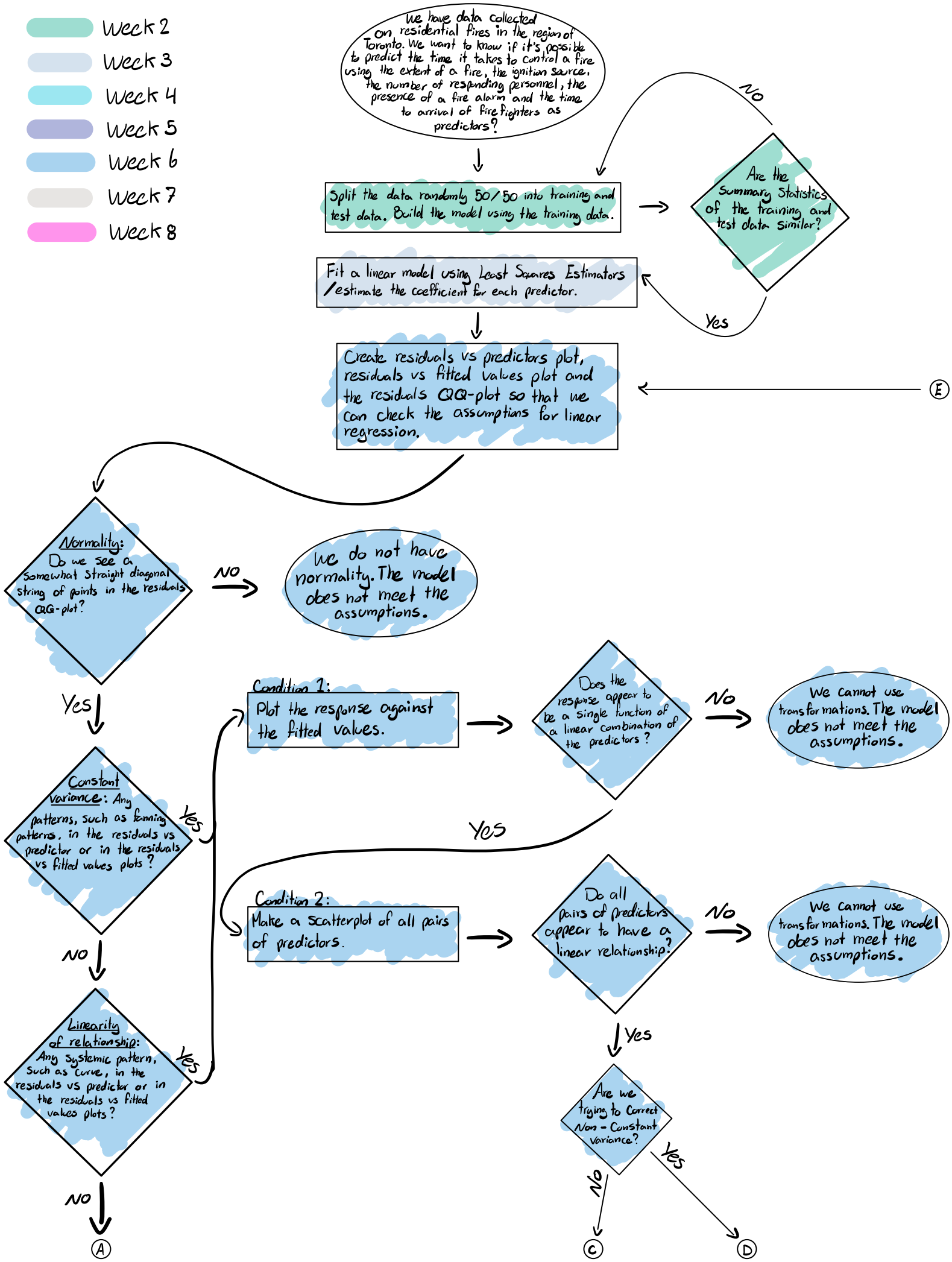
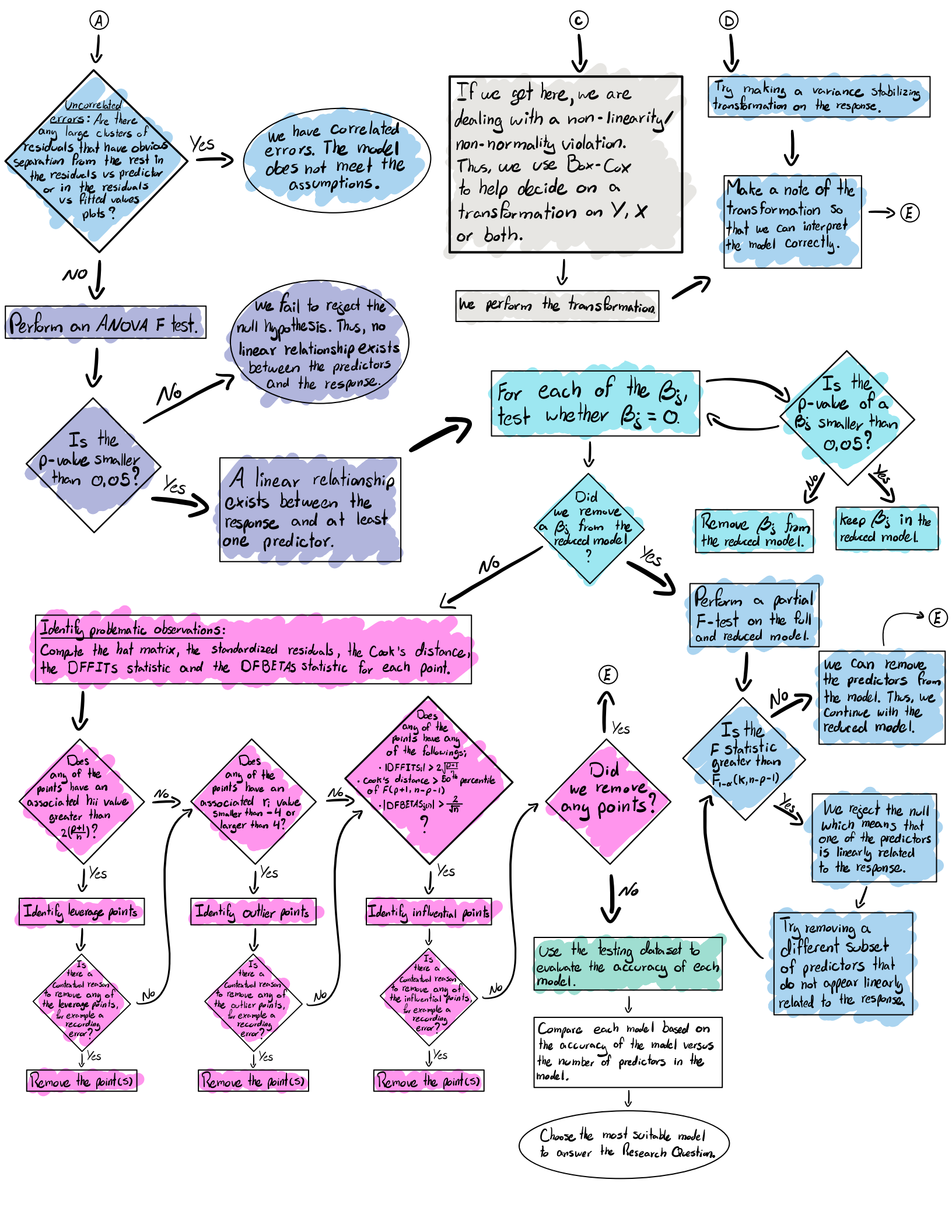

Week 2
Week 3
Week 4
Week 5
Week 6
Week 7
Week 8
We have data collected on residential fires in the region of Toronto. We want to know if it's possible to predict the time it takes to control a fire using the extent of a fire, the ignition source, the number of responding personnel, the presence of a fire alarm and the time to arrival of firefighters as predictors?
Split the data randomly 50/50 into training and test data. Build the model using the training data.
Are the summary statistics of the training and test data similar?
No
Yes
Fit a linear model using Least Squares Estimators /estimate the coefficient for each predictor.
Create residuals vs predictors plot, residuals vs fitted values plot and the residuals QQ-plot so that we can check the assumptions for linear regression.
E
Normality: Do we see a somewhat straight diagonal string of points in the residuals QQ-plot?
NO
We do not have normality. The model does not meet the assumptions.
Yes
Constant variance: Any patterns, such as fanning patterns, in the residuals vs predictor or in the residuals vs fitted values plots?
Yes
NO
Linearity of relationship: Any systemic pattern, such as curve, in the residuals vs predictor or in the residuals vs fitted values plots?
Yes
NO
A
Condition 1:
Plot the response against the fitted values.
Does the response appear to be a single function of a linear combination of the predictors?
No
We cannot use transformations. The model does not meet the assumptions.
Yes
Condition 2:
Make a scatterplot of all pairs of predictors.
Do all pairs of predictors appear to have a linear relationship?
No
We cannot use transformations. The model does not meet the assumptions.
Yes
Are we trying to correct Non-Constant variance?
No
Yes
C
D


A
Uncorrelated errors: Are there any large clusters of residuals that have obvious separation from the rest in the residuals vs predictor or in the residuals vs fitted values plots?
Yes
We have correlated errors. The model does not meet the assumptions.
No
Perform an ANOVA F test.
Is the p-value smaller than 0.05?
No
We fail to reject the null hypothesis. Thus, no linear relationship exists between the predictors and the response.
Yes
A linear relationship exists between the response and at least one predictor.
For each of the β_j, test whether $\beta_j = 0$.
Is the p-value of a β_j smaller than 0.05?
No
Remove β_j from the reduced model.
Yes
Keep β_j in the reduced model.
Did we remove a β_j from the reduced model?
Yes
Perform a partial F-test on the full and reduced model.
Is the F statistic greater than $F_{1-\alpha}(k, n-p-1)$
No
We can remove the predictors from the model. Thus, we continue with the reduced model.
E
Yes
We reject the null which means that one of the predictors is linearly related to the response.
Try removing a different subset of predictors that do not appear linearly related to the response.
No
Identify problematic observations:
Compute the hat matrix, the standardized residuals, the Cook's distance, the DFFITS statistic and the DFBETAS statistic for each point.
Does any of the points have an associated hii value greater than $2(\frac{p+1}{n})$?
Yes
Identify leverage points
Is there a contextual reason to remove any of the leverage points, for example a recording error?
No
Yes
Remove the point(s)
Does any of the points have an associated r_i value smaller than −4 or larger than 4?
Yes
Identify outlier points
Is there a contextual reason to remove any of the outlier points, for example a recording error?
No
Yes
Remove the point(s)
Does any of the points have any of the following:
· $|DFFITS_i| > 2\sqrt{\frac{p+1}{n}}$
· Cook's distance > 50th percentile of $F(p+1, n-p-1)$
· $|DFBETAS_{j(i)}| > \frac{2}{\sqrt{n}}$
?
Yes
Identify influential points
Is there a contextual reason to remove any of the influential points, for example a recording error?
No
Yes
Remove the point(s)
Did we remove any points?
Yes
E
No
Use the testing dataset to evaluate the accuracy of each model.
Compare each model based on the accuracy of the model versus the number of predictors in the model.
Choose the most suitable model to answer the Research Question.
C
If we get here, we are dealing with a non-linearity/non-normality violation. Thus, we use Box-Cox to help decide on a transformation on Y, X or both.
We perform the transformation.
D
Try making a variance stabilizing transformation on the response.
Make a note of the transformation so that we can interpret the model correctly.
E