Stats 500: HW #5

Diana Liang

11/11/2019

**Part 1: Different Models for Robust Regression**

1. **Ordinary Least Squares:**

##   
## Call:  
## lm(formula = total ~ takers + ratio + salary + expend, data = sat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -90.531 -20.855 -1.746 15.979 66.571   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1045.9715 52.8698 19.784 < 2e-16 \*\*\*  
## takers -2.9045 0.2313 -12.559 2.61e-16 \*\*\*  
## ratio -3.6242 3.2154 -1.127 0.266   
## salary 1.6379 2.3872 0.686 0.496   
## expend 4.4626 10.5465 0.423 0.674   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 32.7 on 45 degrees of freedom  
## Multiple R-squared: 0.8246, Adjusted R-squared: 0.809   
## F-statistic: 52.88 on 4 and 45 DF, p-value: < 2.2e-16

Ordinary least squares regression works when the residual errors are normally distributed but not as well when the errors are robust, since only a few outliers could greatly affect the residual sum of squares. In the model above, the only significant predicator variable is takers.

1. **Least Absolute Deviations:**

##   
## Call: rq(formula = total ~ takers + ratio + salary + expend, data = sat)  
##   
## tau: [1] 0.5  
##   
## Coefficients:  
## coefficients lower bd upper bd   
## (Intercept) 1090.89886 920.17149 1151.85075  
## takers -3.13961 -3.38485 -2.66479  
## ratio -7.26632 -10.73796 1.62341  
## salary 3.18313 -0.15788 5.41909  
## expend -0.79753 -8.88001 20.92522

Least absolute deviations regression works for long tailed error distributions since it is not as easily influenced by outliers. As a consequence, the assumption that the errors are normally distributed is unreliable, so significant variables are found by if zero is not included within the boundaries. In this case, only takers does not include zero between the upper and lower bounds.

1. **Huber’s Robust Regression:**

##   
## Call: rlm(formula = total ~ takers + ratio + salary + expend, data = sat)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -92.510 -17.701 -1.002 15.015 77.058   
##   
## Coefficients:  
## Value Std. Error t value   
## (Intercept) 1060.2074 49.8845 21.2533  
## takers -2.9778 0.2182 -13.6470  
## ratio -5.1254 3.0339 -1.6894  
## salary 2.0933 2.2525 0.9293  
## expend 3.9158 9.9510 0.3935  
##   
## Residual standard error: 25.58 on 45 degrees of freedom

Huber’s robust regression is a mixture of OLS and LAD. This way the variables can have meaningful t-values but not p-values. In this model, takers has a much more extreme t-value compared to the other predictor variables, so takers is the most significant variable.

**Part 2: Different Procedures for Variable Selection**

3 different procedures were used to find the variables with the best fit for lpsa in the prostate data: backwards elimination, adjusted R^2, and Mallows’ Cp.

1. **Backwards Elimination:**

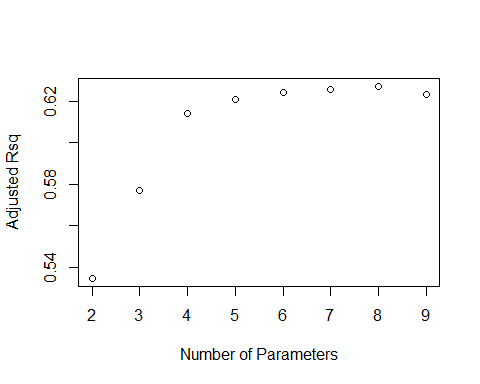
## Call:  
## lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.72964 -0.45764 0.02812 0.46403 1.57013   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.26809 0.54350 -0.493 0.62298   
## lcavol 0.55164 0.07467 7.388 6.3e-11 \*\*\*  
## lweight 0.50854 0.15017 3.386 0.00104 \*\*   
## svi 0.66616 0.20978 3.176 0.00203 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## 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

In this testing-based procedure, each variable with the greatest p-value that is above the significance level of 0.05 is taken away one at a time until either the R^2 value drastically decreases or only significant variables are left over. The left over variables in this case were lcavol, lweight, and svi. All the variables are significant but the R^2 value is the lowest out of all of the methods. This could mean that this model has the worst fit of the three models, but this could also be a consequence of having the least variables.

The selected variables for the criterion-based procedures in part B and C can be found using the summary below.

## Subset selection object  
## Call: regsubsets.formula(lpsa ~ ., data = prostate)  
## 8 Variables (and intercept)  
## Forced in Forced out  
## lcavol FALSE FALSE  
## lweight FALSE FALSE  
## age FALSE FALSE  
## lbph FALSE FALSE  
## svi FALSE FALSE  
## lcp FALSE FALSE  
## gleason FALSE FALSE  
## pgg45 FALSE FALSE  
## 1 subsets of each size up to 8  
## Selection Algorithm: exhaustive  
## lcavol lweight age lbph svi lcp gleason pgg45  
## 1 ( 1 ) "\*" " " " " " " " " " " " " " "   
## 2 ( 1 ) "\*" "\*" " " " " " " " " " " " "   
## 3 ( 1 ) "\*" "\*" " " " " "\*" " " " " " "   
## 4 ( 1 ) "\*" "\*" " " "\*" "\*" " " " " " "   
## 5 ( 1 ) "\*" "\*" "\*" "\*" "\*" " " " " " "   
## 6 ( 1 ) "\*" "\*" "\*" "\*" "\*" " " " " "\*"   
## 7 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" " " "\*"   
## 8 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*" "\*"

1. **Adjusted R^2:**



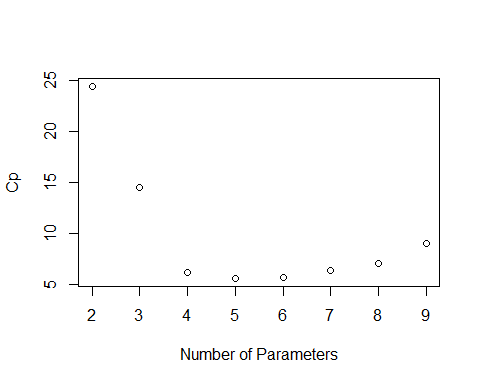
This procedure finds the variables that have the greatest adjusted R^2 value. The adjusted R^2 is just a rescaling of the usual R^2 value, so a greater adjusted R^2 means a greater R^2 value and a better fitted model. Looking at the above summary and graph, the 7 variables that have the greatest adjusted R^2 value are all the variables except gleason. The model summary is below.

## Adjusted R^2: 7

##   
## Call:  
## lm(formula = lpsa ~ . - gleason, data = prostate)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.73117 -0.38137 -0.01728 0.43364 1.63513   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.953926 0.829439 1.150 0.25319   
## lcavol 0.591615 0.086001 6.879 8.07e-10 \*\*\*  
## lweight 0.448292 0.167771 2.672 0.00897 \*\*   
## age -0.019336 0.011066 -1.747 0.08402 .   
## lbph 0.107671 0.058108 1.853 0.06720 .   
## svi 0.757734 0.241282 3.140 0.00229 \*\*   
## lcp -0.104482 0.090478 -1.155 0.25127   
## pgg45 0.005318 0.003433 1.549 0.12488   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## 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

Not all the variables chosen are significant, but the R^2 value is still similar to that from the backward elimination method, so they have a similar level of fit.

1. **Mallows’ Cp:**



The last procedure finds the variables that have the smallest Cp statistic value. The Cp statistic is a rescaling of the RSS, so a smaller Cp value is a smaller RSS and a better fitted model. Looking at the above summary and graph, the 4 variables that lead to the smallest Cp value are lcavol, lweight, lbph, and svi. The model summary is below.

## Mallows Cp: 4

##   
## Call:  
## lm(formula = lpsa ~ lcavol + lweight + lbph + svi, data = prostate)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.82653 -0.42270 0.04362 0.47041 1.48530   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.14554 0.59747 0.244 0.80809   
## lcavol 0.54960 0.07406 7.422 5.64e-11 \*\*\*  
## lweight 0.39088 0.16600 2.355 0.02067 \*   
## lbph 0.09009 0.05617 1.604 0.11213   
## svi 0.71174 0.20996 3.390 0.00103 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## 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

Since the selected variables are very similar to those of the backwards elimination procedure, the R^2 is more similar. All three procedures have similar coefficients of determination, so all the models fit to a similar degree, but the backwards elimination and Mallows’ Cp procedures have the most similar selected variables.