**Problem 1: Best Subset Selection**

1. Table with AIC, BIC, adjusted R2 and Mallow’s Cp for each model size

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | AIC | BIC | Adjusted R2 | Mallow’s Cp |
| M1 | -44.36603 | -39.21661 | 0.5345839 | 26.038827 |
| M2 | -54.95846 | -47.23433 | 0.5868977 | 12.546389 |
| M3 | -63.17744 | -52.87859 | 0.6242063 | 5.092716 |
| M4 | -63.22555 | -50.35199 | 0.6280585 | 5.126817 |
| M5 | -63.72263 | -48.27437 | 0.6335279 | 4.785451 |
| M6 | -63.17571 | -45.15273 | 0.6349654 | 5.450474 |
| M7 | -62.66823 | -42.07054 | 0.6365002 | 6.099923 |
| M8 | -60.77886 | -37.60646 | 0.6327886 | 8.000636 |
| M9 | -58.77957 | -33.03246 | 0.6285705 | 10.000000 |

Model with the smallest AIC: M5

Model with the smallest BIC: M3

Model with largest Adjusted R2: M7

Model with smallest Mallow’s Cp: M5

Yes, they lead to different models. I have chosen Model 5 as the best model because it has followed 2 out of 4 of the criteria.

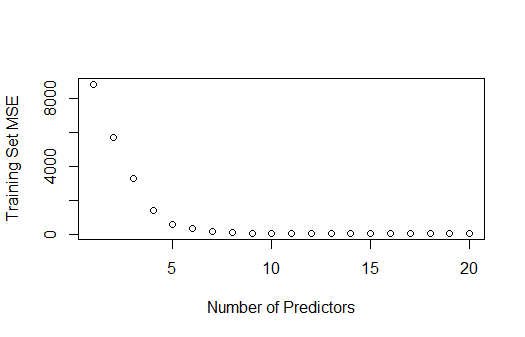
Model 5: lpsa = 0.49472926 + 0.54399786\*lcavol + 0.58821270\*lweight + -0.01644485\*age + 0.10122333\*lbph + 0.71490398\*svi

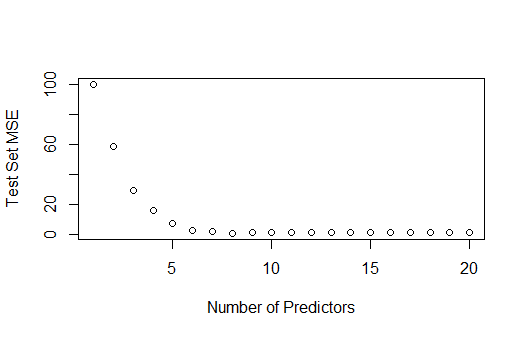
1. I have chosen Model 3 which has the lowest test MSE. The final model is lpsa = -0.7771566 + 0.5258519\*lcavol + 0.6617699\*lweight + 0.6656666\*svi
2. i) The 8 CV errors are: 0.6307298, 0.5929684, 0.6144691, 0.6035124, 0.5904755, 0.5890858, 0.5627177, 0.5693457

ii) Model 7 has the smallest CV error.

Model 7 is lpsa = 0.494 + 0.5695\*lcavol + 0.6144\*lweight – 0.0209\*age + 0.09735\*lbph + 0.7524\*svi – 0.105\*lcp + 0.0053\*pgg45

**Problem 2: Simulation Studies**

1. Refer to code in R script
2. Refer to code in R script
3. Here is the plot showing the training set MSE for models with different numbers of predictors. 
4. Here is the plot showing the test set MSE for models with different numbers of predictors.



1. The minimum test MSE is approximately 1.052 for a model size of 8. The number of true betas that aren’t equal to 0 in my data was 8. So when using 8 predictors, the model performs best in terms of making accurate predictions on the test data.
2. Comparing regression coefficient values

|  |  |  |
| --- | --- | --- |
| Bk | True Coefficient | Estimated Coefficient |
| Intercept | 1 | 0.06431864 |
| X1 | 5 | 4.83977271 |
| X2 | 1 | 0.96811937 |
| X3 | 6 | 5.79061043 |
| X4 | 3 | 2.86968914 |
| X5 | 9 | 8.83376926 |
| X6 | 2 | 2.06199835 |
| X10 | 4 | 4.07937043 |
| X11 | 1 | 1.03718703 |

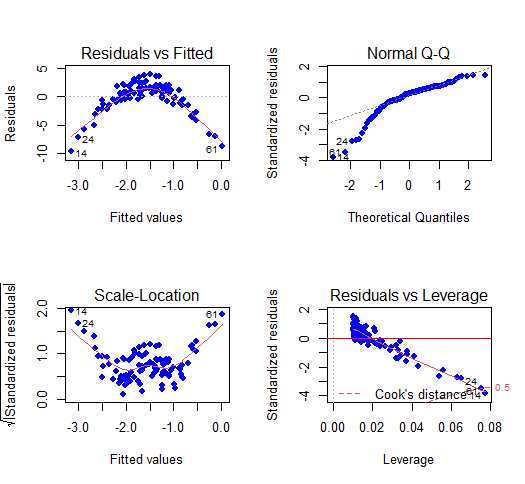
The model at which the test set MSE is minimized is a reasonably good approximation of the true model used to generate the data. Most of the estimated coefficients are close to their true values.

**Problem 3: Cross-validation**

1. The k-fold cross validation is implemented by taking the n number of observations and randomly splitting them into k non-overlapping groups. Each of these divided groups acts as a validation set and the rest is the training set. The test error is then estimated by averaging the k resulting MSE estimates.
2. i) disadvantages of the validation set approach relative to k-fold cross validation is that the estimate of the test error rate can be highly variable depending on which observations are included in the training/validation set. The validation set error rate may tend to overestimate the test error rate for the model fit on the entire data set.

ii) LOOCV is computationally intense since the model must be fit n times. But, LOOCV has higher variance and lower bias than k-fold.

1. Refer to code in R script
2. Diagnostic Plot:



Because of the curvature in the Residuals vs Fitted plot, it suggests that the linearity assumption may not hold. Because of the curvature in the Normal Q-Q, it suggests that there isn’t normality. The Scale-Location graph doesn’t have points randomly scattered around a horizontal line with a roughly constant spread, so homoscedasticity is not met.

1. LOOCV error for M1 = 7.288162

LOOCV error for M2 = 0.9374236

LOOCV error for M3 = 0.9566218

LOOCV error for M4 = 0.9539049

1. LOOCV error for M1 = 7.288162

LOOCV error for M2 = 0.9374236

LOOCV error for M3 = 0.9566218

LOOCV error for M4 = 0.9539049

The results are identical to the results from part e because LOOCV evaluates n folds of a single observation.

1. M2 has the lowest LOOCV error and this could be because the relation between x and y is quadratic in our given model of the simulated data set.
2. Looking at the p-values, the linear and quadratic terms are statistically significant. The cubic and 4th degree terms are not statistically significant. Thus, the results of the cross-validation agree with the conclusion as the minimum LOOCV error was for the quadratic model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Coefficient Estimate | Standard Error | Test statistic | p-value |
| Intercept | -1.55002 | 0.09591 | -16.162 | <2e-16 |
| M1 | 6.18883 | 0.95905 | 6.453 | 4.59e-09 |
| M2 | -23.94830 | 0.95905 | -24.971 | <2e-16 |
| M3 | 0.26411 | 0.95905 | 0.275 | 0.784 |
| M4 | 1.25710 | 0.95905 | 1.311 | 0.193 |

**Problem 4: Concept Review**

1. True. AIC assumes that the true model is not in the candidate pool and tries to mimic it. BIC will eventually lead to a true model if it’s in the candidate pool and if n is large enough. Even then, all the criteria use a different formula and weigh the factors differently. AIC, BIC, adjusted R2 and Mallow’s Cp can lead to different final models, but it’s also up to the data scientist’s understanding of the data as well.
2. False. Model 4 and Model 3 share the predictor X4. However, M3 is not nested in M4. If there are trash predictors in M3 and good predictors in M4, then the RSSM3 ≥ RSSM4 holds. If there are trash predictors in M4 and good predictors in M3, then the RSSM4 ≥ RSSM3
3. True. M2 is nested in M4, so adding predictors to M4 means it can do no worse than M2. So RSSM2 ≥ RSSM4
4. Yes, it is true that in the context of subset selection, if a model Mk+1 is identified as the best model with k+1 predictors, it must contain a subset of the predictors found in Mk. This is because subset selection algorithms systematically add or remove predictors to find the model with the smallest RSS. If Mk+1 is better than Mk, it implies that adding one more predictor improves the model’s fit. However, since Mk already represents the best model with k predictors, it follows that the additional predictor in Mk+1 must be one of the k predictors already present in Mk. Therefore, Mk+1 contains a subset of the predictors in Mk.
5. AIC and BIC offer advantages over using test MSE as model selection criteria because they balance model complexity with goodness of fit. While test MSE only evaluates a model’s predictive accuracy on the data it was not trained on, AIC and BIC consider both goodness of fit and model complexity. AIC penalizes complex models more heavily, encouraging parsimonious models that are less prone to overfitting. BIC, on the other hand, has a stronger penalty for the model complexity, promoting even greater model simplicity. This helps prevent overfitting and often leads to models that generalize better to new data, making AIC and BIC valuable tools for effective model selection.
6. I would disagree with the colleague’s claim. While p-values can provide some guidance, they are not the sole determinants of predictor significance. The choice to include or exclude a predictor should consider various factors, such as domain knowledge, the context of the problem, and the potential impact of the predictor on the model’s overall performance. A p-value slightly above the significance threshold does not necessarily imply that a predictor is not meaningful. It’s essential to assess the practical significance and consider the broader context before excluding a predictor from the model.