Assignment 1

Matthew Dunne

August 9, 2018

Build a Regression Model

1. Build a linear regression model to predict "Amount" as a function of other variables. Choose the variables you think are most appropriate, and be sure to use the entire dataset to build your model.

I chose a model using the step function.

```
setwd("C:/Users/mjdun/Desktop/Data Mining/Assignments")
MyData <- read.csv(file="German.Credit.csv", header=TRUE, sep=",")
#convert integers to Factors
columns<-c(1,2,4,5,7,8,9,10,11,12,13,15,16,17,18,19,20,21)
MyData[columns] <- lapply(MyData[columns], factor)
#lose the Credibility variable
MyData<-subset(MyData, select = 2:21)
first_model<-step(model1<-lm(Credit.Amount~., data=MyData), direction = "both", trace=0)
#get the number of coeffiecients you have in the model, -1 because you don't need intercept
num_coef<-length(first_model$coefficients)-1
```

The resulting model includes the following variables:

Credit.Amount ~ Account.Balance + Duration.of.Credit..month. + Payment.Status.of.Previous.Credit + Purpose + Value.Savings.Stocks + Instalment.per.cent + Sex... Marital.Status + Guarantors + Most.valuable.available.asset + Occupation + Telephone

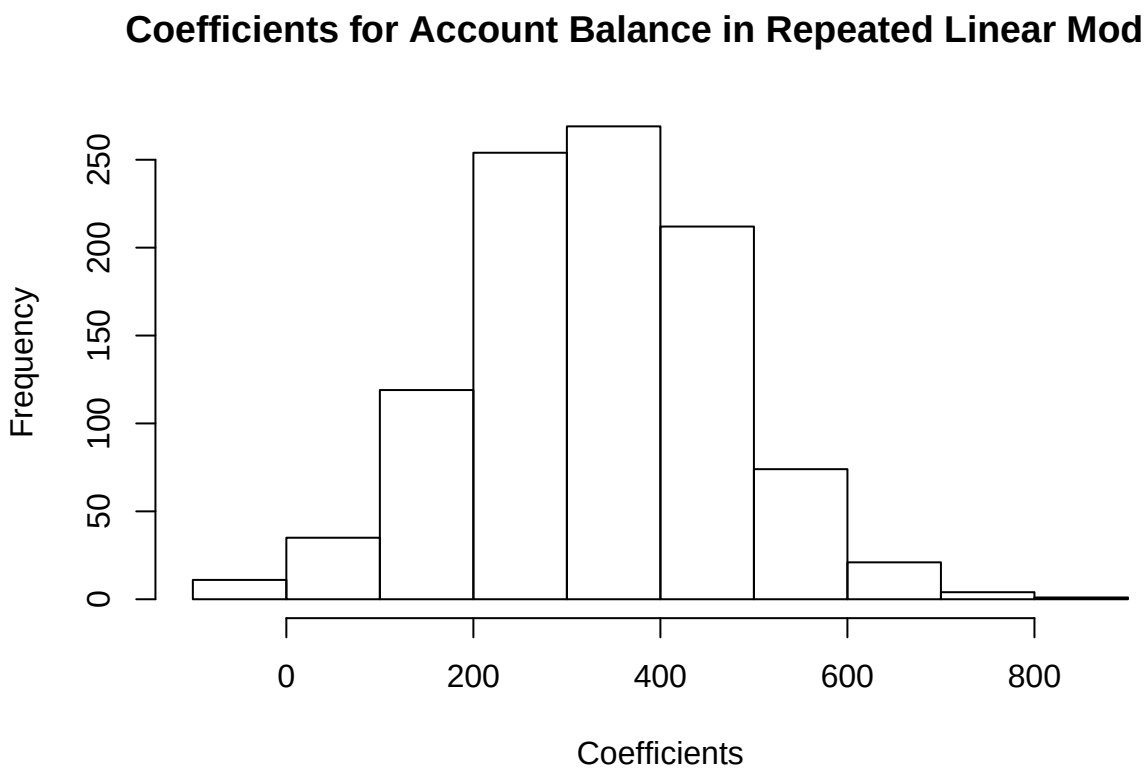
2. Now do the same thing 1000 times while splitting the sample randomly.

```
coeffs<-matrix(NA, 1000, num_coef)
r.sq.training = numeric(1000)
r.sq.test = numeric(1000)
for (i in 1:1000){
  sample<-sample(1:nrow(MyData), 632, replace=FALSE)
  train<-MyData[sample, ]
  test<-MyData[-sample, ]
  linear.model<-lm(Credit.Amount ~ Account.Balance + Duration.of.Credit..month. + Payment.Status.of.Prev
  coeffs[i,]<-linear.model$coefficients[-1]
  r.sq.training[i]<-summary(linear.model)$r.square
  #now R squared in holdout (square of correlation betwee actual and predicted)
  r.sq.test[i]<-cor(test$Credit.Amount, predict(linear.model, newdata = test))^2
}
```

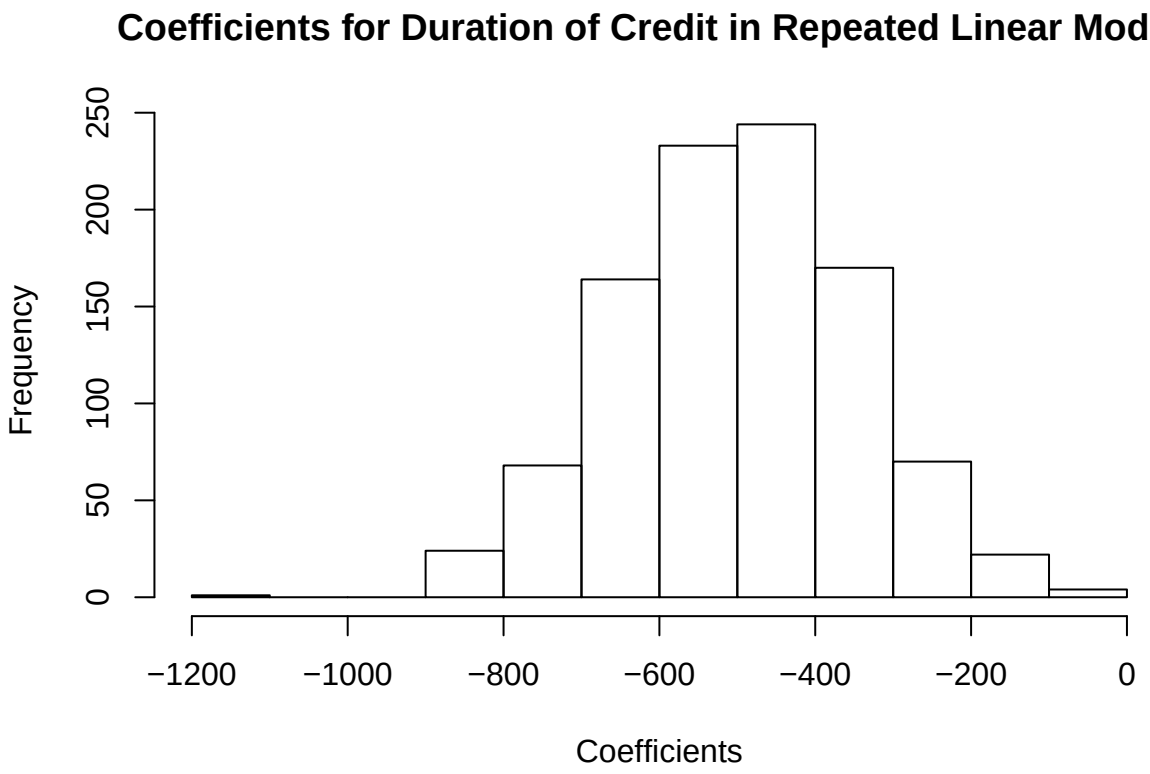
Plot the Distributions

1. Plot the distribution of the coefficients of a few variables from the repeated model.

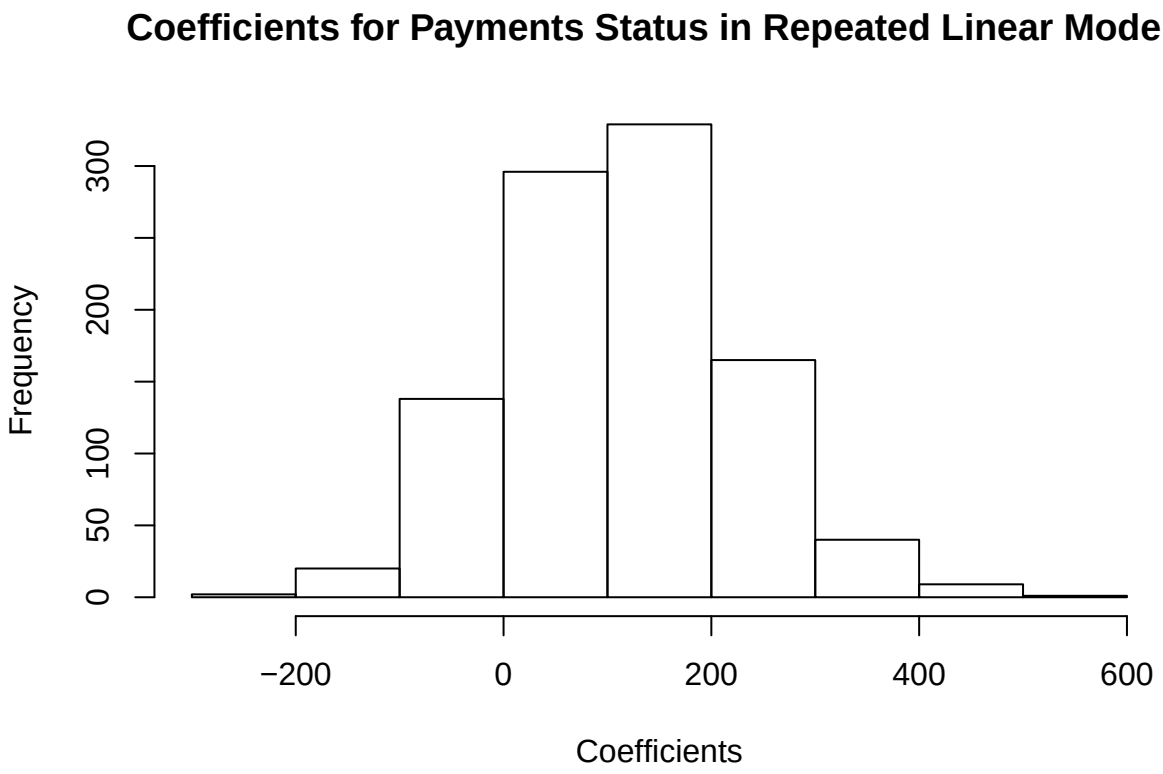
```
hist(coeffs[,1], main = "Coefficients for Account Balance in Repeated Linear Models", xlab="Coefficient
```




```
hist(coeffs[,2], main = "Coefficients for Duration of Credit in Repeated Linear Models", xlab="Coefficie
```

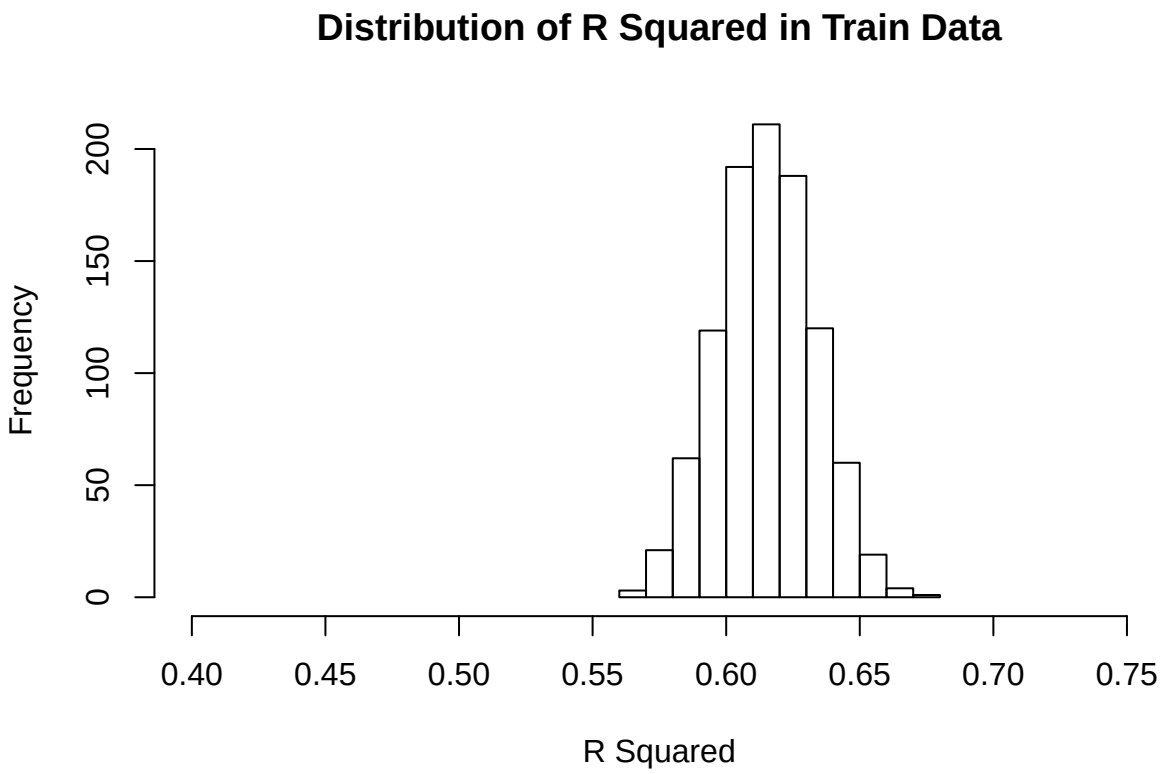



```
hist(coeffs[,3], main = "Coefficients for Payments Status in Repeated Linear Models", xlab="Coefficients
```

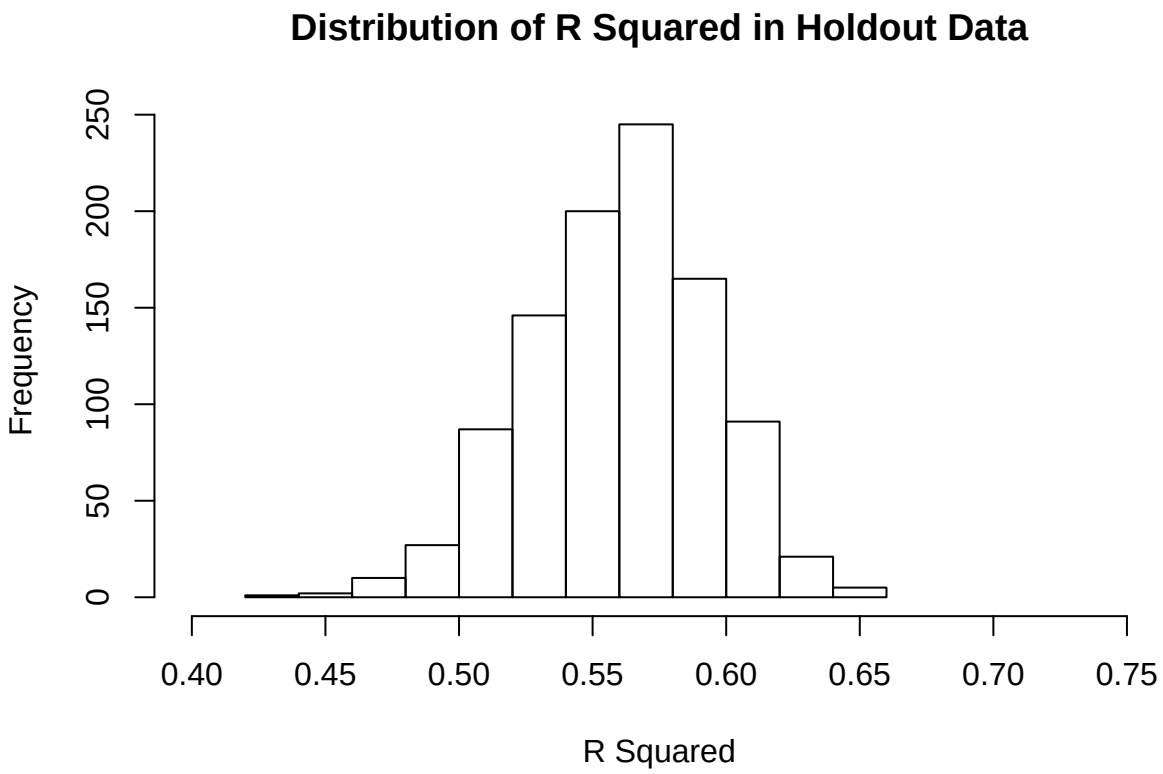


2. From the repeated model, plot the distribution of R Squared from the training data, R Squared from the holdout data, and the difference between the two.

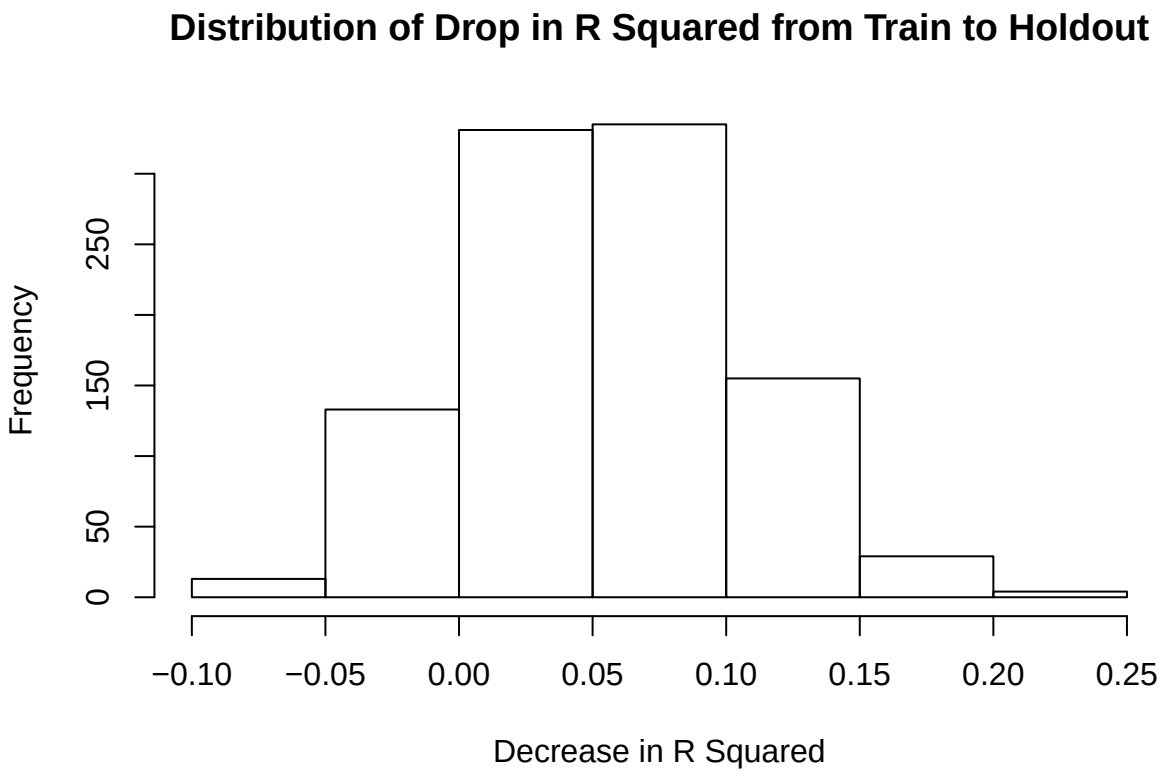
```
#use xlim for first two so the axes/scales are the same
hist(r.sq.training, main="Distribution of R Squared in Train Data", xlab = "R Squared", xlim = c(0.4, 0
```




```
hist(r.sq.test, main="Distribution of R Squared in Holdout Data", xlab = "R Squared", xlim = c(0.4, 0.75
```




```
r.sq.fall<-r.sq.training-r.sq.test
hist(r.sq.fall, main="Distribution of Drop in R Squared from Train to Holdout", xlab = "Decrease in R S
```



If we put these in a table side by side they look something like this.

```
head(cbind("R Sq. Training"=r.sq.training, "R Sq. Holdout"=r.sq.test, "% Drop R Sq."=r.sq.fall))
```

```
##      R Sq. Training R Sq. Holdout % Drop R Sq.
## [1,]      0.6400578     0.5187210  0.121336804
## [2,]      0.6024528     0.6010656  0.001387221
## [3,]      0.5884973     0.6130306 -0.024533228
## [4,]      0.6242124     0.5222730  0.101939412
## [5,]      0.6086232     0.5640239  0.044599353
## [6,]      0.6141359     0.5735027  0.040633206
```

A Further Look at the Coefficients

1. Calculate the mean of each coefficient.

```
variables<-names(first_model$coefficients[-1])
coeff_means<-apply(coeffs, 2, mean)
names(coeff_means)<-variables
```

2. Calculate the standard deviation of each coefficient.

```
coeff_sd<-apply(coeffs, 2, sd)
names(coeff_sd)<-variables
```

3. Compare the means of the coefficients from the repeated sampling model to the coefficients from the model using all the data. Show the percentage difference of the coefficients from the single model from the coefficients of the repeated sampling model.

```r
cbind("Coeff. of Full Sample"=first_model$coefficients[-1], "Mean Coeff. of Rep. Samples"=coeff_means,
```

```
##                                    Coeff. of Full Sample
## Account.Balance2                                333.13609
## Account.Balance3                               -491.28228
## Account.Balance4                                115.88427
## Duration.of.Credit..month.                      127.16914
## Payment.Status.of.Previous.Credit1           -1003.32956
## Payment.Status.of.Previous.Credit2            -904.13569
## Payment.Status.of.Previous.Credit3            -764.79533
## Payment.Status.of.Previous.Credit4            -857.33514
## Purpose1                                        600.62425
## Purpose2                                        -79.54469
## Purpose3                                       -329.52383
## Purpose4                                       -643.19037
## Purpose5                                         55.92755
## Purpose6                                       -144.91939
## Purpose8                                       -482.93286
## Purpose9                                       -233.65448
## Purpose10                                      1630.34124
## Value.Savings.Stocks2                          -249.21864
## Value.Savings.Stocks3                          -356.49210
## Value.Savings.Stocks4                           -57.36226
## Value.Savings.Stocks5                           298.50048
## Instalment.per.cent2                           -816.12532
## Instalment.per.cent3                          -1554.48779
## Instalment.per.cent4                          -2382.49078
## Sex...Marital.Status2                          -192.78732
## Sex...Marital.Status3                           191.79689
## Sex...Marital.Status4                          -460.53303
## Guarantors2                                     671.62405
## Guarantors3                                    -130.80056
## Most.valuable.available.asset2                  243.65658
## Most.valuable.available.asset3                  244.95456
## Most.valuable.available.asset4                  669.41500
## Occupation2                                     421.45784
## Occupation3                                     376.31841
## Occupation4                                    1635.60697
## Telephone2                                      480.25099
##                                    Mean Coeff. of Rep. Samples
## Account.Balance2                                     330.79658
## Account.Balance3                                    -497.45984
## Account.Balance4                                     113.53668
## Duration.of.Credit..month.                           127.26898
## Payment.Status.of.Previous.Credit1                  -1000.64414
## Payment.Status.of.Previous.Credit2                   -896.44037
## Payment.Status.of.Previous.Credit3                   -745.71538
## Payment.Status.of.Previous.Credit4                   -845.69724
## Purpose1                                             596.08666
## Purpose2                                             -76.30026
## Purpose3                                            -322.84023
## Purpose4                                            -629.77998
## Purpose5                                              71.10009
## Purpose6                                            -143.70324
```

```
## Purpose8                                    -482.29193
## Purpose9                                    -246.20247
## Purpose10                                   1677.77898
## Value.Savings.Stocks2                       -247.63681
## Value.Savings.Stocks3                       -366.79756
## Value.Savings.Stocks4                        -62.67487
## Value.Savings.Stocks5                        299.18721
## Instalment.per.cent2                        -809.74207
## Instalment.per.cent3                       -1556.75730
## Instalment.per.cent4                       -2382.97445
## Sex...Marital.Status2                       -193.93703
## Sex...Marital.Status3                        194.58173
## Sex...Marital.Status4                       -462.30000
## Guarantors2                                  653.00542
## Guarantors3                                 -134.52239
## Most.valuable.available.asset2               241.13476
## Most.valuable.available.asset3               242.17561
## Most.valuable.available.asset4               675.52369
## Occupation2                                  404.40046
## Occupation3                                  361.91261
## Occupation4                                 1607.90893
## Telephone2                                   483.42414
##                                          % Diff.
## Account.Balance2                       0.0070723718
## Account.Balance3                      -0.0124182001
## Account.Balance4                       0.0206769682
## Duration.of.Credit..month.            -0.0007844861
## Payment.Status.of.Previous.Credit1     0.0026836927
## Payment.Status.of.Previous.Credit2     0.0085843102
## Payment.Status.of.Previous.Credit3     0.0255861011
## Payment.Status.of.Previous.Credit4     0.0137613051
## Purpose1                               0.0076122951
## Purpose2                               0.0425219917
## Purpose3                               0.0207025022
## Purpose4                               0.0212937695
## Purpose5                              -0.2133969715
## Purpose6                               0.0084629505
## Purpose8                               0.0013289263
## Purpose9                              -0.0509661231
## Purpose10                             -0.0282741273
## Value.Savings.Stocks2                  0.0063876797
## Value.Savings.Stocks3                 -0.0280957802
## Value.Savings.Stocks4                 -0.0847645157
## Value.Savings.Stocks5                 -0.0022953328
## Instalment.per.cent2                   0.0078830621
## Instalment.per.cent3                  -0.0014578494
## Instalment.per.cent4                  -0.0002029660
## Sex...Marital.Status2                 -0.0059282496
## Sex...Marital.Status3                 -0.0143119571
## Sex...Marital.Status4                 -0.0038221318
## Guarantors2                            0.0285122096
## Guarantors3                           -0.0276669625
## Most.valuable.available.asset2         0.0104581354
## Most.valuable.available.asset3         0.0114749069
```

```
## Most.valuable.available.asset4     -0.0090429067
## Occupation2                         0.0421794361
## Occupation3                         0.0398046567
## Occupation4                         0.0172261265
## Telephone2                         -0.0065639002
```

4. Show the 2.5% to 97.5% Confidence Interval for each coefficient from the repeated sample model.

```
rep.lower<-coeff_means + qnorm(.025)*coeff_sd
rep.higher<-coeff_means+qnorm(.975)*coeff_sd
rep.width<-(rep.higher-rep.lower)*sqrt(.632)
full.lower<-(confint(first_model)[,1])[-1]
full.upper<-(confint(first_model)[,2])[-1]
full.width<-full.upper-full.lower
conf_int<-cbind(rep.lower, rep.higher, rep.width, full.lower, full.upper, full.width)
conf_int
```

```
##                                      rep.lower    rep.higher  rep.width
## Account.Balance2                      56.44742   605.1457396  436.20654
## Account.Balance3                    -791.44553  -203.4741446  467.42800
## Account.Balance4                    -108.97206   336.0454103  353.78189
## Duration.of.Credit..month.           116.71497   137.8229861   16.78054
## Payment.Status.of.Previous.Credit1 -1745.02291  -256.2653688 1183.53884
## Payment.Status.of.Previous.Credit2 -1529.68936  -263.1913773 1006.84598
## Payment.Status.of.Previous.Credit3 -1490.59513    -0.8356359 1184.33537
## Payment.Status.of.Previous.Credit4 -1480.41457  -210.9799034 1009.18060
## Purpose1                            183.79218  1008.3811444  655.53526
## Purpose2                           -330.15434   177.5538339  403.62002
## Purpose3                           -542.32242  -103.3580527  348.96977
## Purpose4                          -1193.99655   -65.5634074  897.08660
## Purpose5                           -551.95894   694.1591178  990.64426
## Purpose6                           -605.15760   317.7511157  733.69792
## Purpose8                          -1486.58184   521.9979820 1596.78938
## Purpose9                           -611.82346   119.4185293  581.32589
## Purpose10                          -455.30378  3810.8617329 3391.53450
## Value.Savings.Stocks2              -531.81639    36.5427616  451.83659
## Value.Savings.Stocks3              -664.58415   -69.0109752  473.47131
## Value.Savings.Stocks4              -390.31898   264.9692373  520.94383
## Value.Savings.Stocks5                27.79056   570.5838696  431.51215
## Instalment.per.cent2              -1167.70530  -451.7788422  569.15027
## Instalment.per.cent3              -1924.62966 -1188.8849519  584.90547
## Instalment.per.cent4              -2721.46253 -2044.4863631  538.18541
## Sex...Marital.Status2              -673.66169   285.7876347  762.74713
## Sex...Marital.Status3              -275.52269   664.6861478  747.45124
## Sex...Marital.Status4              -956.41749    31.8174865  785.63126
## Guarantors2                         147.27371  1158.7371323  804.09752
## Guarantors3                        -440.34049   171.2957161  486.24116
## Most.valuable.available.asset2       24.37945   457.8900623  344.63412
## Most.valuable.available.asset3       45.59210   438.7591342  312.56161
## Most.valuable.available.asset4      321.45468  1029.5927056  562.95859
## Occupation2                        -754.26822  1563.0691263 1842.24675
## Occupation3                        -791.11226  1514.9374720 1833.27328
## Occupation4                         378.40998  2837.4078744 1954.86466
## Telephone2                          278.88762   687.9606570  325.20664
##                                    full.lower  full.upper full.width
```

```
## Account.Balance2                      11.10779   655.16439   644.0566
## Account.Balance3                    -999.39575    16.83119  1016.2269
## Account.Balance4                    -188.53599   420.30453   608.8405
## Duration.of.Credit..month.           116.72189   137.61639    20.8945
## Payment.Status.of.Previous.Credit1 -1780.57173  -226.08739  1554.4843
## Payment.Status.of.Previous.Credit2 -1512.02703  -296.24435  1215.7827
## Payment.Status.of.Previous.Credit3 -1452.15434   -77.43633  1374.7180
## Payment.Status.of.Previous.Credit4 -1486.11057  -228.55970  1257.5509
## Purpose1                            158.51355  1042.73495   884.2214
## Purpose2                           -442.30415   283.21476   725.5189
## Purpose3                           -656.90727    -2.14040   654.7669
## Purpose4                          -1705.11065   418.72991  2123.8406
## Purpose5                           -746.44411   858.29921  1604.7433
## Purpose6                           -710.29169   420.45290  1130.7446
## Purpose8                          -1715.44588   749.58016  2465.0260
## Purpose9                           -689.28399   221.97502   911.2590
## Purpose10                           531.04775  2729.63474  2198.5870
## Value.Savings.Stocks2              -641.86256   143.42529   785.2879
## Value.Savings.Stocks3              -840.83985   127.85565   968.6955
## Value.Savings.Stocks4              -603.45857   488.73404  1092.1926
## Value.Savings.Stocks5               -17.17020   614.17117   631.3414
## Instalment.per.cent2              -1209.22195  -423.02869   786.1933
## Instalment.per.cent3              -1981.19680 -1127.77877   853.4180
## Instalment.per.cent4              -2744.45978 -2020.52179   723.9380
## Sex...Marital.Status2              -744.90175   359.32711  1104.2289
## Sex...Marital.Status3              -346.65144   730.24521  1076.8967
## Sex...Marital.Status4             -1101.74997   180.68392  1282.4339
## Guarantors2                          87.34941  1255.89868  1168.5493
## Guarantors3                        -664.19751   402.59640  1066.7939
## Most.valuable.available.asset2      -87.40470   574.71786   662.1226
## Most.valuable.available.asset3      -68.00386   557.91298   625.9168
## Most.valuable.available.asset4      263.13170  1075.69829   812.5666
## Occupation2                        -399.59551  1242.51119  1642.1067
## Occupation3                        -422.82276  1175.45959  1598.2823
## Occupation4                         781.28920  2489.92475  1708.6355
## Telephone2                          221.29923   739.20274   517.9035
```

For how many of these coefficients is the confidence interval of the repeated model smaller than the confidence interval of the full model?

```
nrow(conf_int[conf_int[,3]<conf_int[,6], ])
```

```
## [1] 32
```

Summary of Results

The first thing of note is on the distributions of the coefficients, R Squared's. They are all more or less normal. The distributions of the R Squared for Train and Holdout are obviously different, but for the most part within 0.1.

The means of the coefficients from the repeated samples are, as one might expect, different from the single, full sample. However, they are remarkably close, with only a few coefficients having more than a 10% difference.

The confidence intervals from repeated sample are mostly narrower (32 out of 36 coefficients). This makes sense as we would be more confident in repeated samplings than a single sample.