ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

DEPARTMENT OF MANAGEMENT SCIENCE AND TECHNOLOGY

MSC BUSINESS ANALYTICS

Course: Statistics For Business Analytics

Tutor: I. Ntzoufras

Assignment: Communities and Crime Data Set

Dataset: 28

Full Name: Mallios Charalampos

Student ID: P2821912

Table of Contents

1. Introduction

The data of this assignment refer to crime characteristics of the USA for 1995. The main responses of the study are the number of crimes per 100K population rapes, using the characteristics of each area under consideration. The per capita violent rape crimes variable was calculated using population and the sum of crime variables considered violent crimes in the United States rape. In this particular analysis one of them will be examined: rapesPerPop. There was apparently some controversy in some states concerning the counting of rapes. These resulted in missing values for rape, which resulted in missing values for per capita violent crime. Many of these omitted communities were from the midwestern USA (Minnesota, Illinois, and Michigan have many of these). The aim of current study is to predict the rapes per 100,000 capita for each given state will be given.

1.1 Data Preparation and Data Cleaning

The given train dataset (number 28) contains 147 variables and 100 observations out of the main dataset. It includes some categorical variables that we are not going to use and will drop them from our dataset. These are: communityname, State, countyCode, communityCode. Regarding the remaining variables they are type of int and numeric so will transform them to numeric to be in the same data type. Our dataset contains missing values. For that reason, we are going to transform missing values to NAs. We are going to drop out these variables(columns) that having >30 NA from our dataset. As a next step we delete from our dataset these observations (rows) which are having missing values in the column of our response variable rapesPerPop. So ,the remaining dataset contains 93 observations (rows) and 103 variables (columns). Based on this dataset we are going to examine our analysis in the next sections.

2. Descriptive analysis and exploratory data analysis

In this part of the analysis we are going to explore in depth our given dataset. For example lets see some summary statistics for our response variable rapesPerPop and two independent:

rapesPerPop	whitePerCap	blackPerCap
Min. : 0.00	Min. : 6912	Min. : 0
1st Qu.: 18.83	1st Qu.:13069	1st Qu.: 6405
Median : 31.02	Median :15039	Median : 9106
Mean : 39.13	Mean :16804	Mean : 11523
3rd Qu.: 53.65	3rd Qu.:17144	3rd Qu.: 14139
Max. :148.32	Max. :63220	Max. :120000

Figure 2.1: Summary Statistics of rapesPerPop variables and 2 independents

At a first glance we can observe that mean and median are close enough for the variable rapesPerPop but not so close to assume our data being normal.There is a minimum 0 value for rapes per capita which means there are not recorded rapes in our dataset in these areas. If take a look at other variables we can see differences between white and black people in different areas. In some areas the Min for the black people per capita was zero which indicates that no black race people living there and of course can't be included in a rape type of crime. In other hand, their maximum value overall in our dataset was higher which means there are more places of high density of black people living there. Moreover, it is interesting to observe that per capita rapes that were happened are mainly in a range 0-50 and so there is skewness , which is clearly shown at the histogram below:

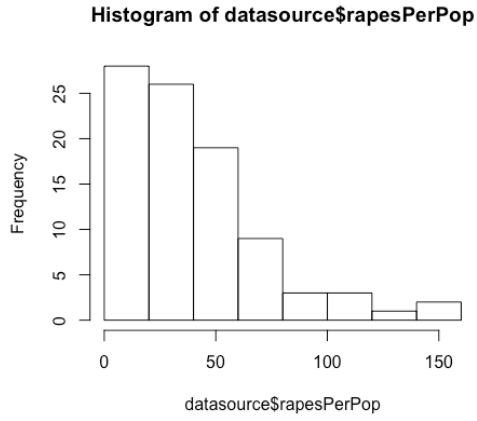


Figure 2.2 Histogram of per capita rapesPerPop

Trying to observe normality of response variable we present the below Q-Q plot which indicates that at first the values are trying to fit to normality and in the slope but at the end of the tail the values are missing the slope.

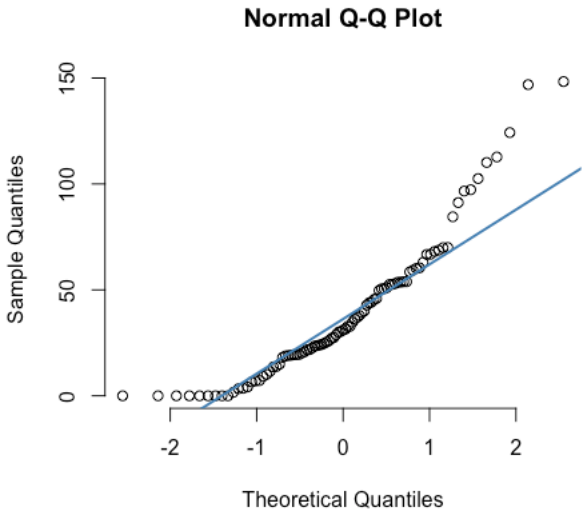


Figure 2.3 QQ plot of per capita rapesPerPop

3. Pairwise comparisons

In order to have a better view regarding correlations of our data, not only between them but also with our response variable we focus on correlations between rapesPerPop and the other independent variables. We choose to visualize only the ones that have positive correlation greater than 0.3 and negative lower than -0.3 with our response variable in order to have the most significant connections visualized.

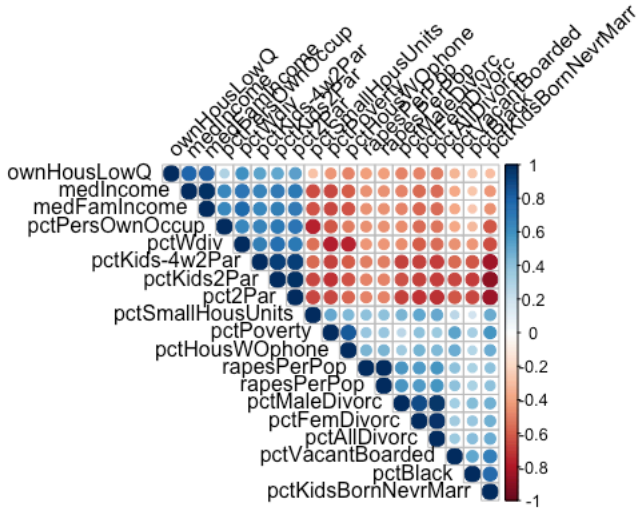


Figure 3.1: Correlation Plot between variables

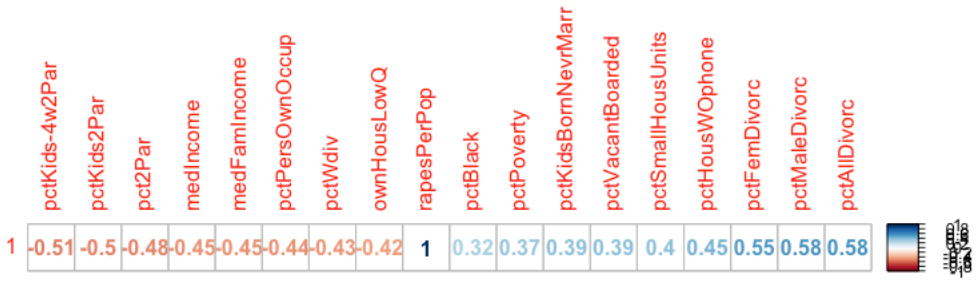


Figure 3.2 Correlation between of per capita rapesPerPop and the significant independents

As we can see there is high positive correlation between rapes and divorces. This means that if a man will get divorced it seems more likely to get involved in a rape crime. In the other hand there is negative correlation between rapes and percentage of kids in family housing with two parents. The higher the percentage of kids living with their two parents the lower will the percentage of rapes be.

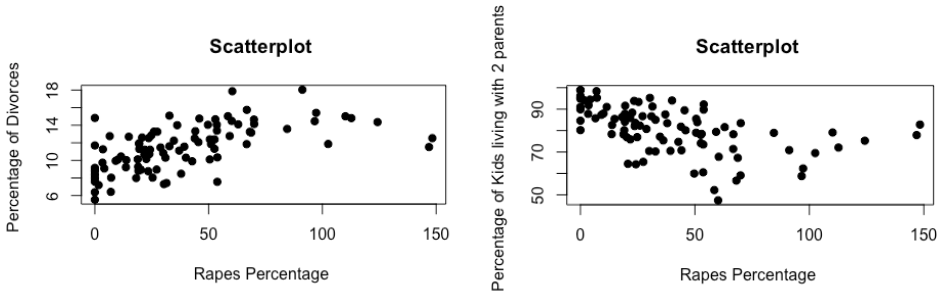


Figure 3.3 Scatterplot rapesPerPop vs Divorces Percentage

The scatterplots confirm the above. The values are scattered in a positive and negative type of slope respectively.

4. Predictive Descriptive models

To begin with, we start reading the dataset given (data28) having done the appropriate cleaning and preparation as we described in section 1. What we have tried firstly is to get the full model out of sample and see the summary of the model. However, the full model appears to have singularities and coefficients not defined because of singularities Some of the variables include in our full model (10 of them) have collinearity between them(ownHousQrange,rentQrange,pctBornStateResid,pctSameHouse.5pctSameState.5,landArea,pctUsePubTrans,pctSameCounty.5,popDensity,pctOfficDrugUnit).It's then needed then to delete them in order to proceed with the next step.

4.1 Lasso variable screening

Now we have cleared singularities we proceed with Lasso variable selection as our primary variable selection method. When we use Lasso, we try to find a lambda that penalizes the predictors in order the final model to have the predictors with mean squared errors. We will get to types of lambda, both min and 1 standard error lamba and we will interpret the results.

lasso1$lambda.1se
14.37605
lasso1$lambda.min
5.670209

Table 4.1 λ 1SE and λ Min values

Running the lasso with 1 standard error λ we interpret the coefficients and get the variables of pctMaleDivorc, pctAllDivorc and constant one.

```
> names(which(variablesoflasso1se[,1]>0))
[1] "(Intercept)"   "pctMaleDivorc" "pctAllDivorc"
```

Figure 4.1.2 variables screening, Lasso with 1 SE

In the other hand running lasso with minimum λ we get as a result the following: constant,pctMaleDivorc,pctAllDivorc,pctVacantBoarded,pctVacant6up,pctHousWOhone,persHomeless:

```
> names(which(variablesoflassomin[,1]>0))
[1] "(Intercept)"      "pctMaleDivorc"    "pctAllDivorc"     "pctVacantBoarded"
[5] "pctVacant6up"     "pctHousWOphone"   "persHomeless"
```

Figure 4.1.3 variables screening, Lasso with min λ

The summary of our two models is described below:

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.034 13.369 -3.144 0.00226 **
pctAllDivorc 4.508 4.133 1.091 0.27830
pctMaleDivorc 3.006 4.240 0.709 0.48017
Residual standard error: 26.94 on 90 degrees of freedom
Multiple R-squared: 0.3419, Adjusted R-squared: 0.3273
F-statistic: 23.38 on 2 and 90 DF, p-value: 6.636e-09

Table 4.1.4 Summary of 1SE lasso model.

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -53.21200 18.60703 -2.860 0.00532 **
persHomeless 0.10971 0.06979 1.572 0.11963
pctHousWOphone 1.62513 0.87734 1.852 0.06741 .
pctVacant6up 0.53132 0.28149 1.888 0.06246 .
pctVacantBoarded -0.22482 1.04986 -0.214 0.83095
pctAllDivorc 2.96555 4.57844 0.648 0.51889
pctMaleDivorc 3.36641 4.42473 0.761 0.44885
Residual standard error: 25.23 on 86 degrees of freedom
Multiple R-squared: 0.4486, Adjusted R-squared: 0.4101
F-statistic: 11.66 on 6 and 86 DF, p-value: 1.609e-09

Table 4.1.5 Summary of min λ lasso model.

As we can observe for 1 standard error Lasso the Adjusted R-squared is 0.32 which is not so good and 0.41 for minimum lambda model which seems to be better in terms of how the proposed model does improve prediction over the mean mode.

None of our models seems to perform well in terms of R-squared but the have a p-value <0.05 which means that they perform better than the constant mode (if we had only the constant variate). Most of the variables selection (screening) are not statistical significant so we are going to apply another additive method to provide a more clear model in terms of performance.

4.2 Stepwise method for variable selection

After we have examined Lasso method, we didn't get what we need to get out final model. We will get the 2 models we have created above and we are going to run stepwise method, with both directions for variable selection (adding and deleting each variable in each step based on AIC score).

Step: AIC=603.61			
rapesPerPop ~ persHomeless + pctHousWOphone + pctVacant6up +			
pctMaleDivorc			
	Df	Sum of Sq	RSS AIC
<none>			55024 603.61
- persHomeless	1	1684.1	56708 604.42
+ pctAllDivorc	1	244.3	54780 605.20
+ pctVacantBoarded	1	6.4	55018 605.60
- pctVacant6up	1	2467.7	57492 605.69
- pctHousWOphone	1	3755.6	58780 607.75
- pctMaleDivorc	1	18836.2	73860 628.99

Table 4.2.1 Stepwise with AIC on lasso with min λ

Step: AIC=614.11			
rapesPerPop ~ pctAllDivorc			
	Df	Sum of Sq	RSS AIC
<none>			65707 614.11
+ pctMaleDivorc	1	365	65342 615.60
- pctAllDivorc	1	33588	99295 650.51

Table 4.2.2 Stepwise with AIC on lasso with 1 St.error λ

We can clearly observe that the process ends up with a slightly better AIC for min λ model (603.61) than for 1SE λ model and with more variables included as for the linear model representation. The lower the AIC , the less the information we are going to lose after the stepwise method. The stepwise AIC method has cleared up pctMaleDivorc, hodling only pctAllDivorc which is the highest correlated variable with our response

as we saw in section 1. In the other hand for the minimum lamda model the variables that finally ends our model are the constant and 4 independent :

persHomeless | pctHousWOphone | pctVacant6up | pctMaleDivorc

Table 4.2.3 Final variable selection of stepwise with AIC on lasso with min λ

We will get the anova results for the models to support the choose of one of them:

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -45.028 12.650 -3.56 0.000593 ***
pctAllDivorc 7.336 1.076 6.82 9.75e-10 ***
Residual standard error: 26.87 on 91 degrees of freedom
Multiple R-squared: 0.3383, Adjusted R-squared: 0.331
F-statistic: 46.52 on 1 and 91 DF, p-value: 9.749e-10

Table 4.2.4 Summary of AIC 1SE lasso model.

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -44.75734 13.01239 -3.440 0.000893 ***
persHomeless 0.10380 0.06325 1.641 0.104333
pctHousWOphone 1.83547 0.74893 2.451 0.016232 *
pctVacant6up 0.44635 0.22468 1.987 0.050078 .
pctMaleDivorc 6.10840 1.11293 5.489 3.86e-07 ***
Residual standard error: 25.01 on 88 degrees of freedom
Multiple R-squared: 0.4459, Adjusted R-squared: 0.4207
F-statistic: 17.7 on 4 and 88 DF, p-value: 1.081e-10

Table 4.2.5 Summary of AIC min λ lasso model.

The model with min λ has better Adjusted R-squared: 0.4207 which it is interpreted as the proportion of total variance that is explained by the model. P- value is also Statistical significant for both of them, which means that the models are better than the constant and the variables are statistical significant except of persHomeless which we are going to drop it out of our final model (min λ lasso model)since it has also better performance in terms of residual standard error(25.01 < 26.87). If we delete the

persHomeless variable we do get only statistical significant covariates as see clearly below :

Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-47.0055	13.0626	-3.598	0.000526 ***
pctHousWOphone	1.9929	0.7498	2.658	0.009320 **
pctVacant6up	0.4471	0.2268	1.971	0.051800 .
pctMaleDivorc	6.3829	1.1107	5.747	1.25e-07 ***
Residual standard error: **25.24** on 89 degrees of freedom				
Multiple R-squared: 0.4289, Adjusted R-squared: **0.4096**				
F-statistic: 22.28 on 3 and 89 DF, **p-value: 7.528e-11**				

Table 4.2.5 Final model Summary.

The p-value for the model remains statistical significant (<0.05) which means is still better than the one only with the constant. The adjusted R-squared was slightly increased and the Residual standard error slightly increased. The only one problem appearing here is the constant variable for the min λ model which appears to be statistical significant but it is negative and affects in high tension our model. To use this model for our analysis we will check the assumptions trying not to violate them.

4.3 Checking for assumptions

We have to check if 4 assumptions are satisfied and those are :

1. Normality of errors
2. Homoscedasticity of errors
3. Independence of variances
4. Linearity of residuals

We will check these assumption with studentized residuals, dedicated tests and with the help of plots.

For **linearity** assumption, we will the result of Tukey test of residualplots and check if linearity is violated. The p-value of Tukey test is > 0.05 so the linearity of residuals **is not violated.**

	Test stat	Pr(>|Test stat|)
pctHousWOphone	-0.9399	0.3498
pctVacant6up	1.2851	0.2021
pctMaleDivorc	0.2743	0.7845
Tukey test	1.2702	**0.2040**

Table 4.3.1 Tukey test p-value = 0.2040

For **normality** of error we are going to investigate it, running the Kolmogrov-Smirnof test and see if the p-value do not reject the H0 hypothesis, that the residuals of errors are normal. The p-value is >0.05 which means we can not reject the Hypothesis 0 , that the errors are normal. Normality for model **is not violated**.

Lilliefors (Kolmogorov-Smirnov) normality test
data: residuals(final)
D = 0.084734, **p-value = 0.09623**

Table 4.3.2 Kolmogorov-Smirnov test p-value = 0.09623

As for the **Homoscedasticity** the ncvtest shows for the model that there we do not have constant variance. When Homoscedasticity assumption is violated the error variance estimator is not estimated correctly and so do the standard errors and the assumption is being violated.

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 8.440905, Df = 1, **p = 0.0036687**

Table 4.3.3 Non-constant Variance Score test p-value = 0.0036687

From the Darwin-Watson test we can see that p-value is > 0.05 for the models, H0 is not rejected, there is independence of errors, and the independence of residuals assumption **is not being violated** for both of the models.

Durbin-Watson test
data: final
DW = 2.2996, **p-value = 0.8933**
alternative hypothesis: true autocorrelation is greater than 0

Table 4.3.4 Durbin-Watson test p-value = 0.8933

For all the assumptions above the only one being violated is the homoscedasticity of errors from the constant variance test. If we use logarithms to interpret our model we could probably solve the problem but it would be more difficult to interpret the results so we proceed with one violation at this point of analysis.

4.4 Interpreting the model

So we have ended up having one model for interpretation and conduct analysis based on their predictive performance it has:

pctrapesPerPop= -47.0055 + 1.9929 * pctHousWOphone + 0.4471 * pctVacant6up + 6.3829 * pctMaleDivorc + ε ~N (0,25.24²)

Table 4.4.1 Final Model

Firstly, starting from the independent variables we can focus on pctHousWOphone. PctHousWOphone indicates percent of occupied housing units without phone. Beacause of its coefficient (1.9929) is points out that the probability of rapes to increased are higher if the houses do not have phone. This seems logical because not having phone indicates poor people which would be more possible to get involved in a crime. The coefficient also is increasing twice and so does the possibility a rape to be increased.

Regarding pctVacant6up that it has not so big coefficient indicates that if someone stays up to 6 months in a vacant it is most possible to get involved in rapes since it may be someone that changes vacation often, which indicates someone who doesn't want to be known in the neighborhood.

Last but not least the pctMaleDivorc which has a very big coefficient indicates how much positive correlated males with a divorce are with the rape crimes. For every per capita increment on Male divorces the rapes percentage will be increased almost 7 times which is huge but also seems logical. Males which are divorced are more likely to get involved in a rape as the criminal.

We observe that the numbers per capita rapes start with a constant negative big number because of the constant variable which indicates that if the coefficents of the other variables are 0 the pctrapesPerPop will propable get a negative value plus the error and this is not so good for our model in terms of predictive ability.

4.5 Cross validation of model

For our model we examined cross-validation with 10 folds on our train dataset to conduct analysis based on the performance of the model. The RMSE is 25.28 close to 25.25 RMSE before cross-validation which means that cross validation didn't affect the residual errors of linear regression.

Linear Regression		
93 samples		
3 predictor		
No pre-processing		
Resampling: Cross-Validated (10 fold)		
Summary of sample sizes: 85, 81, 84, 85, 83, 83, ...		
Resampling results:		
RMSE	Rsquared	MAE
25.28651	0.485482	19.75259

Table 4.5.1 Cross Validation

But to identify the importance of RMSE we introduce the metric called scatter index (SI) to judge whether RMSE is good or not. SI is RMSE normalised to the measured data mean or SI=RMSE/measured data mean. If SI is less than one, your estimations are acceptable. Our SI is less than 1 so our estimations are acceptable.

Scatter index
SI <- mean(modelt$resample$RMSE)/mean(dat28$rapesPerPop)
0.6462323

Table 4.5.2 Scatter Index

4.6 Predicting performance

If we want to identify the predictive performance of our model, we must examine the mean absolute error of the difference between observed data on test dataset and predicted data from our model. We find out the below that their difference is less than our residual standard error so its predictive performance is acceptable:

Prediction Evaluation: mean(abs(crimesdatatest$rapesPerPop - predict(final,crimesdatatest)))
Result : 19.83444

Table 4.6.1 Prediction evaluation

5. Further analysis

Now we have chosen the final model of our examination we can dive into selection of some scenarios of the areas given at the test dataset. We should answer to questions like which is like a typical profile of an area how is the best and the worst area in terms of per capita rapes.

We can get the three scenarios: one of typical profile of an area, which means the mean per capita rapes, if we get the centralized data of the independent variables for our response. Since it is related with pctHousWOphone, pctVacant6up and pctMaleDivorc we need to get mean for them and create the typical area profile.

Typical Profile Area
mean(crimesdatatest$pctHousWOphone)
mean(crimesdatatest$pctVacant6up)
mean(crimesdatatest$pctMaleDivorc)

Table 5.1 Typical Area Profile

predict(final,typical)
rapesPerPop : 34.70566

Table 5.2 prediction for Typical Area Profile

The worst area is that having the higher number of per capita pares so we have to compute the maximum values for our independent variables. The best area is that having the lower number of per capita pares so we have to compute the minimum rapesPerPop from minium values from our independent values.

Worst Area
max(crimesdatatest$pctHousWOphone)
max(crimesdatatest$pctVacant6up)
max(crimesdatatest$pctMaleDivorc)

Table 5.3 Worst Area Profile

predict(final,worst)
rapesPerPop : 106.7531

Table 5.4 prediction for Worst Area Profile

Best Area
min(crimesdatatest$pctHousWOphone)
min(crimesdatatest$pctVacant6up)
min(crimesdatatest$pctMaleDivorc)

Table 5.5 Best Area Profile

predict(final,best)
rapesPerPop : -24.61319

Table 5.6 prediction for best Area Profile. Negative is because of the constant commitment

It would be also interesting to compute the overall estimated crimes per capita from our model fitting on the test dataset. To get this information we must consider how well our model fits our data, examining the fitted values and get the mean value.

mean(final$fitted.values) : 39.12914

Table 5.7 Overall estimation

Since the mean is very close to the mean of the test data set this is an indicator that our model fits well enough to our data. We can also compute from our fitted data the sum per capita rapes :

sum(final$fitted.values) : 3639.01

Table 5.8 Sum estimation

6. Conclusions and Discussion

Having conducted the above analysis we have ended up with a model that fits well on our data and having a good predictive ability but containing a negative feature. Our model is very sensitive to outliers since we have violated constant variance in our assumptions and since our constant variable has great impact with its negative high value. In further analysis we could try use logarithms to adjust the variance of our model and have a better model not being so sensitive in outliers.

7. Appendix

References

Nash, J.E., and Sutcliffe, J.V. 1970. River flow forecasting through conceptual models part I — A discussion of principles, Journal of Hydrology, 10(3), 282–290.

Ris, R.C., Holthuijsen, L.H., and Booij, N. 1999. A third-generation wave model for coastal regions 2, verification. Journal of Geophysical Research, 104(C4) 7667-7681.

Willmott, C.J., Ackleson, S.G., Davis, R.E., Feddema, J.J., Klink, K.M., Legates, D.R., O'Donnell, J., and Rowe, C.M. 1985. Statistics for the evaluation and comparison of models, Journal of Geophysical Research, 90(C5), 8995–9005.

Zambreskey, L., 1988. A verification study of the global WAM model, December 1987 – November 1988. GKSS Forschungzentrum Geesthacht GMBH Report GKSS 89/E/37.

David A. Freedman (2009). Statistical Models: Theory and Practice. Cambridge University Press. p. 26. A simple regression equation has on the right hand side an intercept and an explanatory variable with a slope coefficient. A multiple regression equation has two or more explanatory variables on the right hand side, each with its own slope coefficient

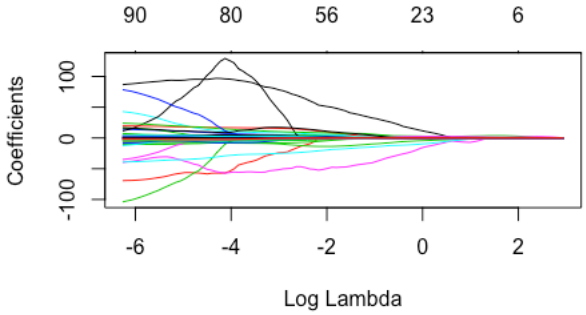
Rencher, Alvin C.; Christensen, William F. (2012), "Chapter 10, Multivariate regression – Section 10.1, Introduction", Methods of Multivariate Analysis, Wiley Series in Probability and Statistics, 709 (3rd ed.), John Wiley & Sons, p. 19, ISBN 9781118391679.

Hilary L. Seal (1967). "The historical development of the Gauss linear model". Biometrika. 54 (1/2): 1–24. doi:10.1093/biomet/54.1-2.1. JSTOR 2333849.

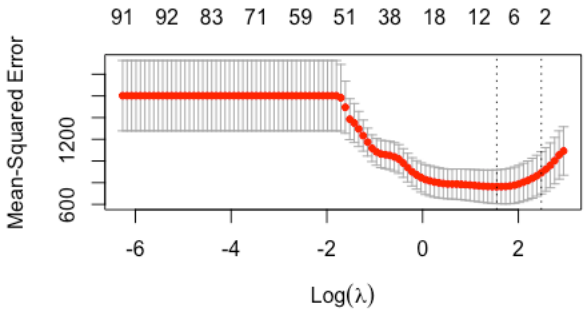
Yan, Xin (2009), Linear Regression Analysis: Theory and Computing, World Scientific, pp. 1–2, ISBN 9789812834119, Regression analysis ... is probably one of the oldest topics in mathematical statistics dating back to about two hundred years ago. The earliest form of the linear regression was the least squares method, which was published by Legendre in 1805, and by Gauss in 1809 ... Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the sun.

Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the Lasso". Journal of the Royal Statistical Society, Series B. 58 (1): 267–288

Additional Figures



Log lambda grows bigger all coefficients turn into zero



Mean square error and the log(lamda) that grow in parallel in positive values of λ

Code in R:

```
#reading data that has been transformed locally in weka
#for question 2 and above we make it manually to present how it is
done with R
datast <- read.table(file.choose(),header = TRUE,sep = ";")
#delete singularites columns,columns with NAs
#target variables except ours
# the first 5 columns which are not interest us
#delete also the rows from rapes perPop contain NA
#93*103 dataframe : datast
datast[,1:103]<-lapply(datast[,1:103], as.numeric)
#summary statistics

########################## Question
1###################################3
summary(datast)
#qqplot for our response
library("car")
qqPlot(datast$rapesPerPop)
qqnorm(datast$rapesPerPop, pch = 1, frame = FALSE)
qqline(datast$rapesPerPop, col = "steelblue", lwd = 2)
#all corellations with our response
```

```
corr<-as.data.frame(cor(datast$rapesPerPop,datast[,1:102]))
sort_corr<-sort(corr)
#split to negative and postive correlations most significant
negative_cor<-sort_corr[,c(1:8)]
positive_cor<-sort_corr[,c(94:102)]
#identify the positive and negative correlations most significant
nd<-datast[c("rapesPerPop","pctKids-4w2Par"
,"pctKids2Par","pct2Par","medIncome","medFamIncome",
"pctPersOwnOccup","pctWdiv","ownHousLowQ"
)]
#pairwise comparisons
pairs(nd)
pd<-
datast[c("rapesPerPop","pctBlack","pctPoverty","pctKidsBornNevrMarr",
"pctVacantBoarded","pctSmallHousUnits",

"pctHousWOphone","pctFemDivorc","pctMaleDivorc","pctAllDivorc")]
#correlations plots between positive and negative
corrplots<-c(nd,pd)
pairs(pd)
#correlations
corrplots<-datast[c("rapesPerPop","pctKids-4w2Par"
,"pctKids2Par","pct2Par","medIncome","medFamIncome",
"pctPersOwnOccup","pctWdiv","ownHousLowQ","rapesPerPop","pctBlack","p
ctPoverty","pctKidsBornNevrMarr","pctVacantBoarded","pctSmallHousUnit
s","pctHousWOphone","pctFemDivorc","pctMaleDivorc","pctAllDivorc")]
#correleogram
res <- cor(corrplots)
round(res, 2)
library(corrplot)
#in use correologram
corrplot(cor(corrplots$rapesPerPop,corrplots[2:19]), method =
"number")
#pairs of postive correlatiosn
pairs(pd[,2:7], pch = 19,  cex = 0.5,
      lower.panel=NULL)

par(mfrow=c(5,5))
library(car)
#scatterplots for response and divorces
plot(datast$rapesPerPop, datast$pctAllDivorc, main="Scatterplot ",
     xlab="Rapes Percentage ", ylab="Percentage of Divorces ",
pch=19)

plot(datast$rapesPerPop, datast$`pctKids-4w2Par`, main="Scatterplot
",
     +      xlab="Rapes Percentage ", ylab="Percentage of Kids living
with 2 parents ", pch=19)
par(mfrow=c(4,4)); n <- nrow(datast)
#histograms
hist(datast$pctBlack)
hist(datast$rapesPerPop)
hist(negative_cor[,2], main=names(negative_cor)[2])
hist(negative_cor[,3], main=names(negative_cor)[4])
hist(negative_cor[,4], main=names(negative_cor)[4])
plot(table(datast[,1])/n, type='h', xlim=range(datast[,1])+c(-1,1),
main=names(datast)[3], ylab='Relative frequency')
plot(table(negative_cor[,2])/n, type='h',
xlim=range(negative_cor[,2])+c(-1,1), main=names(negative_cor)[4],
ylab='Relative frequency')
```

```
plot(table(negative_cor[,3])/n, type='h',
xlim=range(negative_cor[,3])+c(-1,1), main=names(negative_cor)[3],
ylab='Relative frequency')
plot(table(negative_cor[,4])/n, type='h',
xlim=range(negative_cor[,4])+c(-1,1), main=names(negative_cor)[4],
ylab='Relative frequency')

########################## Question 2 ##########################
####################data preparation and cleaning##################
############ in this part we are doing it manually in R
########################
setwd("/Users/xarismallios/Desktop")
crimes_28<- read.csv(file="crime_28.csv", header=TRUE, sep=",")
df<-crimes_28
#We dont need first 4 columns refering states and community
df[,1:4]<-NULL
#we drop columns with many NAs appeared
df[,130:143]<-NULL
df[,126:128]<-NULL
dar<-df
dar[,127]<-NULL
dar[,125]<-NULL
dar[,120:123]<-NULL
dar[,100:116]<-NULL
#delete rows from response where there is NA
dat28<-subset(dar, !is.na(dar$rapesPerPop))

#full model with all variables
fullmodel <- lm(rapesPerPop~.,data=dat28)
#drop the singularities which cause NAs
dat28$ownHousQrange<-NULL
dat28$rentQrange<-NULL
dat28$pctBornStateResid<-NULL
dat28$pctSameHouse.5<-NULL
dat28$pctSameState.5 <-NULL
dat28$landArea<-NULL
dat28$pctUsePubTrans<-NULL
dat28$pctSameCounty.5<-NULL
dat28$popDensity<-NULL
dat28$pctOfficDrugUnit<-NULL
#full model without singularities
fullmodel_nona <- lm(rapesPerPop~.,data=datas28)
datas28<-dat28
datas28[,1:93]<-lapply(dat28[,1:93], as.numeric)

#lasso to fullmodel without singularities
require(glmnet)
X <- model.matrix(fullmodel_nona)[,-1]
lasso1 <- glmnet(X, dat28$rapesPerPop)
lasso1$lambda
#cross validation lasso methos
lasso1 <- cv.glmnet(X, dat28$rapesPerPop)
# lasso with 1 standard error λ
lasso1$lambda.1se
# lasso with min λ
lasso1$lambda.min
coef(lasso1, s = "lambda.1se")
coef(lasso1, s = "lambda.min")
variablesoflasso1se<-coef(lasso1, s = "lambda.1se")
variablesoflassomin<-coef(lasso1, s = "lambda.min")
#get variables screening
```

```
names(which(variablesoflasso1se[,1]>0))
names(which(variablesoflassomin[,1]>0))
#plot lasso
plot(lasso1)
plot(lasso1$glmnet.fit, xvar = "lambda")
abline(v=log(c(lasso1$lambda.min, lasso1$lambda.1se)), lty =2)

# model lasso with 1 SE
fullmodel_lasso_1se<-
lm(rapesPerPop~pctAllDivorc+pctMaleDivorc,data=dat28)
# model lasso with min lamda
fullmodel_lasso_min<-
lm(rapesPerPop~+persHomeless+pctHousWOphone+pctVacant6up+pctVacantBoa
rded+pctAllDivorc+pctMaleDivorc,data=dat28)
#summary for both
summary(fullmodel_lasso_1se)
summary(fullmodel_lasso_min)

# model lasso with min lamda and BIC
fullmodel_stepbic_lasso_min<-step(fullmodel_lasso_min,direction =
"both",k=log(93))
# model lasso with 1 SE and BIC
fullmodel_stepbic_lasso_1se<-step(fullmodel_lasso_1se,direction =
"both",k=log(93))
#summary for both
summary(fullmodel_stepbic_lasso_min)
summary(fullmodel_stepbic_lasso_1se)
# model lasso with min lamda and AIC
fullmodel_stepaic_lasso_min<-step(fullmodel_lasso_min,direction =
"both")
# model lasso with 1 SE and AIC
fullmodel_stepaic_lasso_1se<-step(fullmodel_lasso_1se,direction =
"both")
#summary for both
summary(fullmodel_stepaic_lasso_min)
summary(fullmodel_stepaic_lasso_1se)
#final best model we select
final<-lm(rapesPerPop ~ + pctHousWOphone + pctVacant6up +
            + pctMaleDivorc,data=dat28)
summary(final)
########################## Question 3 ##########################

#assumptions tests
library(nortest)
library(car)
#linearity test, Tukey
residualPlot(final, type="rstudent")
residualPlots(final, plot=F)
plot(rstudent(final), type="l")
#normality test Kolmogorov-Smirnov
lillie.test(residuals(final))
#constant variance test homoscedsticity
ncvTest(final)
library(lmtest)
#independence test
dwtest(final)
########################## Question 4 ##########################

#cross validation using 10 folds
library(caret)
train.control <- trainControl(method = "cv", number = 10)
```

```
# Train the model
modelt <- train(rapesPerPop ~ pctHousWOphone + pctVacant6up +
                  + pctMaleDivorc, data = dat28, method = "lm",
                trControl = train.control)
# Summarize the results
print(modelt)
########################### Question 5,6,7 ###########################

#scatter index
SI<-mean(abs(final$fitted.values - predict(final,crimesdatatest)))
predict(final,dat28)

########################### Question 8###########################

# the three scenarios
typical <-data.frame("pctHousWOphone" =
mean(crimesdatatest$pctHousWOphone) ,"pctVacant6up" =
mean(crimesdatatest$pctVacant6up) ,"pctMaleDivorc" =
mean(crimesdatatest$pctMaleDivorc))
worst <-data.frame("pctHousWOphone" =
max(crimesdatatest$pctHousWOphone) ,"pctVacant6up" =
max(crimesdatatest$pctVacant6up) ,"pctMaleDivorc" =
max(crimesdatatest$pctMaleDivorc))
best <-data.frame("pctHousWOphone" =
min(crimesdatatest$pctHousWOphone) ,"pctVacant6up" =
min(crimesdatatest$pctVacant6up) ,"pctMaleDivorc" =
min(crimesdatatest$pctMaleDivorc))

#3 scenarios of typical profile prediction, max , min
predict(final,typical)
predict(final,worst)
predict(final,best)

# overall estimation and sum
mean(final$fitted.values)
sum(final$fitted.values)
```