**01:960:486:01- Final exam**

**Himesh Buch**

**hnb17**

**5/7/2020**

**Solution of questions: 1, 2, 4, 5, 7, 8, 9, 10, 11, 12, 13 (solutions of 3 and 6 in jpg image format)**

**1.**

a) In backward elimination, we choose the variables with the lowest R squared adj to be removed first. Now, looking at the table,

1. X5
2. X1
3. X4
4. X2
5. X3

The variables should be removed in that order

b) The highest R squared adj should be added first. Now,

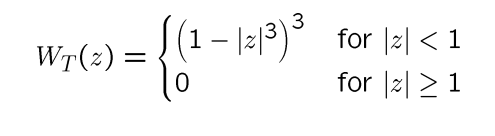
X3 -> X2 -> X4 -> X1 -> x5

Should be the order

c) The first variable with highest R squared to be added is X3. Now we look at only the variables that contain X3 and has the highest R squared adj, which is X3 X2. The next one to be added is, X5 X3 X2, after that, X5 X4 X3 X2, and the final will be, X5 X4 X3 X2 X1. We just have to make sure that the next model that we choose needs to have the variables from the previous model. Our final model is, X5 X4 X3 X2 X1

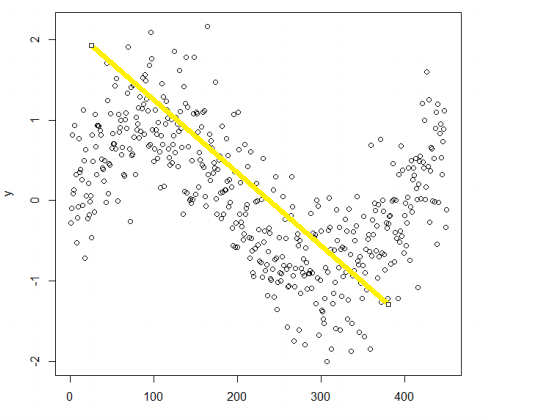
**2.**

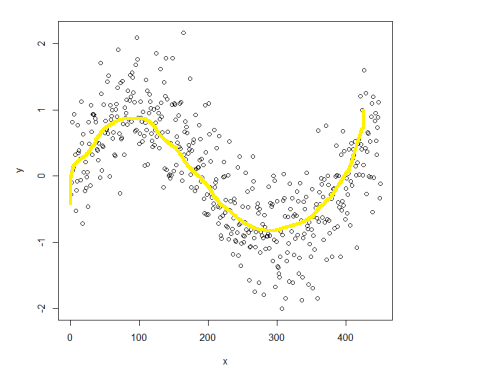
* 1. The formula is,

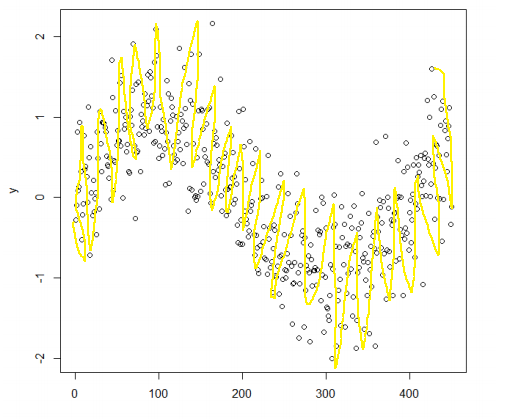


Using the above formula, Z = (xi – x0)/hi, in our case x10 = -0.1 and x0 = x9 = -0.2, and h = 0.013. Hence,

W(z) = (1-|((-0.1) – (-0.2))/0.013|­3) 3 for |z| < 1, and 0 for |z| >= 1







1. For n datapoints, 2n regressions can be performed.

**4**.

The analysis of small data sets in longitudinal studies can lead to power issues and often suffers from biased parameter values. These issues can be solved by using Bayesian estimation. Bayesian methods will be very useful on smaller datasets only if you have some extra knowledge of what your regression coefficients could be. This knowledge can be then incorporated into prior distributions with small variances for your regression coefficients. The posterior distributions of the regression coefficients may then reflect less uncertainty than confidence intervals of your linear regression (OLS) estimate of the same parameters. In those situations, Bayes regression is most likely a better option and likely to perform better than OLS

**5**.

a) Model selection can have multiple advantages like it can really help in finding the best model for any given data. It can either add or remove (depending on the chosen method) data and sort of, filter out only the useful data points or predictors which then can help in selecting the best model. It is also less susceptible to multicollinearity.

b) PCR is basically used for estimating the unknown regression coefficients in a standard linear regression model. The main advantage of PCR is it can perform regression when the explanatory variables are highly correlated or even collinear. It is intuitive, as in, you can replace the basis with an orthogonal basis of principal components, drop the components that do not explain much variance, and regress the response onto the remaining components, which is a very useful aspect of using PCR

**7.**

Basically, Lasso is useful for feature selection, when our dataset has features with poor predictive power. Ridge regression is useful for the grouping effect, in which colinear features can be selected together.

1. We would like to decrease the model complexity, that is the number of predictors. We could use the forward or backward selection for this, but that way we would not be able to tell anything about the removed variables' effect on the response. Removing predictors from the model can be seen as settings their coefficients to zero. Instead of forcing them to be exactly zero. This way, we decrease model complexity while keeping all variables in the model. This, basically, is what Ridge Regression does. In ridge regression, the coefficients of correlated predictors are similar. In a nutshell, in an experiment where there are variables that needed to be pointing to zero in order to come up to a conclusion, we will have to shrink the data, ridge regression can help in that situation
2. In LASSO regression, it has basically the same shrinkage concept as ridge regression but for high values of λ, many coefficients are exactly zeroed under lasso, which is never the case in ridge regression. In lasso, one of the correlated predictors has a larger coefficient, while the rest are (nearly) zeroed. Lasso tends to do well if there are a small number of significant parameters and the others are close to zero. So, in an experiment, where there are few parameter or predictors for a single subject, or where there are only a few predictors that actually influence the response, LASSO performs better than Ridge regression

**8**. b) Ridge has larger bias, smaller variance.

**9.**

By looking at the graph, in the next step we will try to find the best cross validated lambda and get its sum of squared residuals and R squared. We are given that the minimum MSE occurs at lambda=0.47, which is basically what we explained. Now, since we know the lambda, we will compute information criteria, which is AIC and BIC using the selected lambda value. The next step will be to plot the information criteria, AIC and BIC, against lambda. Once that is done, we will find optimal lambdas to both AIC and BIC, fit final models, and get their sum of squared residuals and multiple R-squared. By the definition of Ridge regression, we will see how increasing lambda shrinks the coefficients. The higher the lambda, the more the coefficients are shrunk towards zero. We can also plot the coefficients with lambda (or log of lambda) to visualize the shrinkage.

**10**.

Forward selection method begins with an empty equation and the variables are added to the equation by looking at the p values or R squared adj, depending on the experiment. But, in backward elimination selection method, all variables are added from before, and the algorithm then goes through all of those variables in order to see which of them can be removed by looking at the p values or R squared adj, again, depending on the experiment. Now, if there are only a few useful variables in the entire experiment, the backward elimination method will still have to go through all of them in order to find those valuable variables by eliminating some useless predictors, as they are all added to the equation from the get-go. On the other hand, since, no variables are added in the beginning in the forward selection method, finding useful variables is faster.

**11**.

The over-fitting situation occurs when the model got too many datasets and classification, or prediction is conducted in which some predicators got the noise of other classes. In this fitting, the model supposed to give unexpectedly high accuracy. In over-fitting, the model may fail to classify the variable because of confusion created in logic due to too many conditions. The backward elimination technique curtails out the extraneous feature to circumvent the situation of over-fitting. In that situation, backward elimination will fail.

Also in the case of categorical variables with multiple levels to it, backward elimination will not even be possible to execute.

**12**.

A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill.

The validation dataset is different from the test dataset that is also held back from the training of the model but is instead used to give an unbiased estimate of the skill of the final tuned model when comparing or selecting between final models.

Basically, training data provides useful information regarding optimal fit but is well known to be positively biased. Also, you may suffer from the problem of overfitting or underfitting if you use training dataset. In a nutshell, training dataset is not nicely composed, and it is prone to error whilst test dataset has undergone some algorithmic calculations which makes it more organized and will result in better outcome. And for those reasons, CV metrics will perform far worse in training data sets than in test data sets

**13**.

Naive Bayes assumes that the features are conditionally independent. Real data sets are never perfectly independent, but they can be close. In short Naive Bayes has a higher bias but lower variance compared to logistic regression. If the data set follows the bias, then Naive Bayes will be a better classifier. Both Naive Bayes and Logistic regression are linear classifiers, Logistic Regression makes a prediction for the probability using a direct functional form whereas Naive Bayes figures out how the data was generated given the results. Logistic regression is mainly used in cases where the output is Boolean.

For instance, given the dataset of emails, if we were to find the spam emails, we basically see if the words “buy”, “lottery”, “money”, or “win” are in the emails. Now this dataset has bias, as these emails with the given words might actually not be spams, and in order to jump to conclusions we will have to see how the data was generated given the results. So, in this example, rather than just looking for those words (what logistic regression does), we will have to find how those words are generated and dig more into in what context they were used, which is what Naïve Bayes regression does.