Model Building on the Chicago Redline Dataset: Predicting FAIR Plan Renewals

Oscar Sucre: Team Coordinator, Abstract, Results
Anisha Lal: Models, Work Cited, Methods
Alan Wu: Code, Discussion
Avi Kodali: Introduction, Code

12/6/22

Abstract

Our goal for this project was to find the best model to predict the FAIR model plan for chicago insurance redlining. We used the chredlin dataset from the faraway library package. This dataset is from a 1970's study on the relationship between insurance redlining in Chicago with the variables of race, fire, theft, age, income, side (north or south), and involact which is the response for our models being the new FAIR plan policies and renewals per 100 housing units. This insurance is for housing units in Chicago that are at risk for redlining, which means they are rejected for insurance because they are in a high risk/low income area. We utilized partial regression plots and residuals vs fitted values to analyze the full model. The models were built using all possible regressions, best subset regressions, and stepwise regression procedures. K-fold cross validation was used to validate the models.

Introduction

Using race, fire, theft, age, income, and sides, we wanted to build the best model that would predict the number of new FAIR plan policies and renewals per 100 housing units(involact). To identify the best model, we wanted to find the model that maximized adjusted R^2 and minimized MSE(mean squared error). Adjusted R^2 explains how much of the variance in involact is explained by the regressors we use (while also accounting for the number of regressors included in the model). MSE is the average of the square of the errors, the error being the difference between the values predicted by the model and the actual values. So the best model would be one that explains as much variance in the model as possible, while not deviating from the actual trends in the data. The first step in finding the best model was first to build a model that included all of the regressors (the full model). Then, we used various regression techniques to find which regressors are more important in predicting involact. The first procedure was the all possible regressions procedure, which outputs all possible regression models with one regressor, then all possible regression models with two regressors, then all possible regression models with three regressors, and so on until all possible combinations of regressors are found and outputted. The next procedure was the best subsets regression, which outputs the best model with one regressor, the best model with two regressors, the best model with three regressors, and so on until the full model is reached. The final regression technique was the stepwise regression procedures. Stepwise regression involves adding and/or removing regressors to the model based on some threshold and parameter until the best model is reached. Within stepwise regression, we used forward selection, backward elimination, and bidirectional regression. To analyze the resulting models, we looked at the residual plots for each model, as well as the corresponding diagnostic measures (adjusted R^2, Mallow's Cp, Akaike information criterion, etc.). After identifying the best models from the regression analysis, we used k-folds cross-validation to see the out-of-sample performance of the model and make sure there was no overfitting.

Materials and Methods

Variables and Regressors

Response Variable Y: involact (new FAIR plan policies and renewals per 100 housing units)

X1: race (racial composition in percent minority)

X2: fire (fires per 100 housing units)

X3: theft (theft per 1000 population)

X4: age (percent of housing units built before 1939)

X5: income (median family income in thousands of dollars)

X6: sides (North or South side of Chicago)

Procedure

1. Conduct exploratory analysis on the full model
2. Look through all possible regressions and best subset regressions
3. Perform forward, backward, and stepwise regression methods
4. Select best two models and utilize K-fold Cross Validation to validate model

Methods

OLS Regression: This is used to determine the coefficients of a linear regression equation to describe the relationship between our independent variables (race, fire, theft, age, income, and side) against our dependent variable (involact).

Forward Selection: This is a type of stepwise regression where an independent variable is added to improve the model at each step. This process was performed in R using the function ols_step_forward_p.

Backward Elimination: This is a type of stepwise regression where an independent variable is removed from a full starting model at each step. This process is repeated until removing variables does not make the model better anymore.This process was performed in R using the function ols_step_backwards_p.

Both Directions: This is a stepwise regression that is a combination of forward and backward selection. We start with a blank model and then add variables at each step if it makes the model better. This process was performed in R using the function ols_step_both_p.

All Possible Regressions: This is a regression with all possible variations of the model that are possible to exist. Since there are 8 variables in the data set there are 2^6 possible combinations to be tested.

RMSE: This is the standard deviation of the residuals to show the spread of these data points from the regression line.

Adjusted R^2: This is a modified version of R^2 that is adjusted for the number of predictors in the model.

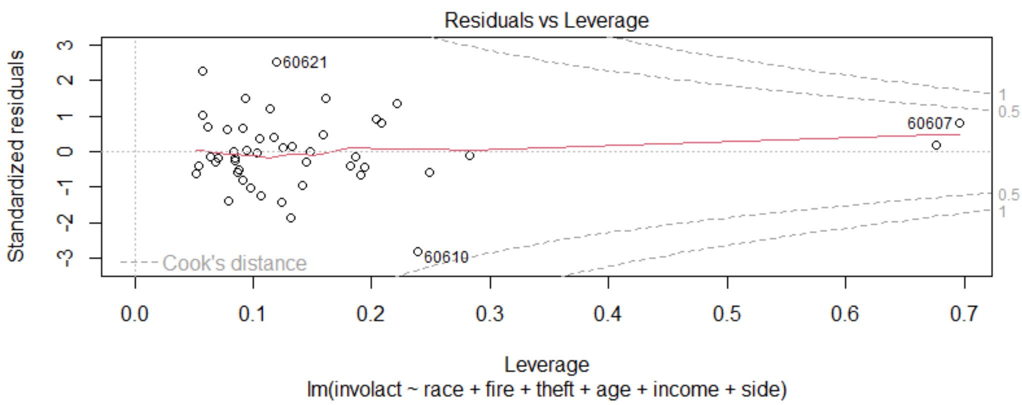
K-Fold Cross Validation: This is a resampling procedure where there are k number of groups that the data sample is split into. This method evaluates the performance of model on different subsets of training data then calculates the average prediction error.

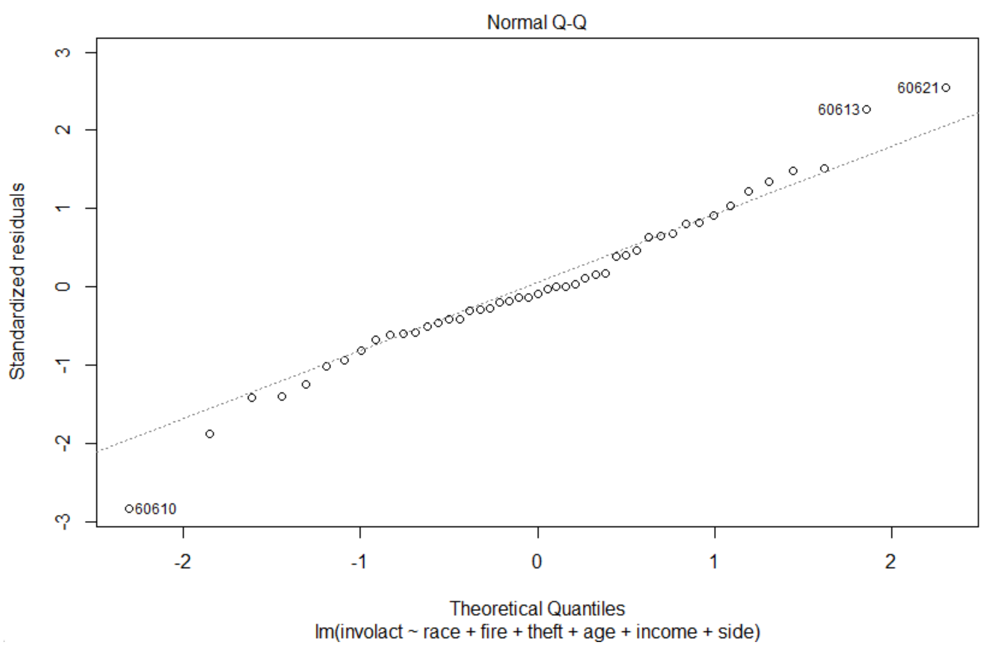
K-Fold Repeated Cross Validation: This is a K-Fold Cross Validation where we repeat the process of splitting the data into K-folds multiple times and take the mean error to give you the final model error. We did 10 repeats with 5 folds.

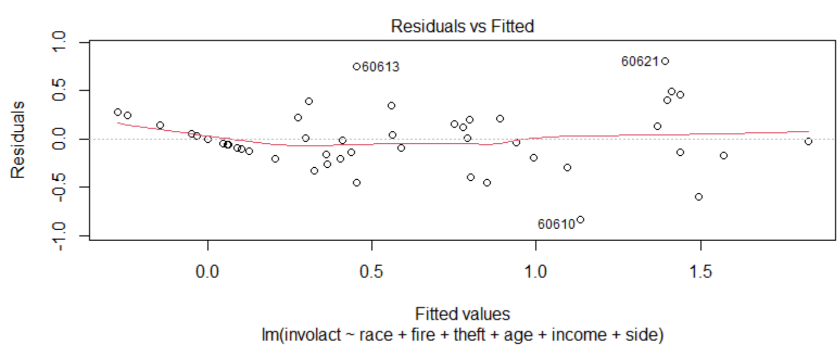
Results

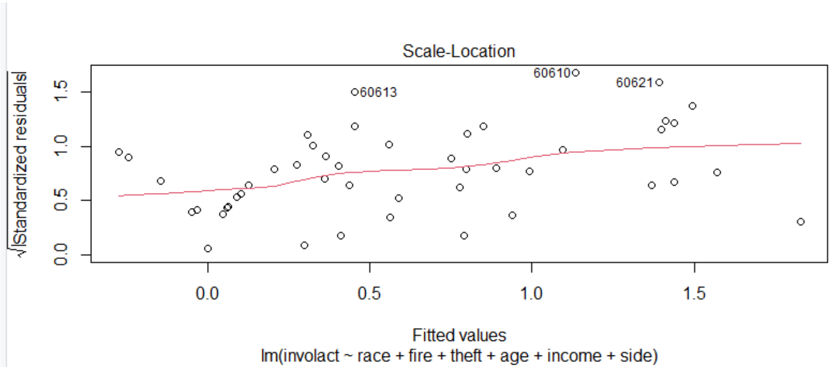
```
Call:
lm(formula = involact ~ race + fire + theft + age + income +
    side, data = chredlin)

Residuals:
     Min       1Q   Median       3Q      Max
-0.83562 -0.16506 -0.02719  0.17675  0.80848

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.629092   0.511993  -1.229  0.22636
race         0.008900   0.002638   3.374  0.00166 **
fire         0.039070   0.008638   4.523 5.33e-05 ***
theft       -0.010210   0.002922  -3.494  0.00118 **
age          0.008419   0.002919   2.884  0.00629 **
income       0.024696   0.032092   0.770  0.44609
sides        0.024031   0.125054   0.192  0.84859
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3391 on 40 degrees of freedom
Multiple R-squared:  0.7511,    Adjusted R-squared:  0.7137
F-statistic: 20.11 on 6 and 40 DF,  p-value: 1.124e-10
```

```
#resiudals vs fitted values, leverage, normal qq plot
plot(lmod)
```

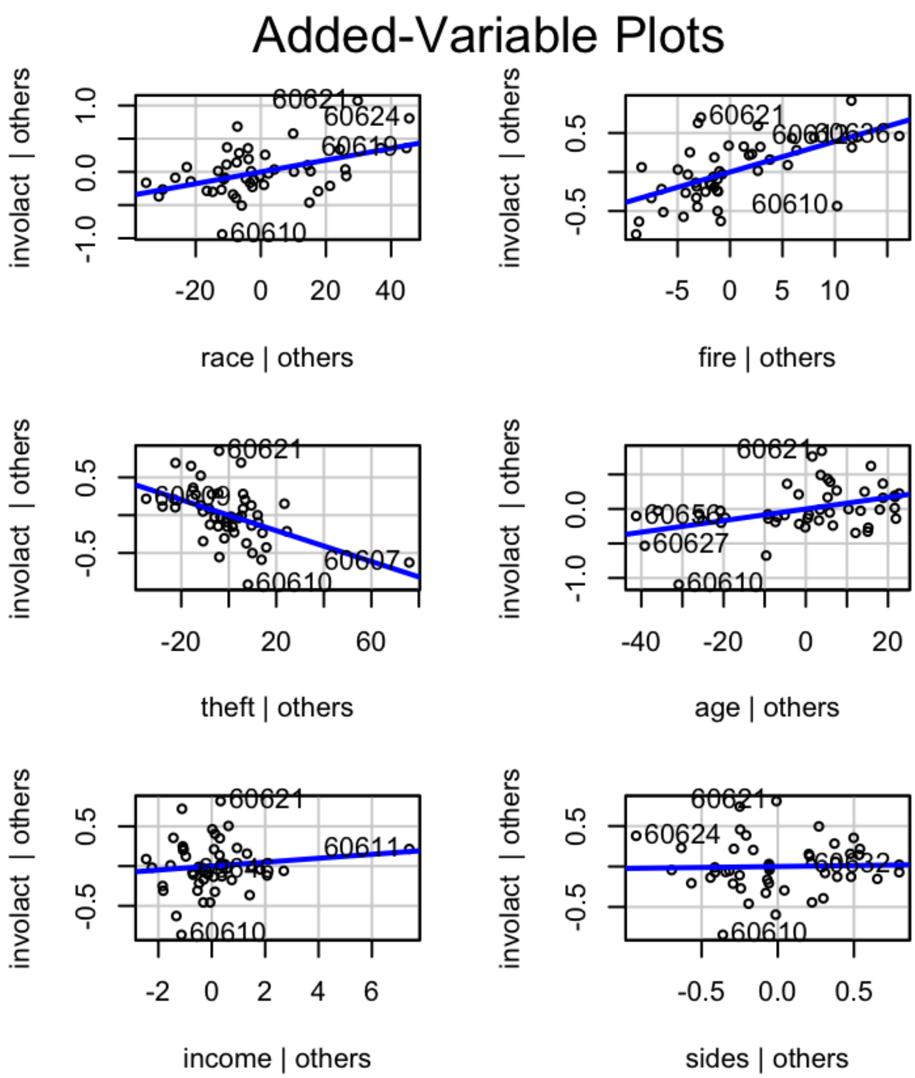









```
#plot of partial regression plots
avPlots(lmod,print_plot=TRUE)
```



Model Selection and Building

```r
#find the best subsets model for each number of regressors
bestsubset <- ols_step_best_subset(lmod)
bestsubset
```

```
> bestsubset
           Best Subsets Regression
----------------------------------------------
Model Index    Predictors
----------------------------------------------
     1         race
     2         race fire
     3         race fire theft
     4         race fire theft age
     5         race fire theft age income
     6         race fire theft age income side
----------------------------------------------
```

Subsets Regression Summary

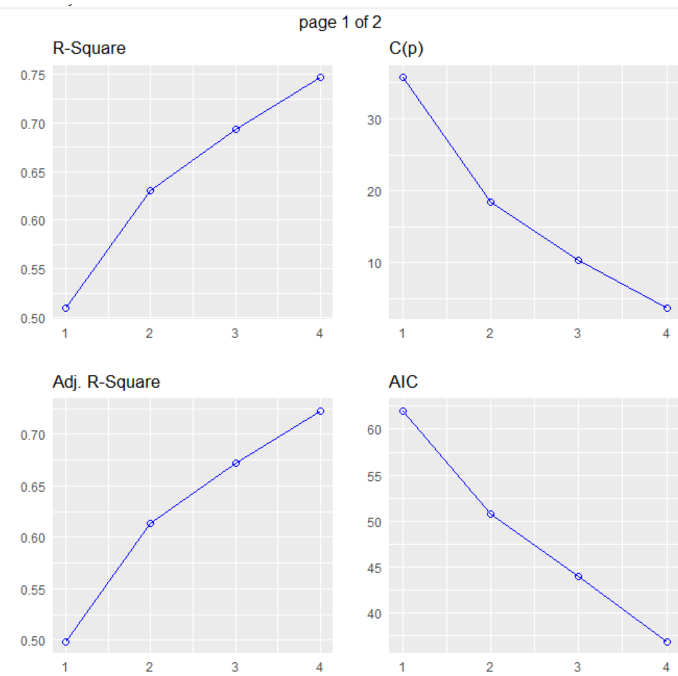
Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC	SBIC	SBC	MSEP	FPE	HSP	APC
1	0.5094	0.4985	0.4611	35.8205	62.0324	-73.2887	67.5828	9.4685	0.2100	0.0046	0.5342
2	0.6303	0.6135	0.5361	18.4083	50.7436	-83.9771	58.1442	7.3026	0.1652	0.0036	0.4202
3	0.6932	0.6718	0.6211	10.2921	43.9701	-89.7865	53.2208	6.2033	0.1431	0.0031	0.3639
4	0.7472	0.7231	0.6717	3.6204	36.8759	-94.9831	47.9767	5.2367	0.1231	0.0027	0.3130
5	0.7508	0.7204	0.6628	5.0369	38.1959	-93.1578	51.1469	5.2905	0.1266	0.0028	0.3221
6	0.7511	0.7137	0.6518	7.0000	40.1525	-90.8389	54.9537	5.4212	0.1321	0.0029	0.3361

```r
#find all possible regressions
allpossible <- ols_step_all_possible(lmod)
allpossible
```

	Index	N	Predictors	R-Square	Adj. R-Square	Mallow's Cp
1	1	1	race	0.50944482	0.4985435890	35.820540
2	2	1	fire	0.49426488	0.4830263212	38.259595
5	3	1	income	0.44202173	0.4296222133	46.653825
4	4	1	age	0.22631819	0.2091252577	81.312249
6	5	1	side	0.03815180	0.0167773947	111.546108
3	6	1	theft	0.02238939	0.0006647144	114.078752
7	7	2	race fire	0.63026093	0.6134546099	18.408266
9	8	2	race age	0.60351278	0.5854906358	22.706057
14	9	2	fire income	0.58280204	0.5638384994	26.033780
12	10	2	fire theft	0.57864789	0.5594955201	26.701254
10	11	2	race income	0.56179469	0.5418762647	29.409162
13	12	2	fire age	0.53590745	0.5148123307	33.568625
15	13	2	fire side	0.53245730	0.5112053620	34.122981
11	14	2	race side	0.52492286	0.5033284395	35.333588
8	15	2	race theft	0.51056919	0.4883223366	37.639880
19	16	2	age income	0.46344652	0.4390577220	45.211372
21	17	2	income side	0.44681658	0.4216718836	47.883406
17	18	2	theft income	0.44326035	0.4179539978	48.454809
20	19	2	age side	0.32366083	0.2929181401	67.671604
16	20	2	theft age	0.22632061	0.1911533627	83.311860
18	21	2	theft side	0.07427738	0.0321990809	107.741587
22	22	3	race fire theft	0.69322108	0.6718178973	10.292070
23	23	3	race fire age	0.67065475	0.6476771705	13.917942
24	24	3	race fire income	0.64426625	0.6194476213	18.157944
32	25	3	fire theft age	0.63605488	0.6106633586	19.477317
33	26	3	fire theft income	0.63226951	0.6066138963	20.085535
25	27	3	race fire side	0.63026623	0.6044708531	20.407415
26	28	3	race theft age	0.61820310	0.5915661077	22.345673
29	29	3	race age income	0.61047756	0.5833015794	23.586983
30	30	3	race age side	0.60356190	0.5759034315	24.698165
36	31	3	fire age side	0.60071474	0.5728576283	25.155636
37	32	3	fire income side	0.59898427	0.5710064328	25.433681
34	33	3	fire theft side	0.59626819	0.5681008530	25.870091
35	34	3	fire age income	0.59226767	0.5638212283	26.512879
31	35	3	race income side	0.56916633	0.5391081678	30.224714
27	36	3	race theft income	0.56277709	0.5322731683	31.251313
28	37	3	race theft side	0.53150579	0.4988201493	36.275866
41	38	3	age income side	0.48258915	0.4464907154	44.135607
38	39	3	theft age income	0.46346751	0.4260350083	47.207999
40	40	3	theft income side	0.44957760	0.4111760383	49.439776
39	41	3	theft age side	0.32512689	0.2780427248	69.436042

```
42    42 4           race fire theft age 0.74719120  0.7231141763     3.620355
43    43 4        race fire theft income 0.69708918  0.6682405324    11.670558
44    44 4          race fire theft side 0.69609843  0.6671554205    11.829748
46    45 4            race fire age side 0.67439182  0.6433815196    15.317483
53    46 4           fire theft age side 0.67344873  0.6423486039    15.469016
45    47 4          race fire age income 0.67163867  0.6403661603    15.759849
52    48 4         fire theft age income 0.65630880  0.6235763081    18.222994
47    49 4         race fire income side 0.64429280  0.6104159269    20.153679
54    50 4        fire theft income side 0.64021861  0.6059537184    20.808304
55    51 4          fire age income side 0.62396240  0.5881492989    23.420290
48    52 4         race theft age income 0.62213571  0.5861486371    23.713796
49    53 4           race theft age side 0.61989214  0.5836913953    24.074284
51    54 4          race age income side 0.61056610  0.5734771587    25.572757
50    55 4        race theft income side 0.57346317  0.5328406102    31.534315
56    56 4         theft age income side 0.48280964  0.4335534098    46.100180
57    57 5    race fire theft age income 0.75082228  0.7204347561     5.036927
58    58 5      race fire theft age side 0.74736649  0.7165575266     5.592191
59    59 5   race fire theft income side 0.69928258  0.6626097280    13.318130
62    60 5    fire theft age income side 0.68019018  0.6411889822    16.385825
60    61 5     race fire age income side 0.67508111  0.6354568532    17.206731
61    62 5    race theft age income side 0.62372079  0.5778330852    25.459111
63    63 6 race fire theft age income side 0.75105211  0.7137099234     7.000000
```

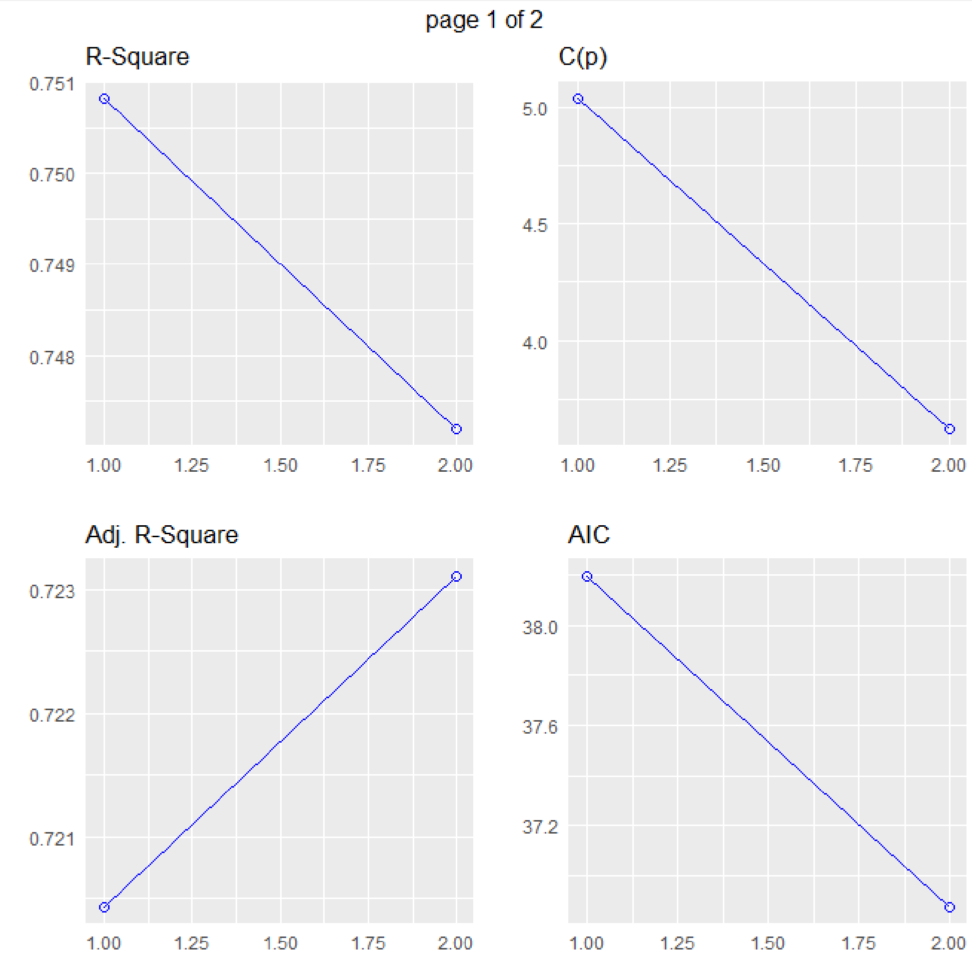
```
forward <- ols_step_forward_p(lmod)
plot(forward)
```



Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	race	0.5094	0.4985	35.8205	62.0324	0.4488
2	fire	0.6303	0.6135	18.4083	50.7436	0.3941
3	theft	0.6932	0.6718	10.2921	43.9701	0.3631
4	age	0.7472	0.7231	3.6204	36.8759	0.3335

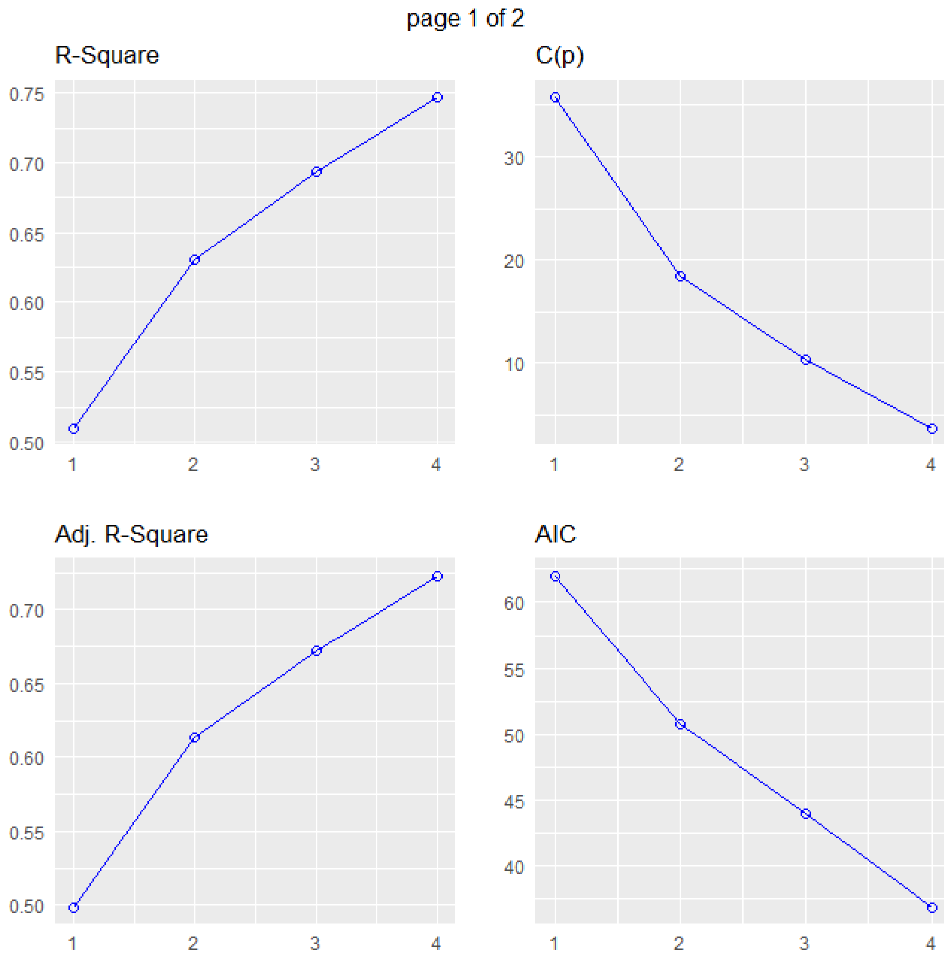
```
backward <- ols_step_backward_p(lmod)
plot(backward)
```



Elimination Summary

Step	Variable Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	side	0.7508	0.7204	5.0369	38.1959	0.3351
2	income	0.7472	0.7231	3.6204	36.8759	0.3335

```
stepwise <- ols_step_both_p(lmod)
plot(stepwise)
```



Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	race	addition	0.509	0.499	35.8210	62.0324	0.4488
2	fire	addition	0.630	0.613	18.4080	50.7436	0.3941
3	theft	addition	0.693	0.672	10.2920	43.9701	0.3631
4	age	addition	0.747	0.723	3.6200	36.8759	0.3335

Model validation with models that seem satisfactory using repeated 5 fold cross validation. Compare the model chosen from the forward/stepwise/backward and the model with the second lowest mallow's Cp

```
> modela <- lm(involact ~ race + fire + theft + age, data=chredlin)
> modelb <- lm(involact ~ race + fire + theft + age + income, data = chredlin)
>
> train.control <- trainControl(method="repeatedcv", number = 5, repeats = 10)
> cva <- train(involact ~ race + fire + theft + age, data=chredlin,
+                  method="lm",trControl = train.control)
> cvb <- train(involact ~ race + fire + theft + age + income, data=chredlin,
+                  method="lm", trControl = train.control)
> cva
Linear Regression

47 samples
 4 predictor

No pre-processing
Resampling: Cross-Validated (5 fold, repeated 10 times)
Summary of sample sizes: 38, 37, 39, 37, 37, 38, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.3511793  0.7171805  0.2706628

Tuning parameter 'intercept' was held constant at a value of TRUE
> cvb
Linear Regression

47 samples
 5 predictor

No pre-processing
Resampling: Cross-Validated (5 fold, repeated 10 times)
Summary of sample sizes: 38, 37, 38, 38, 37, 38, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.3646109  0.7204078  0.2817486

Tuning parameter 'intercept' was held constant at a value of TRUE
>
```

Discussion

When choosing a satisfactory model to predict the Chicago FAIR plan renewals to combat redlining, we took into consideration metrics adjusted R^2, Mallow's Cp, and RMSE. Having three metrics removes the bias of R^2 increasing when adding more regressors to the model and Mallow's Cp allows us to analyze the effect of bias on the model we choose. The initial full model yielded an adjusted R^2 of 0.7137 and Cp value of 7.0. This means that about 70% of the variation in the data can be explained by the model. And since p + 1 = 7, the Mallow's Cp shows no bias. However, when taking a deeper look at the regressors and their impact on the model, we realized that we did not need all regressors.

When performing forwards selection, backwards elimination, and stepwise selection, the result was the same with 4 regressors: race, fire, theft, and age. The RMSE of this model was 0.35. While this model was clearly the best model to use, since it was the same model selected by all three selection procedures and also included in the best subset regressions selection, we wanted to validate and compare this model to another model. The second lowest Mallow's Cp value close to the number of predictors in that model was the model that had all regressors except sides.

We conducted k-fold repeated cross validation with 5 folds and 10 repeats to eliminate error in resampling. We compared model A, the model with 4 regressors, and model B, the model with 5 predictors. Although the R^2 for model B was slightly greater than model A, we can attribute this to model B having a greater amount of regressors, which would make R^2 increase. This is true because when we look at both RMSE and MAE for model A, they are lower than model B, indicating model A has greater predictive power and fits the data better than model B.

Ultimately, we decided to go with model A as it yielded the greatest values for metrics RMSE, adjusted R^2, and Mallow's Cp when compared to other models in all possible regressions.

Works Cited

Chapter10.pdf, Lecture 10 (October 22 2022)

CrossValidationOverview.pdf, Lecture 10 (November 1 2022)

James, Gareth, et al. *An Introduction to Statistical Learning: with Applications in R* Springer, 2017.

Linear Models with R, Julian J. Faraway, (Taylor, 2nd, 2014)