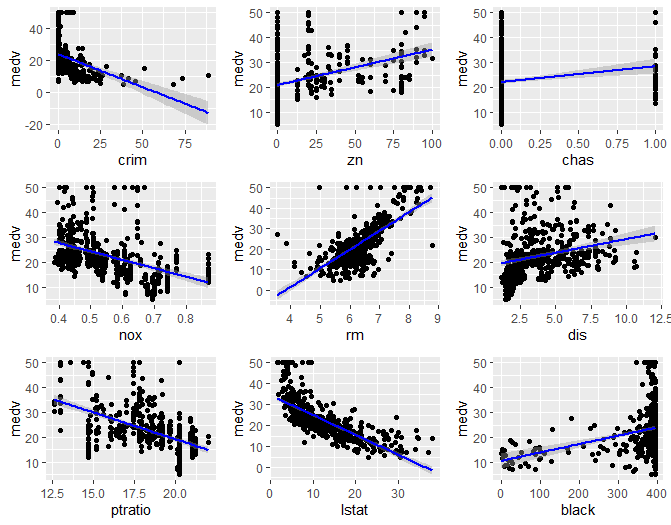
**Problem 1:**

To start, diagnostics were done on the Boston data set. After creating a model with response as medv and all other variables as the explanatory variables, I looked at the multicollinearity vif values. Predictors rad and tax had high vif values of 7.48 and 9.01 respectively. I built another model excluding these two predictors and the vif values were not high, indicating there wasn’t much multicollinearity left.

I then split the dataset into an 80% train and 20% test. I fit the model without rad and tax on the training data and the summary output showed that indus and age were not statistically significant as they had p-values of 0.413 and 0.749 respectively. Building a new model excluding predictors rad, tax, indus, and age on the training data gave an output of statistically significant predictors in predicting medv.

I performed best subset selection to see other candidate models. From this, following the AIC criteria, the best model was Medv = crim + zn + chas + nox + rm + dis + rad + tax + ptratio + black + lstat. But, based on previous data exploration, rad and tax indicated multicollinearity. Which is why I’ve made the decision to exclude these two predictors when coming up with my final model.

To further check for relevant predictors, I created plots comparing how medv and the remaining predictors correlate with each other. The figure below shows this. Predictor lstat has a more non-linear correlation with medv.



To address the non-linear correlation between lstat and medv, I made a new model medv = crim + zn + chas + nox + rm + dis + ptratio + lstat + lstat^2 + black. The summary of this model indicated that zn was not statistically significant, so I built a similar model, but excluded zn. To check validity between these two models, I performed LOOCV on them. The test MSE for the model including zn was 19.5336 while the test MSE for the model excluding zn was 19.52784. The adjusted R2 for the model including zn was 0.7718 while the adjusted R2 for the model excluding zn was 0.7712.

Although the two test MSEs and adjusted R2 values are relatively similar, my final model was medv = crim + chas + nox + rm + dis + ptratio + lstat + lstat^2 + black.

Summary of the final model:

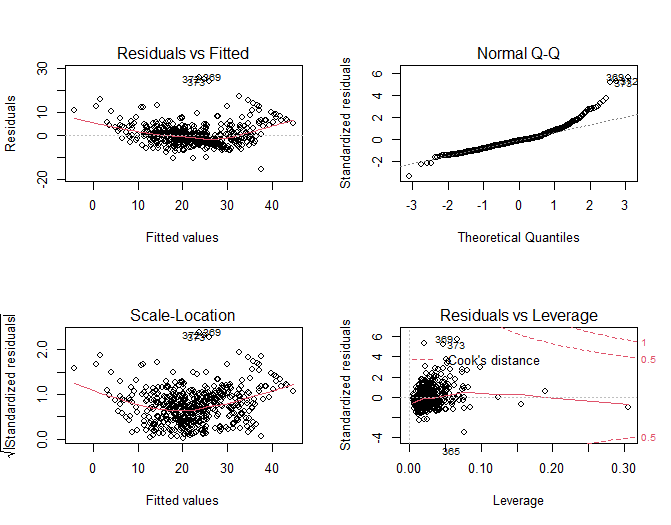
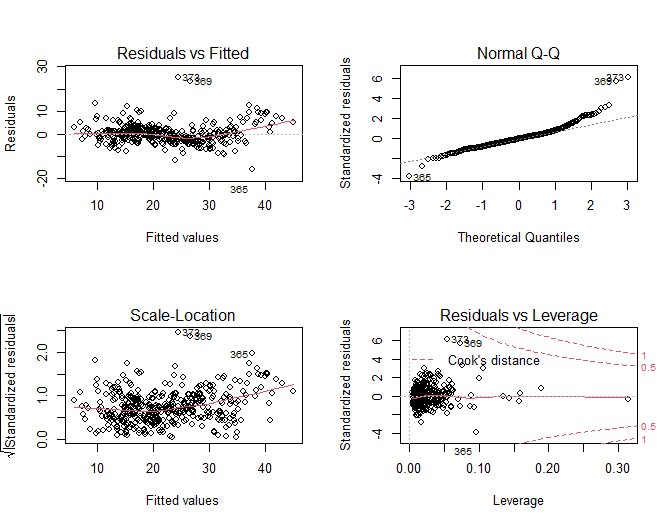
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Predictor | Coefficient | Standard Error | Test statistic | P-value |
| Intercept | 24.726998 | 4.630901 | 5.340 | 1.58e-07 |
| Crim | -0.095297 | 0.028493 | -3.345 | 0.000903 |
| Chas | 2.302368 | 0.876115 | 2.628 | 0.008926 |
| Nox | -13.056991 | 3.339687 | -3.910 | 0.000109 |
| Rm | 3.487381 | 0.400453 | 8.709 | <2e-16 |
| Dis | -1.136923 | 0.169343 | -6.714 | 6.64e-11 |
| Ptratio | -0.797742 | 0.115982 | -6.878 | 2.39e-11 |
| Lstat | -82.083594 | 6.900370 | -11.896 | <2e-16 |
| Lstat^2 | 42.706333 | 4.640684 | 9.203 | <2e-16 |
| Black | 0.005703 | 0.002547 | 2.239 | 0.025710 |

Residual standard error: 4.302 on 394 degrees of freedom.

Multiple R2 = 0.7763, Adjusted R2 = 0.7712

F-statistic: 151.9 on 9 and 394 DF, p-value: < 2.2e-16

The residual vs fitted plot for the full model (right) vs the final model (left) is below.



The four assumptions for multiple linear regression are linearity, independence, homoscedasticity, and normality. As I’ve previously mentioned, lstat showed a non-linear relationship between the response, and to address this issue a polynomial predictor was included. The full model had an issue with homoscedasticity as the spread of the residuals were not constant. The final model distributes the residuals in a constant manner significantly better than the full model.

**Problem 2:**

1. Mk, subset should have the smallest training MSE since best subset selection in the only method where all the predictors of all models are considered.
2. Mk, subset should have the smallest test MSE since best subset selection considers more models than the other techniques we’ve covered.
3. For both the forward and backward selection approaches, Model 12 was the best model based on the AIC criteria.

Forward Selection Model 12: App = -398.66288580 - 508.34782756\*PrivateYes + 1.67744475\*Accept - 0.76280730\*Enroll + 54.15469175 \*Top10perc - 16.01806305\*Top25perc + 0.06190806\*P.Undergrad - 0.09864244\*Outstate + 0.11152696\*Room.Board - 11.61633077\*PhD + 21.15197158\*S.F.Ratio + 0.08795221\*Expend + 9.10501731\*Grad.Rate

Backward Selection Model 12: App = -398.66288580 - 508.34782756\*PrivateYes + 1.67744475\*Accept - 0.76280730\*Enroll + 54.15469175 \*Top10perc - 16.01806305\*Top25perc + 0.06190806\*P.Undergrad - 0.09864244\*Outstate + 0.11152696\*Room.Board - 11.61633077\*PhD + 21.15197158\*S.F.Ratio + 0.08795221\*Expend + 9.10501731\*Grad.Rate

Both models, FWD Model 12 and BWD Model 12 reported a test MSE of 1395597. The final model can be either of the two. If we wanted to look at other criteria like BIC, Mallow’s Cp and RSS, we may be able to have a better understanding of which model is the best.

**Problem 3: A Puzzling Problem**

Issues with the R output:

* Coefficients with large standard errors. The coefficients for both x1 and x2 has large standard errors relative to their estimates. Large standard errors indicate that there is a high level of uncertainty in the coefficient estimates.
* Insignificant coefficients. The p-values associated with x1 and x2 are very high. Typically, we look for p-values less than 0.05 to determine if a predictor is statistically significant.
* Low R-squared value. The multiple R-squared value and adjusted R-squared value are both very low. Here, the model explains very little of the variance in y, suggesting that it might not be a good fit for the data.

Why these issues could happen:

* There may be multicollinearity between the x1 and x2. When there is multicollinearity between two predictors, it becomes difficult to separate the effects of each predictor on the response, leading to unstable coefficient estimates and high standard errors.
* Multiple linear regression assumes a linear relationship between the predictors and response. If the true relationship is nonlinear, the model may not fit well, leading to low R-squared values and insignificant coefficients.

**Problem 4: Interaction Terms**

1. A screenshot of a computer program

   Description automatically generated
2. Summary of model fit:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Coefficient Estimate | Standard Error | t-value | p-value |
| Intercept | 211.1430 | 32.4572 | 6.505 | 2.34e-10 |
| Income | 5.9843 | 0.5566 | 10.751 | <2e-16 |
| StudentYes | 382.6705 | 65.3108 | 5.859 | 9.78e-9 |

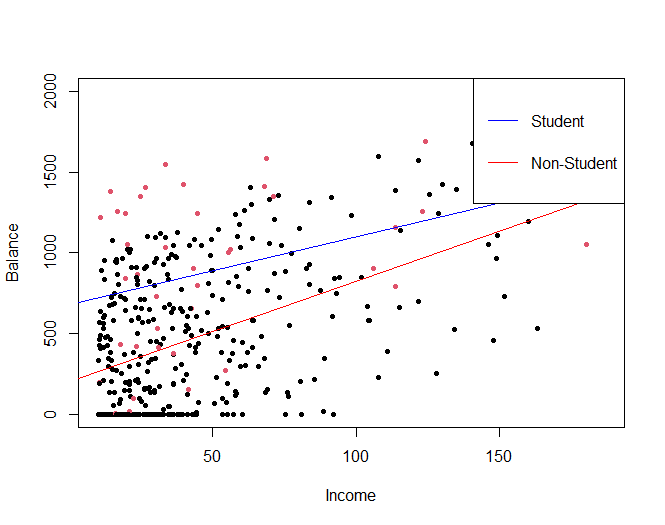
1. Model for Students: Yhat = 211.1430 + 5.9843\*Income + 382.6705

Model for non-students: Yhat = 211.1430 + 5.9843\*Income

1. For students, the coefficient of 5.9843 means that on average, each additional unit increase in income is associated with an increase of 5.9843 units in credit card balance, while keeping all other factors constant.

For non-students, the coefficient of 5.9843 means that on average, each additional unit increase in income is associated with an increase of 5.9843 units in credit card balance, while keeping all other factors constant.

1. Based on this scatterplot, the effect of income on balance is different for students and non-students as there is a noticeable difference in slopes. This suggests that the effect of income varies by student status and the constraint in the initial model may not be reasonable. We should consider using an interaction term between income and student to account for the varying effects.



1. Model for students: Yhat = 200.6232 + (6.2182\*Income – 1.9992\*Income) + 476.6758

= 200.6232 + 4.2182 \* Income + 476.6758

Model for non-students: Yhat = 200.6232 + 6.2182\*Income

1. For students, as the income increases by one unit, their average balance is expected to increase by 4.2128 units, while holding all other variables constant. The additional 476.6758 constant represents the baseline effect of being a student.

For non-students, as the income increases by one unit, their average balance is expected to increase by 6.2182 units, while holding all other variables constant.