Assignment 5 Part I Work

Matthew Dunne

August 12, 2018

The Data

Use the same training and holdout data sets that you used for logistic regression. Use only the numeric variables.

```
setwd("C:/Users/mjdun/Desktop/Data Mining/Assignments")
#using csv which has it in factor variables
MyData <- read.csv(file="German.Credit.csv", header=TRUE, sep=",")
#select just the numeric variables, make Amount the first variable
MyData<-MyData[,c(6,3, 9, 12,14, 17, 19)]
library(caret)
#split data into train and test as you did in logistic regression assignment
set.seed(1234)
s1<-sample(1:nrow(MyData), nrow(MyData)*.7, replace=FALSE)
train<-MyData[s1, ]
test<-MyData[-s1, ]
```

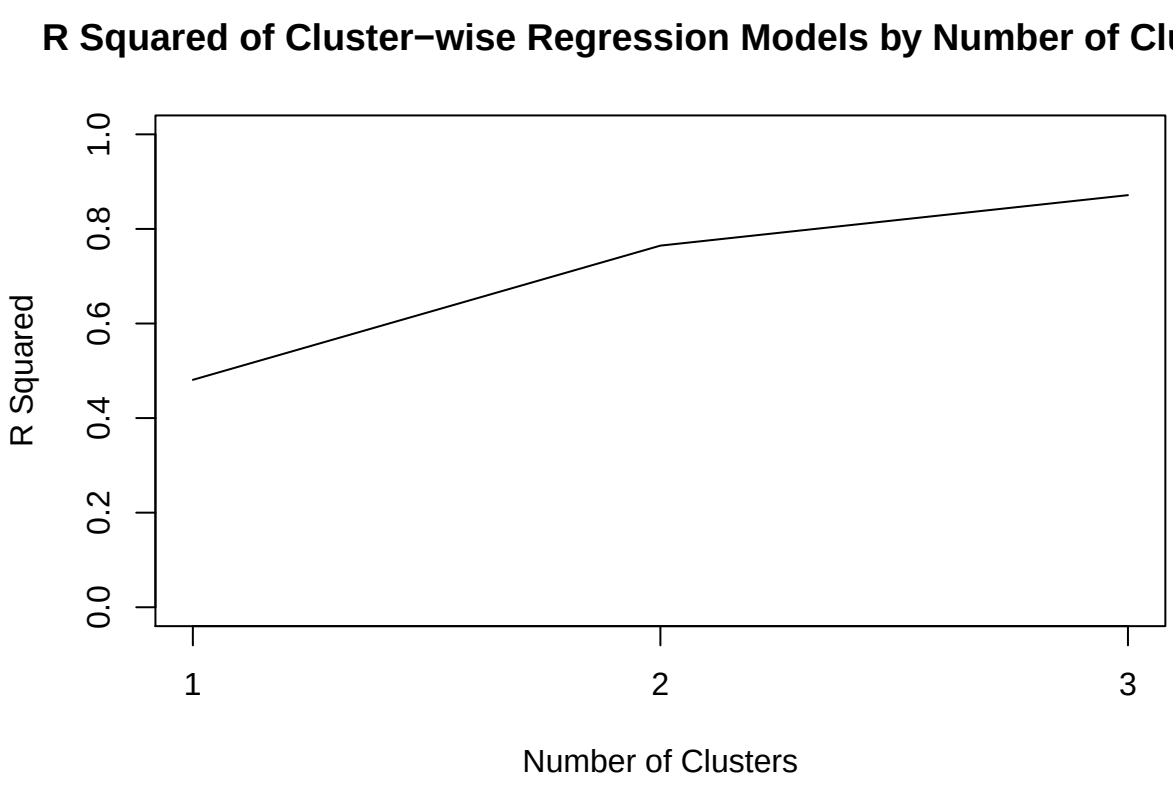
Clusterwise Regression Model

Build a clusterwise regression model with Amount as the dependent variable and the numeric variables as independent variables. Build 1, 2, and 3 cluster solutions.

```
source("clustreg.R")
source("clustregpredict.R")
#for one cluster, one random start, one iteration in each random start (because there is only one clust
clust.reg.1<-clustreg(train, k=1, tries=1, sed=12345, niter=1)
#for two clusters, two random starts, tend iterations in each random start
clust.reg.2<-clustreg(train, k=2, tries=2, sed=12345, niter=10)
#for three clusters
clust.reg.3<-clustreg(train, k=3, tries=2, sed=12345, niter=10)
```

And then plot R Squared as a function of the number of clusters.

```
plot(c(1,2,3), c(clust.reg.1$rsq.best,clust.reg.2$rsq.best, clust.reg.3$rsq.best), type = "l", ylim = c
axis(side = 1, at=c(1,2,3), labels = c(1,2,3))
```




```
train.rsq<-cbind(c("1 Cluster"=clust.reg.1$rsq.best,"2 Clusters"=clust.reg.2$rsq.best, "3 Clusters"=clus
train.rsq
```

```
##                  [,1]
## 1 Cluster  0.4808622
## 2 Clusters 0.7646923
## 3 Clusters 0.8714568
```

In terms of R Squared the 3 Cluster model is probably best. On the chart the line between 2 and 3 clusters looks relatively flat, making it seem not much in R Squared is added by adding the extra cluster, but this is deceiving. Moving from 0.76 to 0.87 is significant and should not result in much loss of interpretability.

Validation

Perform holdout validation testing of the cluster-wise regression using clustreg.predict().

```
predict.cluster1<-clustreg.predict(clust.reg.1, test)
predict.cluster2<-clustreg.predict(clust.reg.2, test)
predict.cluster3<-clustreg.predict(clust.reg.3, test)
test.rsq<-cbind(c("1 Cluster"=predict.cluster1$rsq,"2 Clusters"=predict.cluster2$rsq, "3 Clusters"=predi
test.rsq
```

```
##                  [,1]
## 1 Cluster  0.5452174
## 2 Clusters 0.7485304
## 3 Clusters 0.8579283
```

Choosing a Model

Let us compare the R Squared values from the train and holdout data side by side.

```
data.frame("Train R Sq."=train.rsq, "Holdout R Sq."=test.rsq)
```

```
##            Train.R.Sq. Holdout.R.Sq.
## 1 Cluster    0.4808622     0.5452174
## 2 Clusters   0.7646923     0.7485304
## 3 Clusters   0.8714568     0.8579283
```

The two and three cluster models hold up very well, losing only a small amount of R Squared. The R Squared for the one cluster model actually goes up in the holdout data, but we are choosing between the two and three cluster models anyway, deciding whether the increase in R Squared is worth the loss in interpretability if any.

So we have to look at what those clusters are made of. We will use the clusters made using the training data.

First the two cluster model:

```
#the first cluster
clust.reg.2$results[[1]]
```

```
##
## Call:
## lm(formula = dat[c.best == i, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3533.4  -876.4  -290.0   504.0  8407.7
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -2299.481    608.274  -3.780 0.000198 ***
## Duration.of.Credit..month.     305.265      9.580  31.864  < 2e-16 ***
## Instalment.per.cent           -303.206     96.280  -3.149 0.001848 **
## Duration.in.Current.address     -3.132    105.141  -0.030 0.976262
## Age..years.                     58.526     10.402   5.626  5.2e-08 ***
## No.of.Credits.at.this.Bank    -361.652    189.853  -1.905 0.058008 .
## No.of.dependents               432.845    308.078   1.405 0.161341
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1669 on 236 degrees of freedom
## Multiple R-squared:  0.8189, Adjusted R-squared:  0.8143
## F-statistic: 177.9 on 6 and 236 DF,  p-value: < 2.2e-16
```

```
#the second cluster
clust.reg.2$results[[2]]
```

```
##
## Call:
## lm(formula = dat[c.best == i, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3368.2  -651.7   -83.1   483.1 11697.4
##
## Coefficients:
```

```
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                3197.9461   344.9118   9.272   <2e-16 ***
## Duration.of.Credit..month.   97.8084     4.7413  20.629   <2e-16 ***
## Instalment.per.cent        -799.3666    55.1159 -14.503   <2e-16 ***
## Duration.in.Current.address  -9.3902    54.2727  -0.173    0.863
## Age..years.                   0.5543     5.3209   0.104    0.917
## No.of.Credits.at.this.Bank  -22.4742   100.6255  -0.223    0.823
## No.of.dependents            -99.2938   154.3842  -0.643    0.520
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1216 on 450 degrees of freedom
## Multiple R-squared:  0.5756, Adjusted R-squared:   0.57
## F-statistic: 101.7 on 6 and 450 DF,  p-value: < 2.2e-16
```

For both of these clusters we see that Amount up goes as the Duration of the Loan increases. This makes sense. If you take a longer time to pay back a loan it is probably a larger loan. Also the loan Amount decreases as Installment Percentage goes up. This also makes sense. The larger percentage of a person's income the loan represents, the riskier they are. The riskier they are, the less money then can borrow.

Distinguishing between the two clusters is a little difficult. The only other significant coefficient in the first cluster is for Age. Signifying a group of people who are loaned more money, i.e. are more credit worthy, as they get older. The other cluster can best be interpreted as "everyone else".

So the first cluster we can say is people for whom we can predict the Amount based on (1) the Duration of the Loan, (2) the Percentage of the person's income the loan represents, and (3) the person's Age. The second cluster is people for whom only the first two elements have any predictive value.

Now let us examine the make up of the three cluster model:

```
#the first cluster
clust.reg.3$results[[1]]
```

```
##
## Call:
## lm(formula = dat[c.best == i, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2800.7  -444.0    21.6   405.0  3438.5
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2267.772    255.900   8.862  < 2e-16 ***
## Duration.of.Credit..month.   95.183      3.716  25.617  < 2e-16 ***
## Instalment.per.cent        -822.847     41.223 -19.961  < 2e-16 ***
## Duration.in.Current.address  14.202     42.806   0.332  0.74025
## Age..years.                  11.016      4.180   2.636  0.00876 **
## No.of.Credits.at.this.Bank  -67.519     79.482  -0.849  0.39617
## No.of.dependents            151.866    116.689   1.301  0.19393
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 849.9 on 362 degrees of freedom
## Multiple R-squared:  0.7207, Adjusted R-squared:  0.7161
## F-statistic: 155.7 on 6 and 362 DF,  p-value: < 2.2e-16
```

```r
#the second cluster
clust.reg.3$results[[2]]
```

```
##
## Call:
## lm(formula = dat[c.best == i, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2773.2  -470.8  -154.5   380.5  7760.2
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  12362.994    668.058  18.506   <2e-16 ***
## Duration.of.Credit..month.     163.388      8.224  19.867   <2e-16 ***
## Instalment.per.cent          -2933.504    104.703 -28.017   <2e-16 ***
## Duration.in.Current.address    -31.388     90.543  -0.347   0.7293
## Age..years.                    -22.503      9.062  -2.483   0.0141 *
## No.of.Credits.at.this.Bank    -238.902    168.832  -1.415   0.1591
## No.of.dependents               -68.149    284.067  -0.240   0.8107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1182 on 156 degrees of freedom
## Multiple R-squared:  0.8922, Adjusted R-squared:  0.888
## F-statistic: 215.1 on 6 and 156 DF,  p-value: < 2.2e-16
```

```r
#the third
clust.reg.3$results[[3]]
```

```
##
## Call:
## lm(formula = dat[c.best == i, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2299.8  -578.3  -177.4   439.7  7208.7
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -2542.280    538.563  -4.720 5.08e-06 ***
## Duration.of.Credit..month.     281.002      8.302  33.849  < 2e-16 ***
## Instalment.per.cent            332.283     89.115   3.729 0.000266 ***
## Duration.in.Current.address     68.588     94.043   0.729 0.466857
## Age..years.                     31.205      9.368   3.331 0.001073 **
## No.of.Credits.at.this.Bank    -142.540    162.747  -0.876 0.382421
## No.of.dependents              -284.599    279.522  -1.018 0.310128
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1252 on 161 degrees of freedom
## Multiple R-squared:  0.8833, Adjusted R-squared:  0.8789
## F-statistic: 203.1 on 6 and 161 DF,  p-value: < 2.2e-16
```

These clusters are considerably more difficult to interpret, but we might make some headway.

Cluster 1: The Average Borrower - Amount goes up as Duration of Loan goes up and down as the bigger the loan is relative to the person's other income. The size of the loan increase (moderately) as the person gets older.

Cluster 2: The Risky Borrower - As before Amount goes up as Duration of Loan goes up, and down as the size of the loan increases relative to the person's other income, except that the effect of the latter is far more pronounced than in Cluster 1. Also the Amount goes *down* as the person ages, meaning they are allowed to borrow less as they get older. This is probably risky borrowers taking out short term loans for relatively small amounts to make ends meet.

Cluster 3: Special Purpose Borrower - These people are loaned more money as the loan increase in duration and as the loan represents a larger percentage of the person's income. It also increases as the person ages. This could be people barrowing for a specific purpose, perhaps a business venture where higher amounts at longer duration are acceptable, provided the interest rate suffices.

So which model do we choose? Although the three cluster model is harder to interpret, it does offer more insight, at least potentially than the two cluster model, which only tells us that some people can borrow more as they get older. It also has a significanly higher R Squared, both on the training and holdout data. **So, I choose the three cluster model.**

Summary

The R Squared increased substantially in moving from one cluster to two clusters. It increased again, less substantially though still significantly, in moving from two clusters to three. So we would have to choose between two clusters and three. For both the two and three cluster models, the drop in R Squared was quite small from train to holdout data. Therefore the question is whether the increase in R Squared in moving from two clusters to three is worth the loss of interpretability of results by adding an additional cluster.

The two cluster model is more interpretable but not very insightful. The three cluster model is difficult to interpret, and those interpretations are somewhat ambiguous. However, if true those interpretations do provide some real insight. It is for that reason that I would choose the three cluster model.