scatterplot is now properly displaying what appear to be randomized results. Overall, it looks like the residuals are a proper spread.

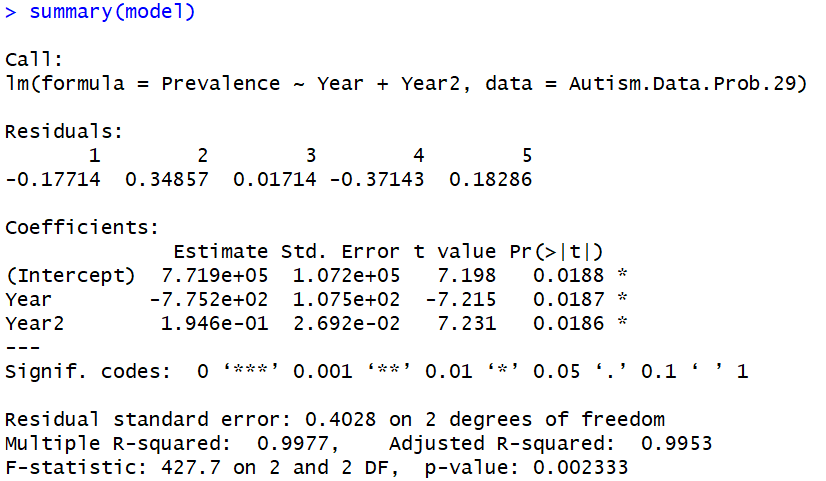
Normality – The quadratic model appears to have addressed the lack of normality in the residuals.

Independence – Without any additional knowledge about the sampling methods, we are unable to truly determine if the data is independent. There is concern for the year element in the data and whether the samples are repeats of prior years, etc. However, we will proceed as if data is independent.



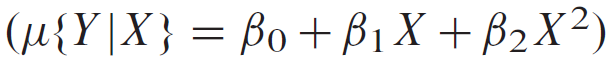


* 1. Once a reasonable model is found (possibly using a transformation), provide a table showing the t-statistics and p-values for the significance of the regression parameters .





* 1. The estimate regression equation. Make sure the dependent variable is noted as the predicted value or predicted mean value, not just the dependent variable.



* 1. Interpretation of the model, paying special attention if you used a transformation (hint!). That is, interpret the slope as well as the **confidence interval**.

The model indicates that as the time goes by, the prevalence of autism increases slowly. Yet, as time progresses, prevalence of autism begins to increase exponentially.

* 1. A measure of the proportion of variation in the response that is accounted for by the explanatory variable. Interpret this measure clearly.

It is estimated that Year and Year2 account for 99% of the Prevalence variance.

1. Bonus!

Consider the steer data in Display 7.3 on page 179 (Chapter 7) of the textbook (third edition). Perform a lack of fit test comparing the regression model and a separate means model. Because we have at least two points in at least one group (replication to estimate the variance), we can perform ANOVA. (ANOVA does not make sense if no values of the independent variable are repeated.) During live session, we already addressed the assumptions and determined that a linear-log model is best for regression. Perform this lack of fit test (all parts) on the transformed data. Use the software of your choice. Specifically, include the following:

* 1. Hypotheses
  2. The ANOVA table you created
  3. Decision

d. Conclusion in nonstatistical terms

e. Code and relevant output