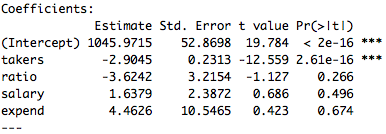
Israel Diego

3/22/16

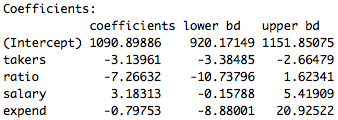
STATS 500 Homework 5

1. Using sat data, fit a model with total as the response and takers, ratio, salary and expend, with the following methods:

a) Ordinary least squares

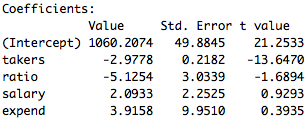


b) Least absolute deviations



c) Huber’s robust regression

|  |
| --- |
| Pr(>|t|) |
| 0 |
| 0 |
| 0.098 |
| 0.358 |
| 0.70 |

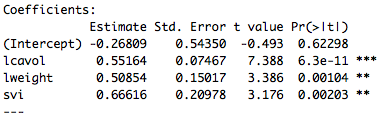


In the OLS example, the variable **takers** is the only statistically significant variable at the .1% level. In the LAD regression, the **takers**, **ratio**, and **expend** coefficients decreased, while **salary** increased. Similar to OLS, only **takers** has a statistically significant coefficient, because its confidence interval does not contain the zero term. In Huber’s Robust regression, we see more similar results to the OLS regression. The changes in the coefficients of variables are not as drastic as those in the LAD regression. The **takers** coefficient did not change much at all, **ratio** coefficient decreased a bit, **salary** increased a little, and **expend** decreased compared to the OLS. Also, the **takers** coefficient is statistically significant at the .1% level, and **ratio** seems to be significant now at the 10% level, if we assume asymptotic normality. All other variable coefficients are not significant.

2. Use the prostate data with **lpsa** as the response and the other variables as predictors. Implement the following variable selection methods:

a) **Backward Elimination**

The model selected from Backward Elimination chooses: **lcavol**, **lweight**, and **svi** as predictors. The regression model is shown below:

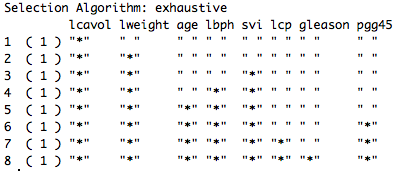


Residual standard error: 0.7168 on 93 degrees of freedom

Multiple R-squared: 0.6264, Adjusted R-squared: 0.6144

F-statistic: 51.99 on 3 and 93 DF, p-value: < 2.2e-16

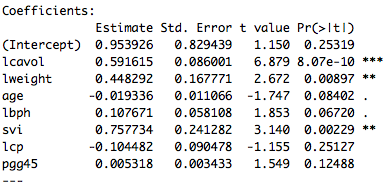
For the Adjusted method and Mallows’ method, we refer to the regsubsets table summary below.



b) **Adjusted**

Macintosh HD:Users:rdiego:Documents:Senior Year:Winter 2016:Stats 500:homeworks:HW5:Adj R2 plot.pdf

The plot of the Adjusted method shows that it is optimal to use 7 parameters in order to get the maximum Adjusted . The table above says the best possible model with 7 regressors has: **lcavol**, **lweight**, **age**, **lbph**, **svi**, **lcp**, and **pgg45**.



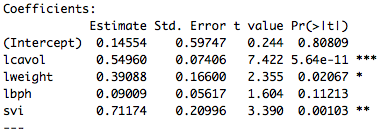
Residual standard error: 0.7048 on 89 degrees of freedom

Multiple R-squared: 0.6544, Adjusted R-squared: 0.6273

F-statistic: 24.08 on 7 and 89 DF, p-value: < 2.2e-16

Macintosh HD:Users:rdiego:Documents:Senior Year:Winter 2016:Stats 500:homeworks:HW5:Cp plot.pdfc) **Mallows’**

The plot of the Mallows’ method shows that it is optimal to use 4 parameters in order to get the minimum statistic. The table above says the best possible model with 4 regressors has: **lcavol**, **lweight**, **lbph**, and **svi**.

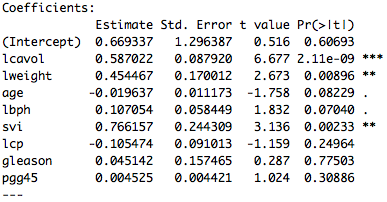


Residual standard error: 0.7108 on 92 degrees of freedom

Multiple R-squared: 0.6366, Adjusted R-squared: 0.6208

F-statistic: 40.29 on 4 and 92 DF, p-value: < 2.2e-16

**The Full Model:**



Residual standard error: 0.7084 on 88 degrees of freedom

Multiple R-squared: 0.6548, Adjusted R-squared: 0.6234

F-statistic: 20.86 on 8 and 88 DF, p-value: < 2.2e-16

Looking at all the models selected compared to the full model, the Adjusted method had the largest Adjusted of 0.6273, followed by the Mallows’ model selected with Adjusted 0.6208, and finally the Backward Elimination method with Adjusted of 0.6144. The Backward Elimination model is the simplest model with only three regressors and is almost the same as the Mallows’ , model except for the regressor **lbph**, which is the additional regressor in the Mallows’ , model. The full model seems to have too many variables, since most variables are not significant. Though the Adjusted model chosen is an improvement of the full model, it still has too many regressors. The Mallows’ model is the borderline case in relation to the Backward Elimination model. In my opinion the Mallows’ model is best out of all the models chosen, since it includes the most significant variables: **lcavol**, **lweight**, and **svi**, plus **lbph** which does seem to have some explanatory power,