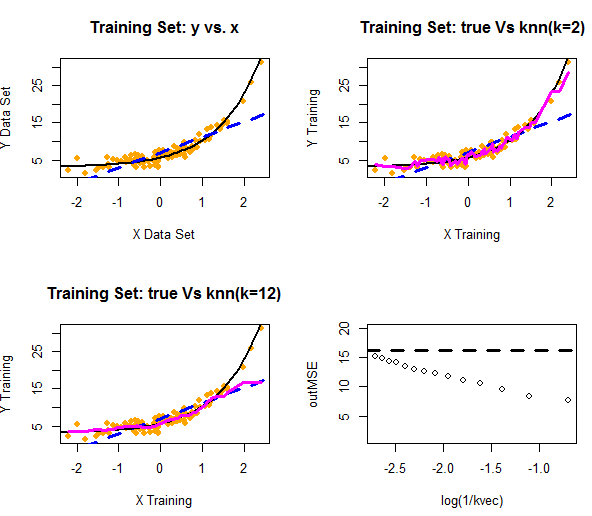


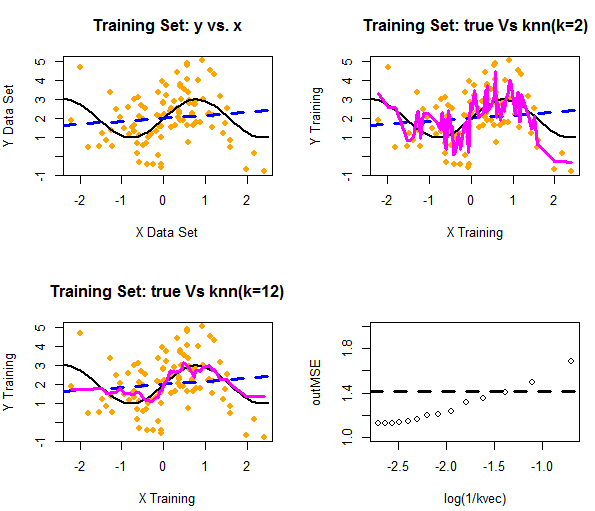
Question #5

The optimal result is to utilize the regression. The logistic regression function was far less tuned to the variability in the model (while there is essentially no variability, this results in significant bias). Nonetheless, the logistic regression’s lack of bias made it more consistent across board (though especially at the tail ends of our data set). The K-means function, in return, appeared to capture far more of the bias resulting in poorer model performance upon tuning. While the optimal k-result is k=10, the logistic regression was still a superior model overall. Anecdotally this makes sense, as the structure of a logistic regression is nearly identical to the structure of the question upon presentation.



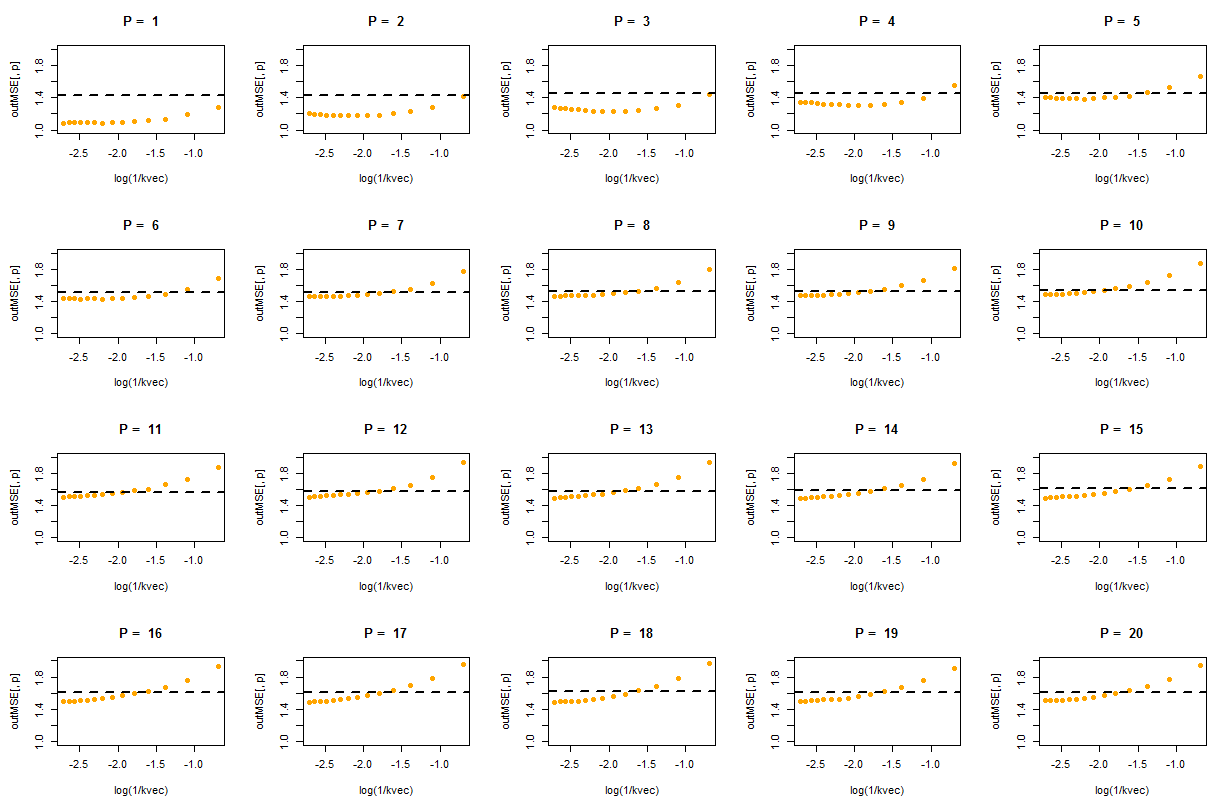
Question #6

Simulating data from a model that added an exponential variable resulted in the k-means model being far superior to that of logistic regression. The average MSE of a straight line fitted to an exponential curve is predictably poor. While the k-means methodology performed similar to the logistic regression at the tail-end of one part of our model, it was far superior over the remainder of the data, with nearly half the error term of the regression line at some points. In this instance, the optimal value was k=2.



Question #7

The addition of a trigonometric function to our model creates a mixed result due to the variability of the output. Logistic regression may appear to be more accurate in some facets as the *average* least squares error would be smaller than otherwise considered. However, the MSE of the logistic regression function is far higher than K-means function at nearly all data points. The optimal k, as a result, is 13, and the k-means model is far superior even at most p-values.



Question 8

The result of this simulation provides that the higher the noise, the more difficult the k-means algorithm has in predicting subsequent results. Because the additional noise in the model introduces variance (and because k-means operates more effectively in clustered examples than the logistic regression), this caused problems with the k-means fitting model. This is why, in the chart above, as p increases the errors associated with k-means increases, and leads to logistic regression’s model superiority.

In other words, while logistic regression is optimal for relationships with higher variability and lower bias, k-means is optimal for functions with higher bias and lower variability. As we force the K-means model to require additional clusters, we over-fit. This results in logistic regression being optimal at some points in lieu of k-means clustering.

BONUS

An increase in the number of observations may have an impact on some of the results. Most notably in the final question that bifurcated the efficacy of the models. In models with higher training samples, this increase in observations would reduce dimensionality. The reduction in dimensionality would likely lead to more optimal results for the k-means algorithm, particularly for non-parametric models. As dimensionality reduces, bias decreases due to more accurate clusters and overfitting is reduced. As a result, the efficacy of the -means algorithm increases. In general, larger data sets allow for more fine-tuning without overfitting, allowing us to utilize models that may have higher incidents of bias.