FTAP HW 7

Zachary Fogelson

July 21, 2015

Lasso Homework

Problem 1

1.

```
train <- read.csv("train.csv")
trainMeaningFull <-subset(train, select=-c(id, member_id))
trainMeaningFull <- trainMeaningFull[complete.cases(trainMeaningFull),]
kable(head(train))
```

id	member_id	funded_amnt	term_m	int_rate	installment	emp_length	annual_inc	dti	revol_util
54734	80364	25000	36	0.1189	829.10	0	85000	19.48	0.521
55742	114426	7000	36	0.1071	228.22	0	65000	14.29	0.767
57245	138150	1200	36	0.1311	40.50	15	54000	5.47	0.404
57416	139635	10800	36	0.1357	366.86	6	32000	11.63	0.256
58915	153417	5025	36	0.1008	162.34	3	85000	8.10	0.732
59006	154254	3000	36	0.1426	102.92	3	80800	14.97	0.395

```
kable(rbind(summary(train), sapply(train, sd)))
```

id	member_id	funded_amnt	term_m	int_rate	installment
Min. : 54734	Min. : 70699	Min. : 500	Min. :36.00	Min. :0.0542	Min. : 15.69
1st Qu.:496560	1st Qu.: 635543	1st Qu.: 5000	1st Qu.:36.00	1st Qu.:0.0925	1st Qu.: 163.67
Median :623258	Median : 798198	Median : 9250	Median :36.00	Median :0.1171	Median : 272.80
Mean :626300	Mean : 783617	Mean :10654	Mean :42.22	Mean :0.1191	Mean : 317.15
3rd Qu.:767345	3rd Qu.: 967828	3rd Qu.:15000	3rd Qu.:60.00	3rd Qu.:0.1435	3rd Qu.: 417.36
Max. :975902	Max. :1198245	Max. :35000	Max. :60.00	Max. :0.2459	Max. :1305.19
NA	NA	NA	NA	NA	NA
170023.183351294	225109.217961916	6974.75731097762	10.5166250234748	0.0362426792330949	205.33220291880

2.

a)

```
ols <- lm(int_rate ~ ., trainMeaningFull)
olsSum <- summary(ols)
```

Based on the OLS regression, all of the columns which are not IDs or have Na's appear to be significant

b)

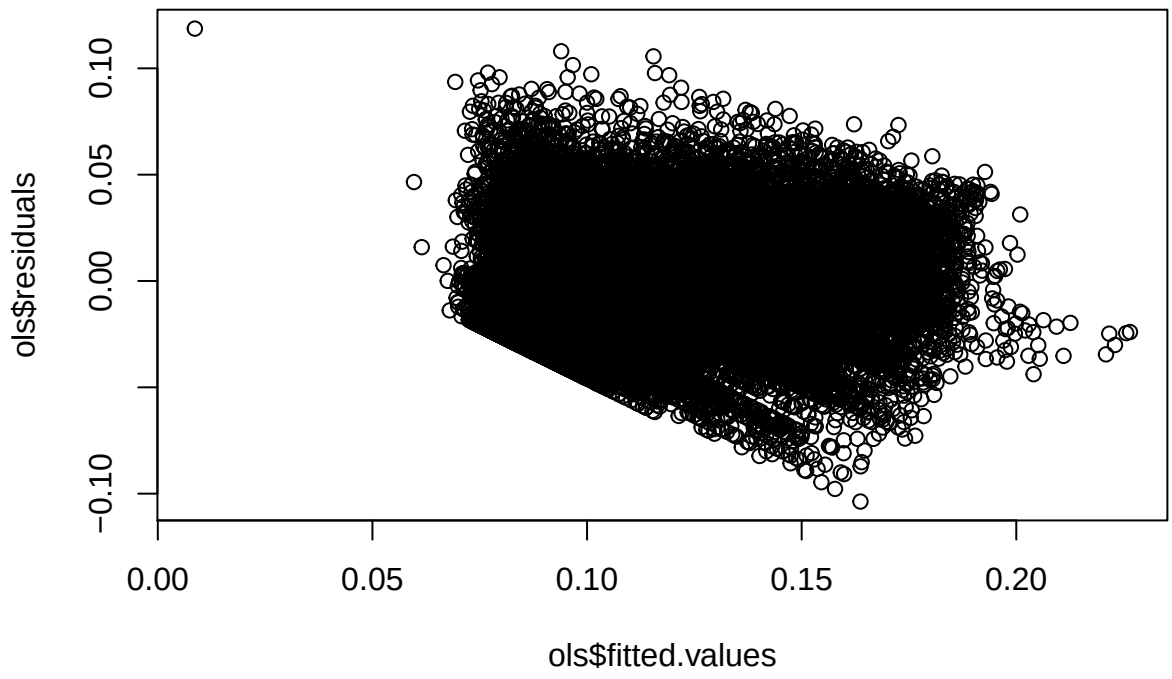
```
olsSum$r.squared
```

```
## [1] 0.5120423
```

The baseline model has an R^2 of about .39. It is hard to judge if the baseline model is successful because, we do not have anything to compare it to. But using all of the available information we can explain 39% of the variance which is not terrible.

c)

```
plot(ols$fitted.values, ols$residuals)
```



d)

```
rmse <- mean((ols$residuals)^2)
rmse
```

```
## [1] 0.0006375506
```

e)

```
test <- read.csv("test.csv")
test <- subset(test, select=-c(id, member_id))
test <- test[complete.cases(test),]
```

```
predOLS <- predict(ols,test, interval = "none")
rmseOLS <- mean((test$int_rate - predOLS)^2)
cat("RMSE OLS: ", rmseOLS)
```

```
## RMSE OLS:  0.0007511019
```

3.

```
interactDF <- as.data.frame(model.matrix(~(.)^2,subset(trainMeaningFull, select=-c(int_rate))))
interactTest <- as.data.frame(model.matrix(~(.)^2,subset(test, select=-c(int_rate))))
```

a)

```
smallest = lm(trainMeaningFull$int_rate ~ 1, interactDF)
biggest = as.formula(lm(trainMeaningFull$int_rate ~ ., interactDF))
stepForwardAIC <- step(smallest, scope = biggest, k = 2, direction ="forward", trace = F)
stepForwardBIC <- step(smallest, scope = biggest, k = log(length(trainMeaningFull$int_rate)), direction
summary(stepForwardAIC)
```

```
##
## Call:
## lm(formula = trainMeaningFull$int_rate ~ `term_m:revol_util` +
##     `term_m:installment` + funded_amnt + installment + `funded_amnt:term_m` +
##     `funded_amnt:revol_util` + `funded_amnt:installment` + term_m +
##     revol_util + `annual_inc:dti` + `installment:revol_util` +
##     emp_length + `funded_amnt:emp_length` + `term_m:emp_length` +
##     `installment:emp_length` + `term_m:dti` + `funded_amnt:dti` +
##     `emp_length:revol_util` + `funded_amnt:annual_inc` + `installment:annual_inc` +
##     `term_m:annual_inc` + annual_inc + `emp_length:annual_inc` +
##     `annual_inc:revol_util` + `installment:dti` + dti, data = interactDF)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.108140 -0.007357 -0.000139  0.006408  0.109870
##
## Coefficients:
##                              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                 1.701e-02  1.736e-03    9.798  < 2e-16 ***
## `term_m:revol_util`         6.979e-05  5.079e-05    1.374 0.169442
## `term_m:installment`       -2.173e-05  3.623e-07  -59.971  < 2e-16 ***
## funded_amnt                -9.614e-05  6.347e-07 -151.464  < 2e-16 ***
## installment                 2.563e-03  2.026e-05  126.480  < 2e-16 ***
## `funded_amnt:term_m`        1.138e-06  1.081e-08  105.219  < 2e-16 ***
## `funded_amnt:revol_util`   -7.406e-07  2.723e-07   -2.720 0.006537 **
## `funded_amnt:installment` -2.319e-09  4.888e-11  -47.448  < 2e-16 ***
## term_m                      1.887e-03  4.322e-05   43.652  < 2e-16 ***
## revol_util                  4.854e-02  2.105e-03   23.053  < 2e-16 ***
## `annual_inc:dti`           -1.322e-09  2.268e-10   -5.831 5.56e-09 ***
## `installment:revol_util`   -9.782e-05  8.767e-06  -11.157  < 2e-16 ***
## emp_length                  1.352e-04  1.072e-04    1.261 0.207461
## `funded_amnt:emp_length`    1.149e-07  1.314e-08    8.741  < 2e-16 ***
## `term_m:emp_length`        -1.916e-05  2.657e-06   -7.211 5.69e-13 ***
## `installment:emp_length`   -2.863e-06  4.269e-07   -6.707 2.02e-11 ***
## `term_m:dti`               -1.867e-05  2.148e-06   -8.693  < 2e-16 ***
## `funded_amnt:dti`           9.503e-08  1.119e-08    8.489  < 2e-16 ***
## `emp_length:revol_util`     1.872e-04  5.526e-05    3.388 0.000704 ***
```

```
## `funded_amnt:annual_inc`   2.884e-11  1.185e-12   24.340  < 2e-16 ***
## `installment:annual_inc`  -8.707e-10  3.746e-11  -23.241  < 2e-16 ***
## `term_m:annual_inc`       -5.291e-09  2.642e-10  -20.029  < 2e-16 ***
## annual_inc                 1.602e-07  1.065e-08   15.036  < 2e-16 ***
## `emp_length:annual_inc`    1.521e-09  2.224e-10    6.838 8.18e-12 ***
## `annual_inc:revol_util`    2.574e-08  5.350e-09    4.811 1.51e-06 ***
## `installment:dti`         -2.384e-06  3.672e-07   -6.493 8.51e-11 ***
## dti                        4.700e-04  8.879e-05    5.293 1.21e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01457 on 33093 degrees of freedom
## Multiple R-squared:  0.8377, Adjusted R-squared:  0.8376
## F-statistic:  6570 on 26 and 33093 DF,  p-value: < 2.2e-16
```

```
summary(stepForwardBIC)
```

```
##
## Call:
## lm(formula = trainMeaningFull$int_rate ~ `term_m:revol_util` +
##     `term_m:installment` + funded_amnt + installment + `funded_amnt:term_m` +
##     `funded_amnt:revol_util` + `funded_amnt:installment` + term_m +
##     revol_util + `annual_inc:dti` + `installment:revol_util` +
##     emp_length + `funded_amnt:emp_length` + `term_m:emp_length` +
##     `installment:emp_length` + `term_m:dti` + `funded_amnt:dti` +
##     `emp_length:revol_util` + `funded_amnt:annual_inc` + `installment:annual_inc` +
##     `term_m:annual_inc` + annual_inc + `emp_length:annual_inc` +
##     `annual_inc:revol_util` + `installment:dti` + dti, data = interactDF)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.108140 -0.007357 -0.000139  0.006408  0.109870
##
## Coefficients:
##                             Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                1.701e-02  1.736e-03    9.798  < 2e-16 ***
## `term_m:revol_util`        6.979e-05  5.079e-05    1.374 0.169442
## `term_m:installment`      -2.173e-05  3.623e-07  -59.971  < 2e-16 ***
## funded_amnt               -9.614e-05  6.347e-07 -151.464  < 2e-16 ***
## installment                2.563e-03  2.026e-05  126.480  < 2e-16 ***
## `funded_amnt:term_m`       1.138e-06  1.081e-08  105.219  < 2e-16 ***
## `funded_amnt:revol_util`  -7.406e-07  2.723e-07   -2.720 0.006537 **
## `funded_amnt:installment` -2.319e-09  4.888e-11  -47.448  < 2e-16 ***
## term_m                     1.887e-03  4.322e-05   43.652  < 2e-16 ***
## revol_util                 4.854e-02  2.105e-03   23.053  < 2e-16 ***
## `annual_inc:dti`          -1.322e-09  2.268e-10   -5.831 5.56e-09 ***
## `installment:revol_util`  -9.782e-05  8.767e-06  -11.157  < 2e-16 ***
## emp_length                 1.352e-04  1.072e-04    1.261 0.207461
## `funded_amnt:emp_length`   1.149e-07  1.314e-08    8.741  < 2e-16 ***
## `term_m:emp_length`       -1.916e-05  2.657e-06   -7.211 5.69e-13 ***
## `installment:emp_length`  -2.863e-06  4.269e-07   -6.707 2.02e-11 ***
## `term_m:dti`              -1.867e-05  2.148e-06   -8.693  < 2e-16 ***
## `funded_amnt:dti`          9.503e-08  1.119e-08    8.489  < 2e-16 ***
## `emp_length:revol_util`    1.872e-04  5.526e-05    3.388 0.000704 ***
```

```
## `funded_amnt:annual_inc`   2.884e-11  1.185e-12   24.340  < 2e-16 ***
## `installment:annual_inc`  -8.707e-10  3.746e-11  -23.241  < 2e-16 ***
## `term_m:annual_inc`       -5.291e-09  2.642e-10  -20.029  < 2e-16 ***
## annual_inc                 1.602e-07  1.065e-08   15.036  < 2e-16 ***
## `emp_length:annual_inc`    1.521e-09  2.224e-10    6.838 8.18e-12 ***
## `annual_inc:revol_util`    2.574e-08  5.350e-09    4.811 1.51e-06 ***
## `installment:dti`         -2.384e-06  3.672e-07   -6.493 8.51e-11 ***
## dti                        4.700e-04  8.879e-05    5.293 1.21e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01457 on 33093 degrees of freedom
## Multiple R-squared:  0.8377, Adjusted R-squared:  0.8376
## F-statistic:  6570 on 26 and 33093 DF,  p-value: < 2.2e-16
```

The in-sample R^2 of the AIC function is .988.

b) See summaries above

c)

```
rmseAIC <- mean((stepForwardAIC$residuals)^2)
rmseBIC <- mean((stepForwardBIC$residuals)^2)
cat("RMSE AIC: ", rmseAIC, "RMSE BIC: ", rmseBIC)
```

```
## RMSE AIC:  0.0002120341 RMSE BIC:  0.0002120341
```

d)

```
predAIC <- predict(stepForwardAIC,interactTest, interval = "none")
predBIC <- predict(stepForwardBIC,interactTest, interval = "none")
rmseAIC <- mean((test$int_rate - predAIC)^2)
rmseBIC <- mean((test$int_rate - predBIC)^2)
cat("RMSE AIC: ", rmseAIC, "RMSE BIC: ", rmseBIC)
```

```
## RMSE AIC:  0.0002731841 RMSE BIC:  0.0002731841
```

4. a)

```
biggest = lm(trainMeaningFull$int_rate ~ ., interactDF)
smallest = as.formula(lm(trainMeaningFull$int_rate ~ 1, interactDF))
stepBackwardAIC <- step(biggest, scope = smallest, k = 2, direction ="backward", trace = F)
stepBackwardBIC <- step(biggest, scope = smallest, k = log(length(trainMeaningFull$int_rate)), directio
summary(stepBackwardAIC)
```

```
##
## Call:
## lm(formula = trainMeaningFull$int_rate ~ funded_amnt + term_m +
##     installment + annual_inc + dti + revol_util + `funded_amnt:term_m` +
##     `funded_amnt:installment` + `funded_amnt:emp_length` + `funded_amnt:annual_inc` +
##     `funded_amnt:dti` + `funded_amnt:revol_util` + `term_m:installment` +
```

```
##     `term_m:emp_length` + `term_m:annual_inc` + `term_m:dti` +
##     `installment:emp_length` + `installment:annual_inc` + `installment:dti` +
##     `installment:revol_util` + `emp_length:annual_inc` + `emp_length:revol_util` +
##     `annual_inc:dti` + `annual_inc:revol_util`, data = interactDF)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.108450 -0.007354 -0.000135  0.006420  0.109663
##
## Coefficients:
##                              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                 1.671e-02  1.527e-03   10.938  < 2e-16 ***
## funded_amnt                -9.612e-05  6.320e-07 -152.074  < 2e-16 ***
## term_m                      1.896e-03  3.809e-05   49.775  < 2e-16 ***
## installment                 2.562e-03  2.019e-05  126.913  < 2e-16 ***
## annual_inc                  1.610e-07  1.058e-08   15.225  < 2e-16 ***
## dti                         4.463e-04  8.588e-05    5.197 2.04e-07 ***
## revol_util                  5.130e-02  6.311e-04   81.290  < 2e-16 ***
## `funded_amnt:term_m`        1.137e-06  1.079e-08  105.387  < 2e-16 ***
## `funded_amnt:installment` -2.320e-09  4.888e-11  -47.454  < 2e-16 ***
## `funded_amnt:emp_length`    1.022e-07  8.512e-09   12.011  < 2e-16 ***
## `funded_amnt:annual_inc`    2.894e-11  1.174e-12   24.654  < 2e-16 ***
## `funded_amnt:dti`           9.245e-08  1.090e-08    8.480  < 2e-16 ***
## `funded_amnt:revol_util`  -4.486e-07  1.674e-07   -2.679 0.007383 **
## `term_m:installment`      -2.170e-05  3.613e-07  -60.050  < 2e-16 ***
## `term_m:emp_length`       -1.605e-05  9.636e-07  -16.654  < 2e-16 ***
## `term_m:annual_inc`       -5.317e-09  2.615e-10  -20.330  < 2e-16 ***
## `term_m:dti`              -1.809e-05  2.073e-06   -8.728  < 2e-16 ***
## `installment:emp_length`  -2.445e-06  2.691e-07   -9.086  < 2e-16 ***
## `installment:annual_inc`  -8.741e-10  3.712e-11  -23.550  < 2e-16 ***
## `installment:dti`         -2.304e-06  3.580e-07   -6.435 1.25e-10 ***
## `installment:revol_util`  -1.070e-04  5.593e-06  -19.131  < 2e-16 ***
## `emp_length:annual_inc`    1.544e-09  2.215e-10    6.968 3.27e-12 ***
## `emp_length:revol_util`    1.888e-04  5.525e-05    3.416 0.000636 ***
## `annual_inc:dti`          -1.309e-09  2.264e-10   -5.781 7.49e-09 ***
## `annual_inc:revol_util`    2.564e-08  5.350e-09    4.793 1.65e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01457 on 33095 degrees of freedom
## Multiple R-squared:  0.8377, Adjusted R-squared:  0.8376
## F-statistic:  7117 on 24 and 33095 DF,  p-value: < 2.2e-16
```

```
summary(stepBackwardBIC)
```

```
##
## Call:
## lm(formula = trainMeaningFull$int_rate ~ funded_amnt + term_m +
##     installment + annual_inc + dti + revol_util + `funded_amnt:term_m` +
##     `funded_amnt:installment` + `funded_amnt:emp_length` + `funded_amnt:annual_inc` +
##     `funded_amnt:dti` + `term_m:installment` + `term_m:emp_length` +
##     `term_m:annual_inc` + `term_m:dti` + `installment:emp_length` +
##     `installment:annual_inc` + `installment:dti` + `installment:revol_util` +
##     `emp_length:annual_inc` + `annual_inc:dti` + `annual_inc:revol_util`,
```

```
##     data = interactDF)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.109568 -0.007344 -0.000133  0.006419  0.109796
##
## Coefficients:
##                              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                 1.693e-02  1.526e-03   11.095  < 2e-16 ***
## funded_amnt                -9.675e-05  5.785e-07 -167.255  < 2e-16 ***
## term_m                      1.881e-03  3.792e-05   49.610  < 2e-16 ***
## installment                 2.584e-03  1.824e-05  141.658  < 2e-16 ***
## annual_inc                  1.625e-07  1.057e-08   15.375  < 2e-16 ***
## dti                         4.471e-04  8.590e-05    5.205 1.95e-07 ***
## revol_util                  5.206e-02  5.761e-04   90.354  < 2e-16 ***
## `funded_amnt:term_m`        1.148e-06  9.896e-09  116.000  < 2e-16 ***
## `funded_amnt:installment` -2.316e-09  4.888e-11  -47.378  < 2e-16 ***
## `funded_amnt:emp_length`    9.148e-08  7.934e-09   11.530  < 2e-16 ***
## `funded_amnt:annual_inc`    2.912e-11  1.173e-12   24.825  < 2e-16 ***
## `funded_amnt:dti`           8.814e-08  1.075e-08    8.195 2.59e-16 ***
## `term_m:installment`       -2.210e-05  3.268e-07  -67.615  < 2e-16 ***
## `term_m:emp_length`        -1.390e-05  7.371e-07  -18.861  < 2e-16 ***
## `term_m:annual_inc`        -5.364e-09  2.613e-10  -20.525  < 2e-16 ***
## `term_m:dti`               -1.808e-05  2.073e-06   -8.722  < 2e-16 ***
## `installment:emp_length`   -2.082e-06  2.481e-07   -8.389  < 2e-16 ***
## `installment:annual_inc`   -8.802e-10  3.707e-11  -23.741  < 2e-16 ***
## `installment:dti`          -2.164e-06  3.532e-07   -6.128 8.97e-10 ***
## `installment:revol_util`   -1.211e-04  1.558e-06  -77.691  < 2e-16 ***
## `emp_length:annual_inc`     1.533e-09  2.215e-10    6.920 4.60e-12 ***
## `annual_inc:dti`           -1.294e-09  2.265e-10   -5.714 1.11e-08 ***
## `annual_inc:revol_util`     2.706e-08  5.323e-09    5.085 3.70e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01457 on 33097 degrees of freedom
## Multiple R-squared:  0.8376, Adjusted R-squared:  0.8375
## F-statistic:  7760 on 22 and 33097 DF,  p-value: < 2.2e-16
```

The R^2 term for the best fitting model in the backward model is .988.

b)

See summaries above.

c)

```
rmseBackAIC <- mean((stepBackwardAIC$residuals)^2)
rmseBackBIC <- mean((stepBackwardBIC$residuals)^2)
cat("RMSE AIC: ", rmseBackAIC, "RMSE BIC: ", rmseBackBIC)
```

```
## RMSE AIC:  0.0002120568 RMSE BIC:  0.0002121667
```

d)

```
predBackAIC <- predict(stepBackwardAIC,interactTest, interval = "none")
predBackBIC <- predict(stepBackwardBIC,interactTest, interval = "none")
rmseBackAIC <- mean((test$int_rate - predBackAIC)^2)
rmseBackBIC <- mean((test$int_rate - predBackBIC)^2)
cat("RMSE AIC: ", rmseBackAIC, "RMSE BIC: ", rmseBackBIC)
```

```
## RMSE AIC:  0.0002732995 RMSE BIC:  0.0002731435
```

5.

a)

```
model=as.formula(paste("~", paste(names(interactDF)[-1], collapse= "+")))
x=model.matrix(model,interactDF);
lassoFit <- glmnet(x,trainMeaningFull$int_rate)
coef(lassoFit, s=.0001)
```

```
## 30 x 1 sparse Matrix of class "dgCMatrix"
##                                      1
## (Intercept)              2.539734e-02
## (Intercept)              .
## funded_amnt             -2.693844e-05
## term_m                   1.572056e-03
## installment              3.912436e-04
## emp_length              -4.509901e-04
## annual_inc              -1.140216e-08
## dti                     -1.850803e-04
## revol_util               5.564668e-02
## funded_amnt:term_m       .
## funded_amnt:installment -1.266817e-09
## funded_amnt:emp_length   1.304751e-08
## funded_amnt:annual_inc   .
## funded_amnt:dti          .
## funded_amnt:revol_util  -4.823517e-06
## term_m:installment       1.392770e-05
## term_m:emp_length        .
## term_m:annual_inc        .
## term_m:dti               .
## term_m:revol_util        .
## installment:emp_length   1.015833e-07
## installment:annual_inc   7.992317e-12
## installment:dti          4.234663e-07
## installment:revol_util   6.463018e-05
## emp_length:annual_inc    5.289935e-10
## emp_length:dti           .
## emp_length:revol_util    1.753555e-05
## annual_inc:dti          -1.428128e-09
## annual_inc:revol_util    .
## dti:revol_util           .
```

There is no R^2 for the lasso model.

b)

```
insamplePred <- predict(lassoFit, newx = as.matrix(interactDF), s=.0001)
lassoRMSE <- mean((insamplePred - trainMeaningFull$int_rate)^2)
cat("RMSE Lasso: ", lassoRMSE)
```

```
## RMSE Lasso:  0.0002951606
```

c)

```
predLasso <- predict(lassoFit,newx = as.matrix(interactTest), s=.0001)
lassoRMSE <- mean((test$int_rate - predLasso)^2)
cat("RMSE Lasso: ", lassoRMSE)
```

6.

a)

```
lassoCVFit <- cv.glmnet(x,trainMeaningFull$int_rate, nfolds = 10)
lassoCVFit$lambda.min
```

```
## [1] 1.332991e-05
```

b)

```
insampleCVPred <- predict(lassoCVFit, newx = as.matrix(interactDF), s="lambda.min")
lassoCVRMSE <- mean((insampleCVPred - trainMeaningFull$int_rate)^2)
cat("RMSE CVLasso: ", lassoCVRMSE)
```

```
## RMSE CVLasso:  0.000226073
```

c)

```
predCVLasso <- predict(lassoCVFit,newx = as.matrix(interactTest[-1]), s="lambda.min")
lassoCVRMSE <- mean((test$int_rate - predCVLasso)^2)
cat("RMSE CVLasso: ", lassoCVRMSE)
```

Logit Homework

```
trades <- read.csv("detailed_trades_est.csv")
```

1)

```
logit <- glm(PCHANGE~LASK+LBID+RETURN+SIGN_VOL, family=binomial, data=trades)
AIC(logit)
```

```
## [1] 3754.378
```

```
BIC(logit)
```

```
## [1] 3784.334
```

2)

```
interactions <- subset(trades,select = -c(PCHANGE,PCHANGE0))
interactions <- as.data.frame(model.matrix(~(.)^2, interactions))
```

3)

```
trades.pca <- prcomp(interactions)
myModel <- lm(trades$PCHANGE~trades.pca$x[,1]+trades.pca$x[,2]+trades.pca$x[,3])
```

4)

```
error <- trades$PCHANGE-(as.numeric(myModel$fitted.values > .5))
mean(abs(error))
```

```
## [1] 0.3431472
```