Project 1 - Data Science for Economists

Lin, Oliver; Nazarov, Nazar; Priolo, Robert

Professor Rojas - Spring 2021

Question 1

vegas5 dataset with variables *default*, *arm*, *refinance*, *lien2*, *term30*, *underwater*, *ltv*, *rate*, *amount* and *fico*.

Observations: 10,000 mortgage loan observations from Las Vegas, Nevada, single family homes, 2010

```
stat_sum <- describe(vegas5)[,-c(1,6,7,10)]
kable(stat_sum,
      caption = "Statistical Summary")
```

Observations: 10,000 mortgage loan observations from Las Vegas, Nevada, single family homes, 2010

default = 1 if payment late by 90+ days, arm = 1 if adjustable rate mortgage, 0 if fixed, refinance = 1 if loan is for a refinance of any type (0 if for purchase), lien2 = 1 if 2nd lien mortage (0 if 1st lien), term30 = 1 if 30 year mortgage (0 if 15 year mortgage), underwater = 1 if borrower estimated to owe more than property worth at time of observing (0 otherwise), ltv loan to value ratio of property at origination (percent), rate current interest rate on loan (percent), amount loan amount in $10,000 units, fico borrower's credit score at origination.

Based on the statistical summary table, 37% of households' payments were late by 90+ days, 38% of households had adjustable rate mortgages, almost more than half of households had loans for refinance of any type. In addition, 10% of households had 2nd lien mortgages, 85% of households had 30-yr mortgages, 82% of households owed more than the property worth at time of observing. Mean loan-to-value ratio is 69% which is pretty good, with median 78% and max 109%. Mean rate on loans is 5.98%, with min, median and max 0.5%, 6.2% and 17%, respectively. A mean amount of borrowing is $245,600, with min of $11,000, median $211,000, max $7,500,000. Regarding FICO scores, mean score is 685, with min, median and max

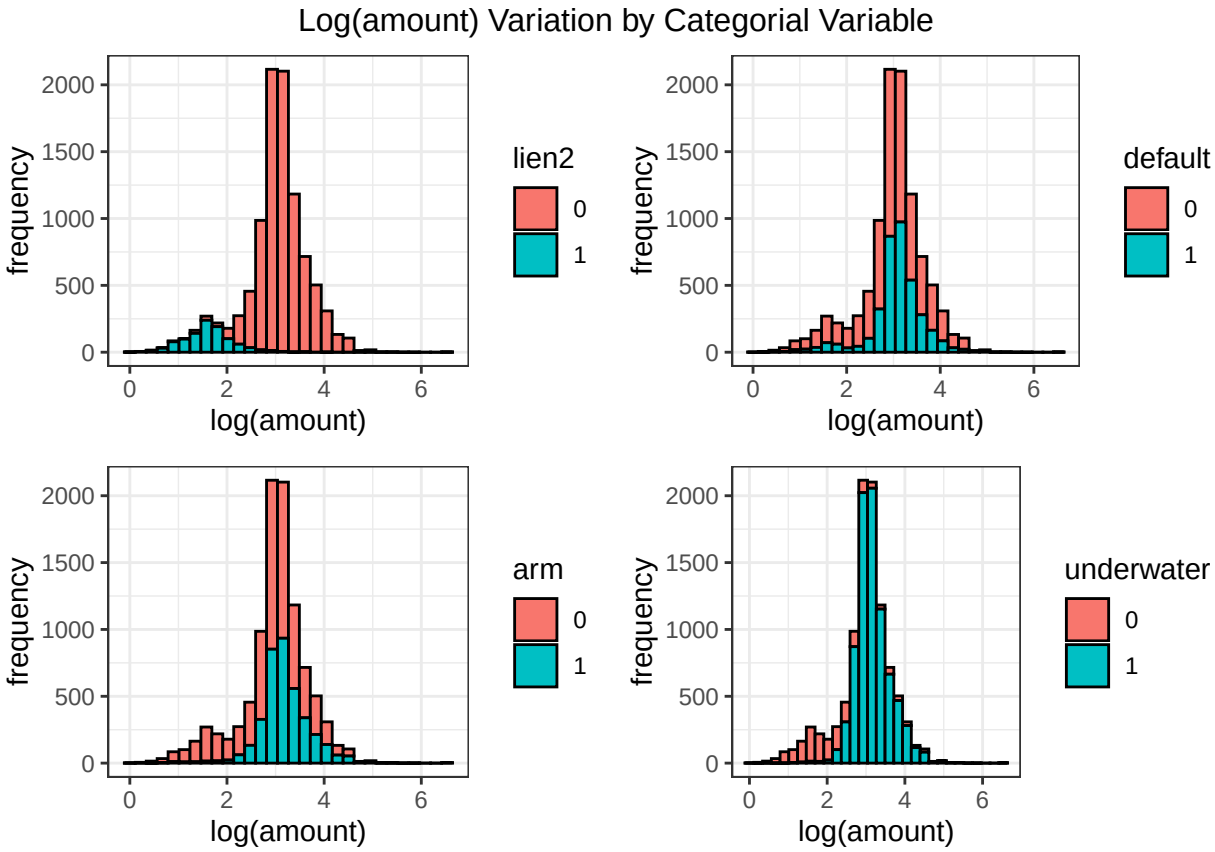
Table 1: Statistical Summary

	n	mean	sd	median	min	max	skew	kurtosis	se
default	10000	0.369900	0.4828015	0.00	0.0000	1.000	0.5388839	-1.7097750	0.0048280
arm	10000	0.378300	0.4849872	0.00	0.0000	1.000	0.5018187	-1.7483528	0.0048499
refinance	10000	0.545100	0.4979867	1.00	0.0000	1.000	-0.1811112	-1.9673954	0.0049799
lien2	10000	0.103500	0.3046260	0.00	0.0000	1.000	2.6029314	4.7757293	0.0030463
term30	10000	0.853500	0.3536245	1.00	0.0000	1.000	-1.9990962	1.9965853	0.0035362
underwater	10000	0.822800	0.3818570	1.00	0.0000	1.000	-1.6905163	0.8579313	0.0038186
ltv	10000	69.255427	21.0332339	78.25	5.0000	109.000	-1.5370436	1.5616383	0.2103323
rate	10000	5.979238	2.0634614	6.25	0.5000	17.375	0.3058574	0.5317328	0.0206346
amount	10000	24.561284	20.6957365	21.10	1.0948	750.000	12.5964812	365.0683257	0.2069574
fico	10000	684.731400	66.1149497	687.00	442.0000	823.000	-0.3497857	-0.3290018	0.6611495

being 442, 687, 823, respectively. Variables that deal with dollar values are not normally distributed based on skewness and kurtosis stats.

Histograms

```
a5 <- ggplot(vegas5) +
  geom_histogram(aes(x = log(amount), fill = factor(lien2)), colour = "black")+
            theme_bw() + ylab("frequency") + labs(fill = "lien2")

a6 <- ggplot(vegas5) +
  geom_histogram(aes(x = log(amount), fill = factor(default)),
                 colour = "black") + labs(fill = "default") +
  theme_bw() + ylab("frequency")

a7 <- ggplot(vegas5) +
  geom_histogram(aes(x = log(amount), fill = factor(arm)),
                 colour = "black") + labs(fill = "arm") + theme_bw() +
  ylab("frequency")

a8 <- ggplot(vegas5) +
  geom_histogram(aes(x = log(amount), fill = factor(underwater)),
                 colour = "black") + labs(fill = "underwater") + theme_bw() +
  ylab("frequency")

grid.arrange(a5,a6,a7,a8, top = "Log(amount) Variation by Categorial Variable")
```

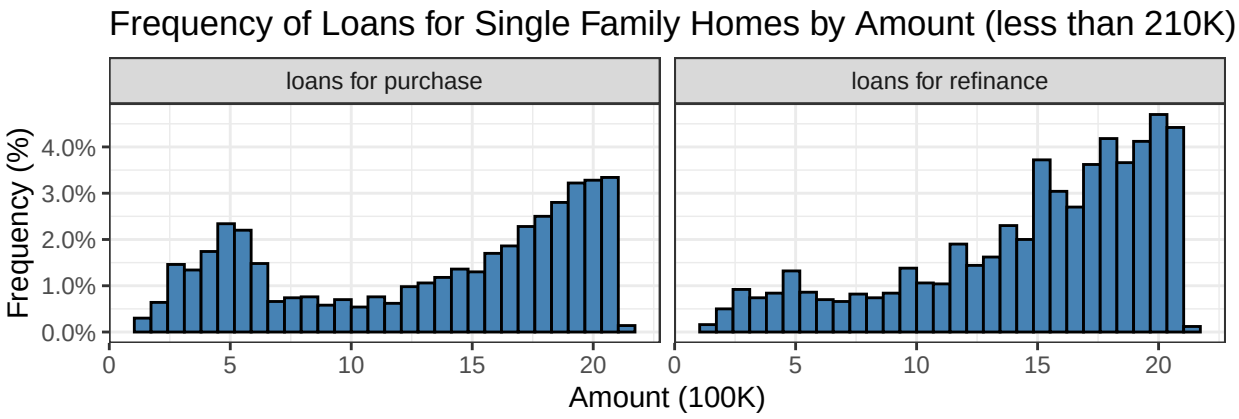


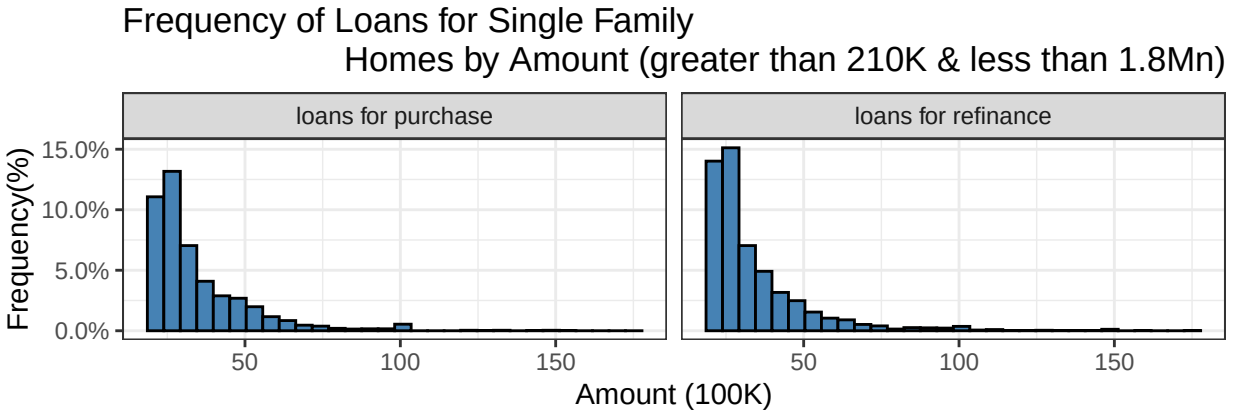
The lien graph shows that there are far less number of 2nd lien mortgages than 1st lien mortgages. And the

loan amount of those 2nd lien mortgages are also smaller than the amount of the 1st lien mortgages. From the other 3 graphs we can see that:

1) there are less number of defaults than those that are paid on time;
2) there are less adjustable rate mortgages than fixed rate mortgages;
3) there are slightly less underwater mortgages than those that are not;
4) interestingly, the two cases for the 3 variables, default, arm, underwater, exhibits the similar variation.

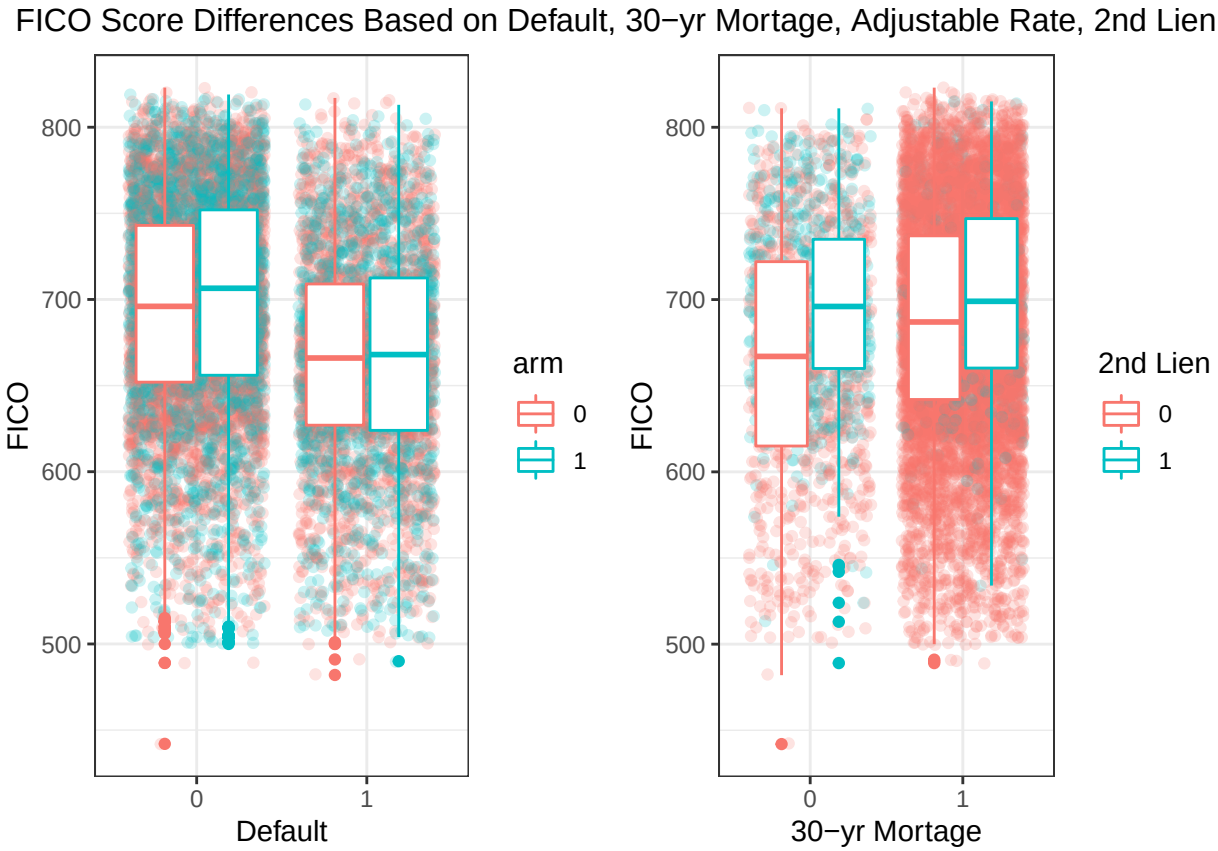
```
fd  <-  function(x) {
  n=length(x)
  r=IQR(x)
  2*r/n^(1/3)
}

a1 <- vegas5 %>% filter(amount < median(amount)) %>% ggplot(aes(x = amount)) +
  geom_histogram(colour = "black", fill = "steelblue") +
  aes(y = stat(count)/sum(stat(count))) +
  scale_y_continuous(labels=scales::percent) +
  ylab("Frequency (%)") +
  xlab("Amount (100K)") +
  facet_wrap(~refinance,
             labeller = labeller(refinance = c(`0`="loans for purchase",
                                               `1`="loans for refinance"))) +
  theme_bw() + ggtitle("Frequency of Loans for Single Family Homes by Amount (less than 210K)")


a2 <- vegas5 %>% filter(amount > median(amount) & amount < 180) %>% ggplot(aes(x = amount)) +
  geom_histogram( colour = "black", fill = "steelblue") +
  aes(y = stat(count)/sum(stat(count))) +
  scale_y_continuous(labels=scales::percent) + xlab("Amount (100K)") +
  ylab("Frequency(%)") + facet_wrap(~refinance,
                          labeller = labeller(refinance = c(`0`="loans for purchase",
                                                            `1`="loans for refinance"))) +
  theme_bw() + ggtitle("Frequency of Loans for Single Family
                       Homes by Amount (greater than 210K & less than 1.8Mn)")
grid.arrange(a1, a2)
```





In this plot, three variables are plotted: amount, frequency of occurrence and refinance. We can observe a peculiar trend when plotting observations based on the filters. Households with less than 210K borrowed seem to exhibit a left-skewness whereas households who borrowed between 210K and 1.8Mn, skewness to the right, which indicates that both distributions are not normal.

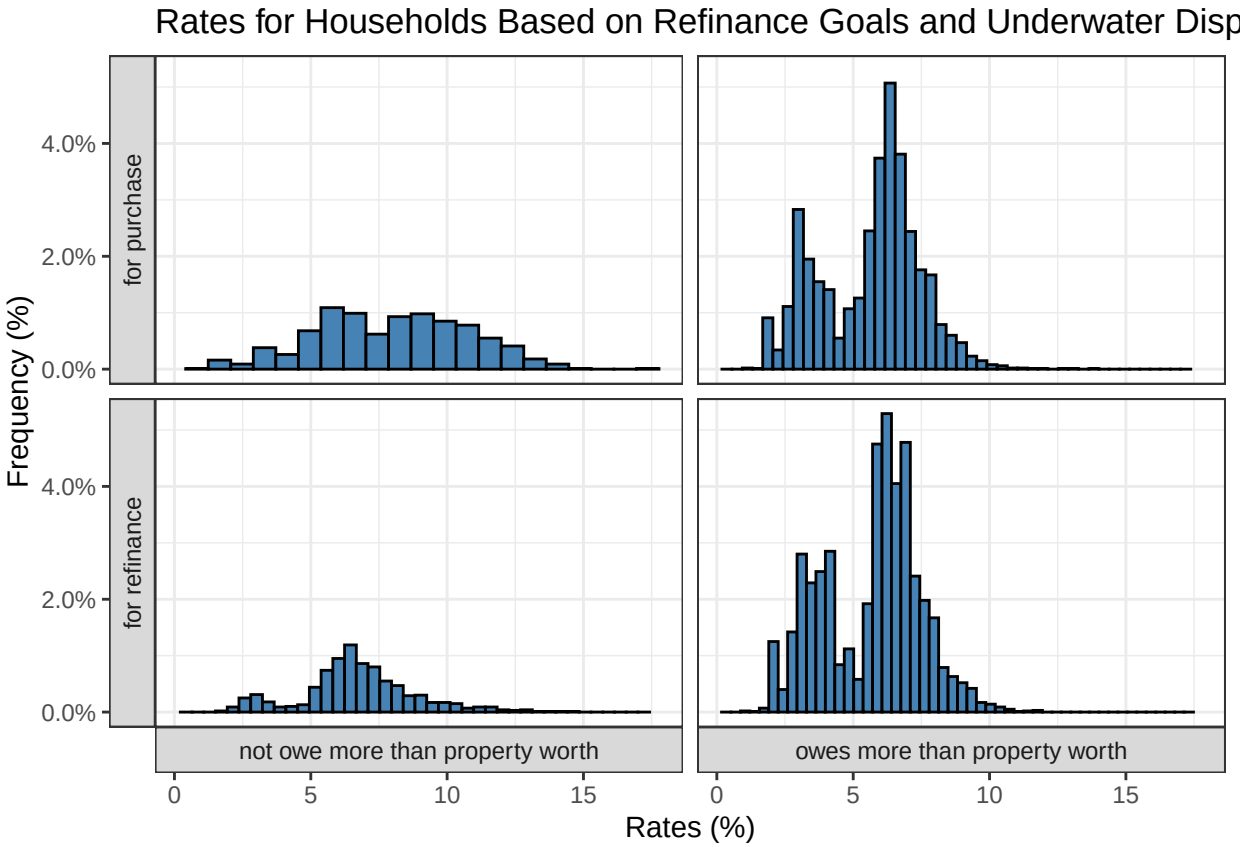
```
a4 <- ggplot(vegas5, aes(x = factor(default),
                    y = fico, colour = factor(arm))) + geom_jitter(alpha = 0.20) +
   geom_boxplot() + xlab("Default") +
  theme_bw() + labs(colour = "arm") +
  ylab("FICO")


a3 <- ggplot(vegas5,aes(x = factor(term30), y = fico, colour = factor(lien2))) +
  geom_jitter(alpha = 0.20) +
 geom_boxplot() + ylab("FICO") + xlab("30-yr Mortage") +
  labs(colour = "2nd Lien") +
   theme_bw()

grid.arrange(a4,a3, ncol = 2,
             top = "FICO Score Differences Based on Default, 30-yr Mortage, Adjustable Rate, 2nd Lien")
```



In this plot, five variables are plotted in the form of boxplots. On the left, we can see FICO scores tend to be higher for those who were not late on their payments (90+ days) and with adjustable rate mortgages (quite surprising). This may be due to the fact that interest rates might have been lower during that period. What is not surprising to observe is that FICO scores for those who were late on their payments are lower, but tend to be similar whether those with lower scores had an adjustable rates or not.

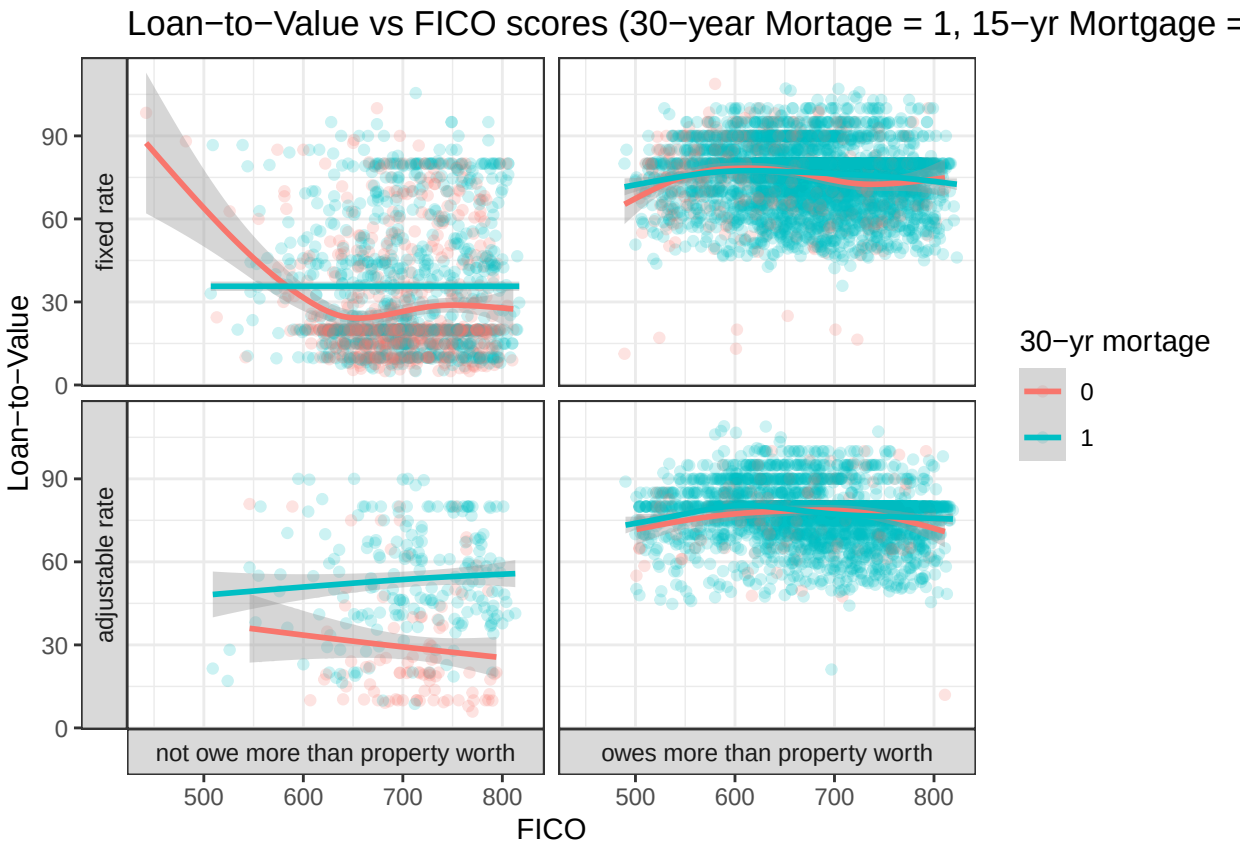
On the right, we can see that individuals who applied for a 15-year mortgage have a slightly higher FICO score than the ones with 30-year mortgage. This is intuitive because 15-year mortgages tend to have a higher periodic payment than 30-year, which banks would understandably require a higher FICO score. And in the same boxplot, we can see that 1st lien mortgages have a noticeably lower FICO score than the 2nd lien mortgages. This is the case probably because in the case of default, 1st lien would be paid in full before 2nd lien would be paid. Therefore 2nd lien would require a higher FICO score.

```
fd(rate)
ggplot(vegas5) +
  geom_histogram(aes(x = rate), binwidth = fd, colour = "black", fill = "steelblue") +
  aes(y = ..count../sum(count)) +
  scale_y_continuous(labels = scales::percent) +
  ylab("Frequency (%)") + xlab("Rates (%)") + facet_grid(refinance~underwater, switch = "both",
                                                         labeller = labeller(
                                                           refinance = c(`0` = "for purchase",
                                                                         `1`= "for refinance"),
                                                           underwater = c(`0` = "not owe more than property worth",
                                                                          `1` = "owes more than property worth")
                                                         )) + theme_bw() +
  ggtitle("Rates for Households Based on Refinance Goals and Underwater Disposition")
```



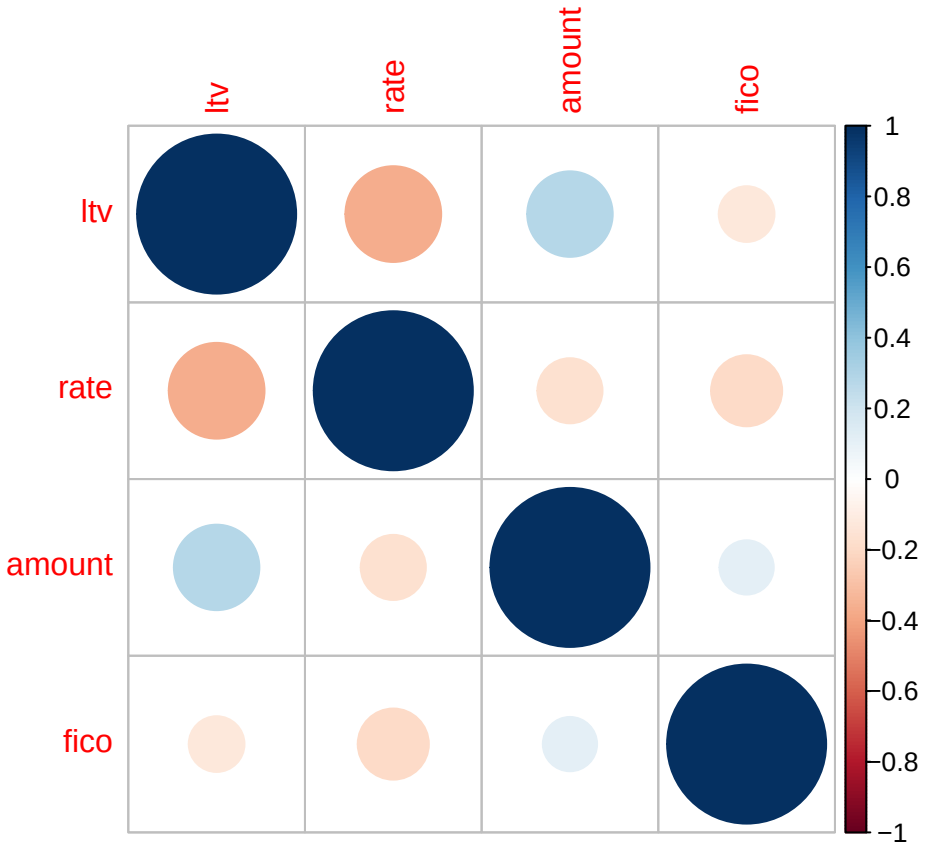
As we can see, households who owe more than property worth tend to have higher mortgage rates than their non-underwater counterparts.

```
ggplot(vegas5, aes(x = fico, y = ltv, colour = factor(term30))) +
  geom_point(alpha =0.20) +
  geom_smooth() +
  facet_grid(arm~underwater,
             switch = "both",
             labeller = labeller(arm = c(`0` = "fixed rate",
                                         `1`= "adjustable rate"),
               underwater = c(`0` = "not owe more than property worth",
                              `1` = "owes more than property worth"))) +
  xlab("FICO") + ylab("Loan-to-Value") +
  labs(colour = "30-yr mortage") +
  ggtitle("Loan-to-Value vs FICO scores (30-year Mortage = 1, 15-yr Mortgage = 0)") +
  theme_bw()
```



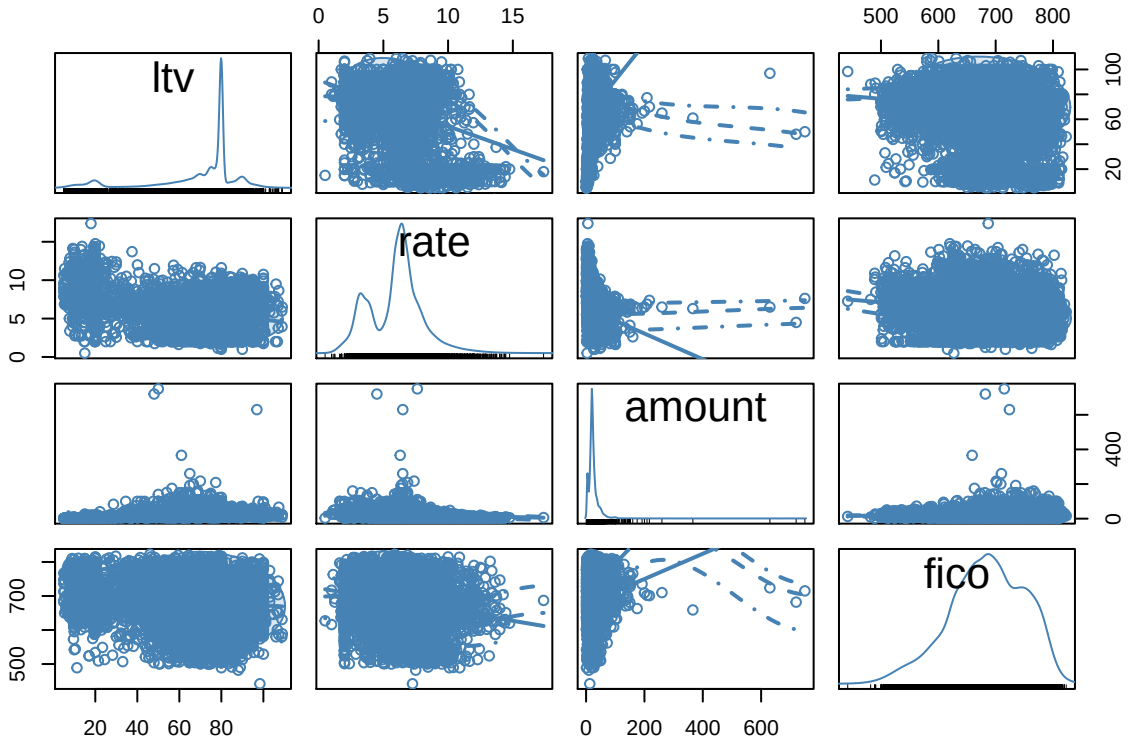
Here we can observe that households who owe more than property worth have higher loan-to-value ratios, but the relationship betwee FICO scores and ltv's is unclear where as for those who do not owe more than property worth tend to have lower ltv ratios and the relationship between FICO scores and ltv's is somewhat negative, specifically for those who did not take out a 30-yr mortgage in the first place. Although it looks somewhat positive for those who took out 30-yr mortgages. We can deduce that there is a larger concentration of households who possess 30-yr mortages and owe more than property worth. This potentially can be a byproduct of the 2007-2008 housing market crash, i.e. people owing more than their house is worth.

```
#correlation plot
corrplot(cor(vegas5[,7:10]))
```



There are some interesting correlations between variables in the correlation plot. First, we can see a slight positive correlation between FICO score and loan amount. This is explainable because banks want higher credibility for larger loans. We can also see a negative correlation between FICO score and mortgage rate. This might be explained by banks being more willing to give low rates to more creditworthy borrowers. Additionally, we can see a small positive correlation between FICO score and loan amount. Obviously, banks tend to grant larger loans to borrowers with a higher credit score.

```
#scatterplot
scatterplotMatrix(vegas5[,7:10], col = "steelblue",
                  ellipse = TRUE)
```



Similar to descriptions for the correlation plot above, this scatterplot matrix confirms our initial observations regarding variable correlations. First, looking at the upper diagonal of the matrix, we can see that rate and loan-to-value are negatively correlated. Second, amount and loan-to-value tend to be positively correlated since the regression line is seen on the left corner inside the micro-scatter plot. Third, there might be a negative correlation between FICO and loan-to-value, it is so small that it looks almost horizontal, which leads us to believe that this correlation is negligible.

Question 2

Estimate a multiple linear regression model that includes all the main effects only (i.e., no interactions nor higher order terms). We will use this model as a baseline. Comment on the statistical and economic significance of your estimates. Also, make sure to provide an interpretation of your estimates.

```
model.lm<- lm(default~arm+refinance+lien2+term30+underwater+ltv+rate+amount+fico)
#testing against GLM model due to binomial coef
model.glm<- glm(default~arm+refinance+lien2+term30+underwater+ltv+rate+amount+fico, family = "binomial")
#standard model summaries
summary(model.lm)
```

```
##
## Call:
## lm(formula = default ~ arm + refinance + lien2 + term30 + underwater +
##     ltv + rate + amount + fico)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8530 -0.3712 -0.2104  0.4883  1.1859
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.514e-01  6.942e-02   6.503 8.26e-11 ***
## arm         -2.389e-02  1.004e-02  -2.379   0.0174 *
## refinance   -4.835e-02  9.737e-03  -4.965 6.97e-07 ***
## lien2        1.821e-01  2.615e-02   6.962 3.57e-12 ***
## term30      -1.744e-02  1.426e-02  -1.223   0.2212
## underwater   1.782e-01  1.937e-02   9.199  < 2e-16 ***
## ltv          4.986e-03  4.174e-04  11.947  < 2e-16 ***
## rate         4.003e-02  2.586e-03  15.484  < 2e-16 ***
## amount      -2.322e-04  2.349e-04  -0.988   0.3230
## fico        -1.133e-03  7.528e-05 -15.049  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4543 on 9990 degrees of freedom
## Multiple R-squared:  0.1153, Adjusted R-squared:  0.1145
## F-statistic: 144.7 on 9 and 9990 DF,  p-value: < 2.2e-16
```

```
summary(model.glm)
```

```
##
## Call:
## glm(formula = default ~ arm + refinance + lien2 + term30 + underwater +
##     ltv + rate + amount + fico, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1174  -0.9403  -0.6598   1.1405   2.7382
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.8306594  0.3495402  -2.376  0.01748 *
## arm         -0.1434127  0.0483486  -2.966  0.00301 **
## refinance   -0.2404201  0.0476051  -5.050 4.41e-07 ***
## lien2        1.2776049  0.1580227   8.085 6.22e-16 ***
## term30      -0.0698832  0.0724898  -0.964  0.33503
## underwater   1.2418800  0.1213201  10.236  < 2e-16 ***
## ltv          0.0254911  0.0021861  11.660  < 2e-16 ***
## rate         0.1906607  0.0128088  14.885  < 2e-16 ***
## amount      -0.0013225  0.0013039  -1.014  0.31045
## fico        -0.0052253  0.0003692 -14.154  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13178  on 9999  degrees of freedom
## Residual deviance: 11912  on 9990  degrees of freedom
## AIC: 11932
```

```
##
## Number of Fisher Scoring iterations: 4
```

```
#stargazer quick comparison
stargazer(model.lm, model.glm,
          type = "text",
          column.labels = c("LM", "GLM"),
          model.names=FALSE, dep.var.labels.include=FALSE,
          dep.var.caption      =c("models"), digits=5,
          intercept.top = TRUE, intercept.bottom = FALSE)
```

```
##
## ===============================================================
##                                     models
##                        ---------------------------------------
##                                    LM                  GLM
##                                    (1)                 (2)
## ---------------------------------------------------------------
## Constant                        0.45142***          -0.83066**
##                                  (0.06942)           (0.34954)
##
## arm                             -0.02389**          -0.14341***
##                                  (0.01004)           (0.04835)
##
## refinance                       -0.04835***         -0.24042***
##                                  (0.00974)           (0.04761)
##
## lien2                           0.18205***           1.27760***
##                                  (0.02615)           (0.15802)
##
## term30                           -0.01744            -0.06988
##                                  (0.01426)           (0.07249)
##
## underwater                      0.17818***           1.24188***
##                                  (0.01937)           (0.12132)
##
## ltv                             0.00499***           0.02549***
##                                  (0.00042)           (0.00219)
##
## rate                            0.04003***           0.19066***
##                                  (0.00259)           (0.01281)
##
## amount                           -0.00023            -0.00132
##                                  (0.00023)           (0.00130)
##
## fico                            -0.00113***         -0.00523***
##                                  (0.00008)           (0.00037)
##
## ---------------------------------------------------------------
## Observations                      10,000              10,000
## R2                                0.11530
## Adjusted R2                       0.11450
## Log Likelihood                                     -5,956.21100
## Akaike Inf. Crit.                                  11,932.42000
```

```
## Residual Std. Error     0.45432 (df = 9990)
## F Statistic         144.65730*** (df = 9; 9990)
## ===============================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

Insignificant variables: arm, term30, amount

Arm: interesting that arm is not significant since our intuition is that people who obtain an arm loan with a lower initial payment would cause higher defaults if the arm causes mortgage payments to increase from its initial value

term30: may be insignificant for a few reasons. First, the amount of records with 15 years is low compared the entire data set. Second, borrowers with 15 year loans may be more financially fit than 30 year borrowers. Third, the borrow can refinance to a 30 years to lower payments if needed.

amount: we are surprised that amount is insignificant as my intuition would be borrowers with high loan amounts would have a higher likelihood to default

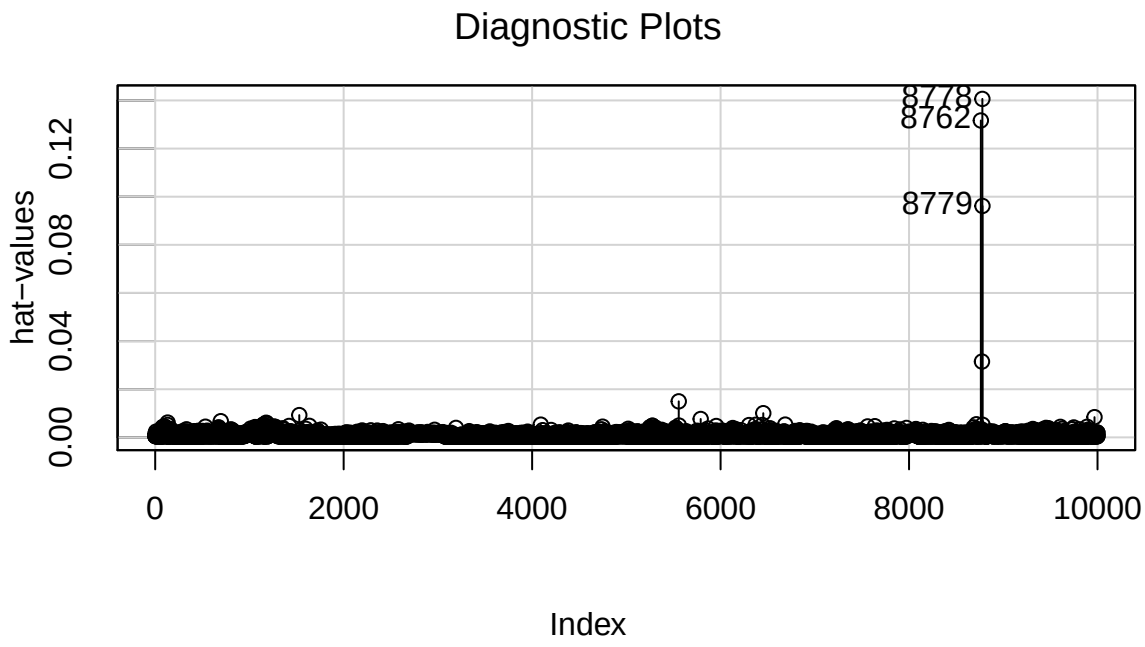
Question 3

Identify if there are any outliers, high leverage, and or influential observations worth removing. If so, remove them but justify your reason for doing so and re-estimate your model.

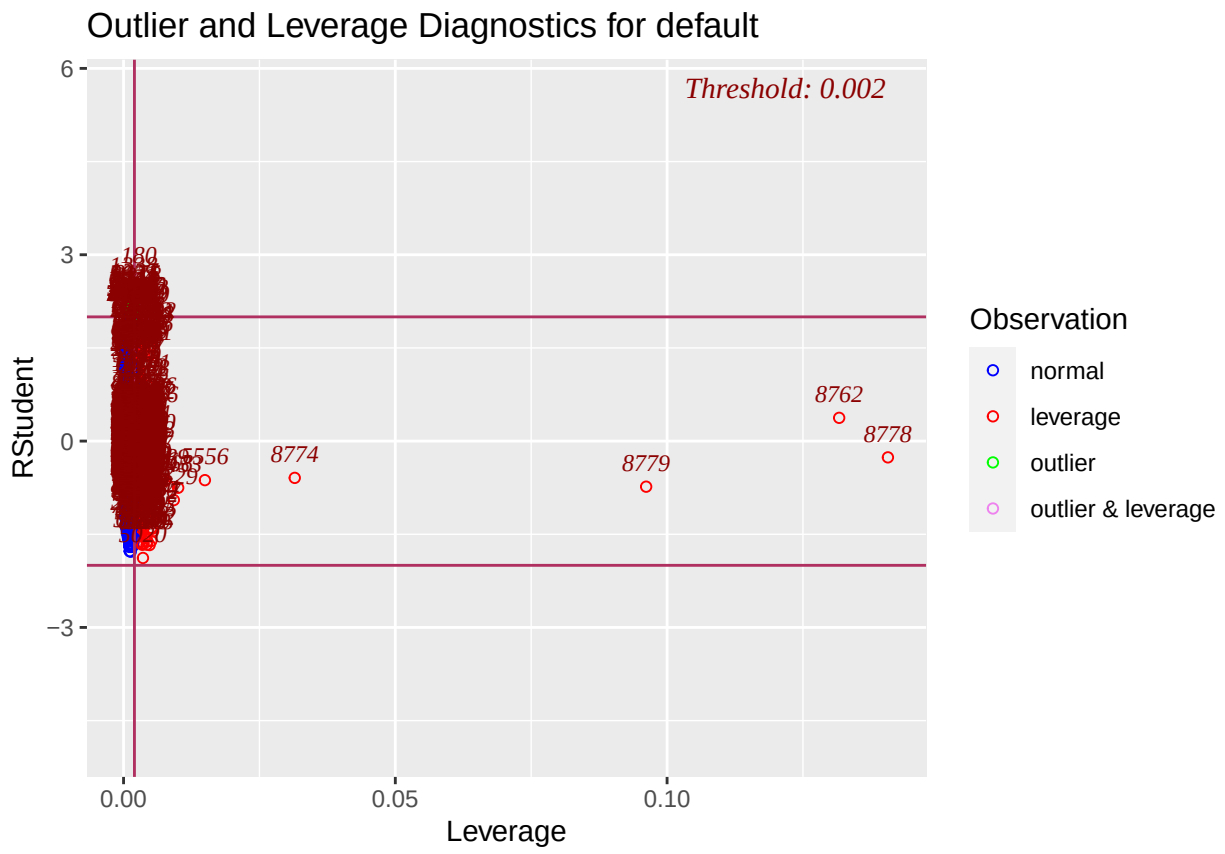
```
model.lm<- lm(default~arm+refinance+lien2+term30+underwater+ltv+rate+amount+fico)
#visual
# qqPlot(model.lm, id=list(n=3))
#Bonferroni test
outlierTest(model.lm)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 180 2.614688          0.0089443           NA
```

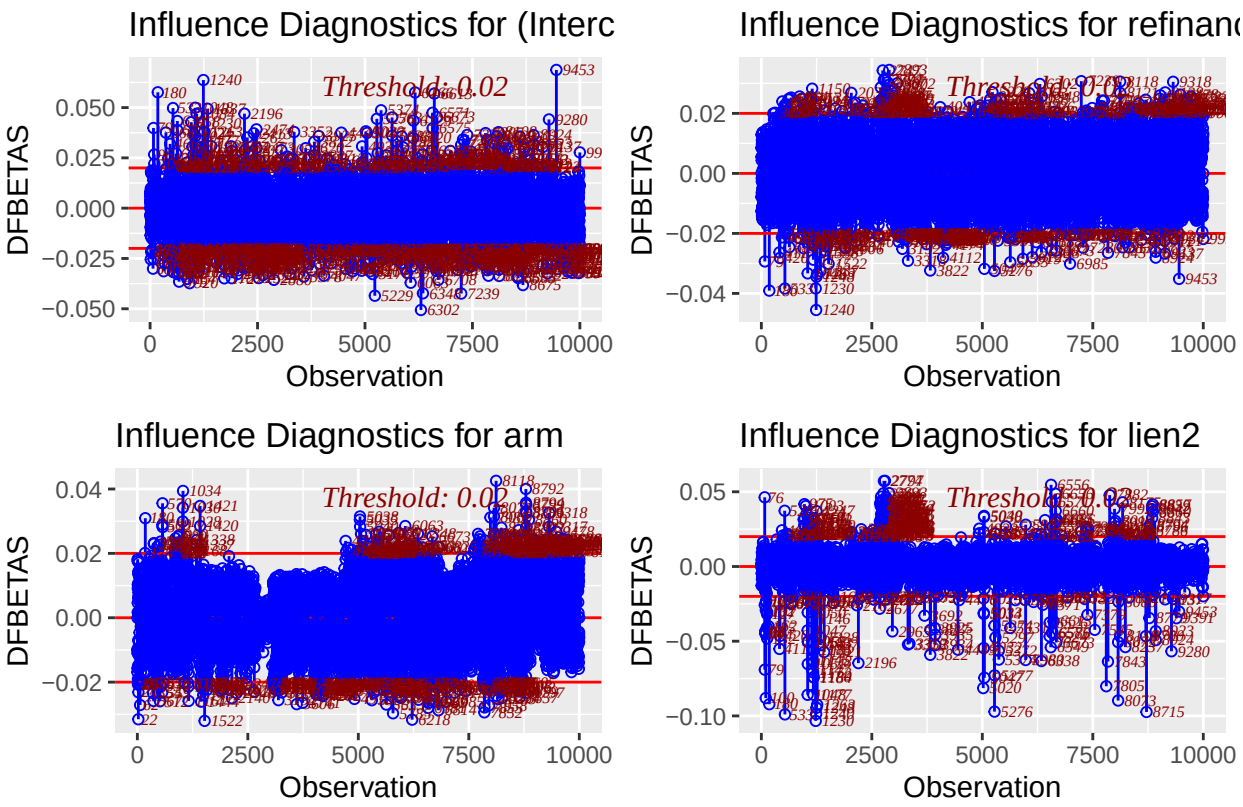
```
influenceIndexPlot(model.lm, id=list(n=3),vars="hat")
```

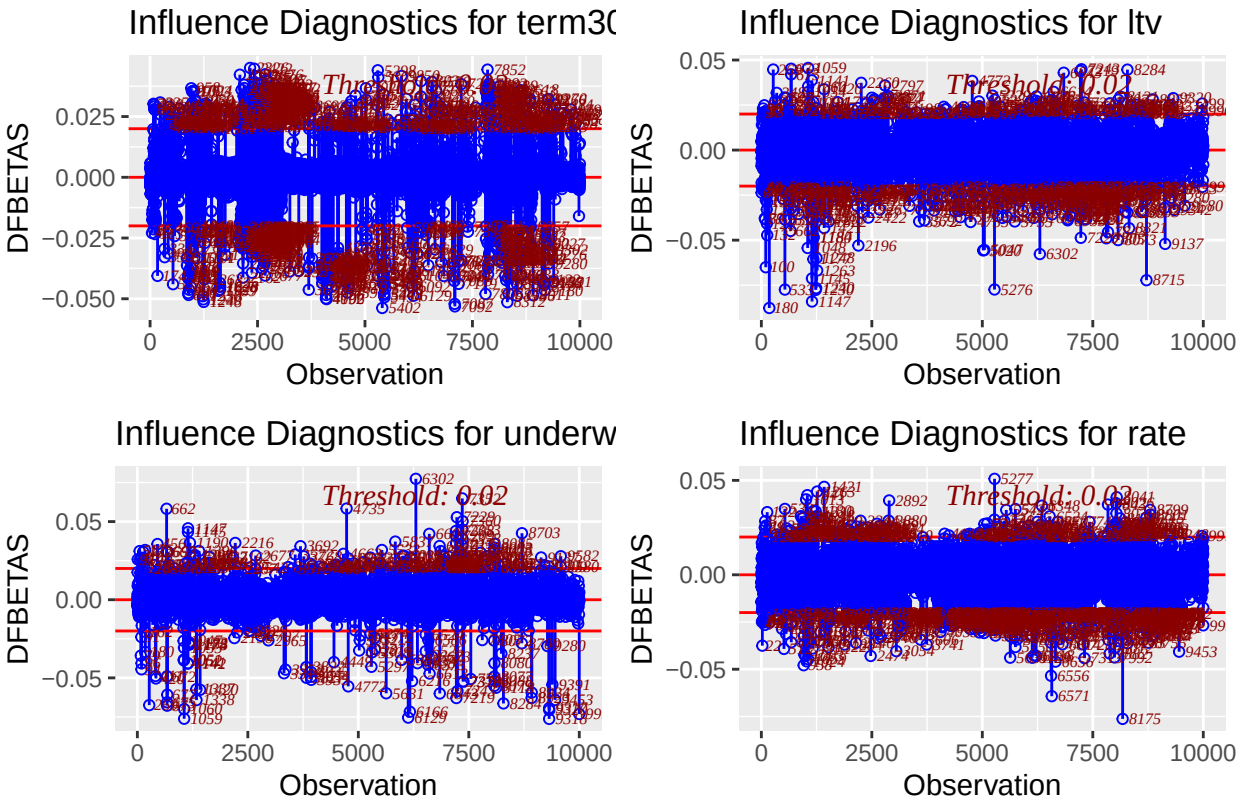


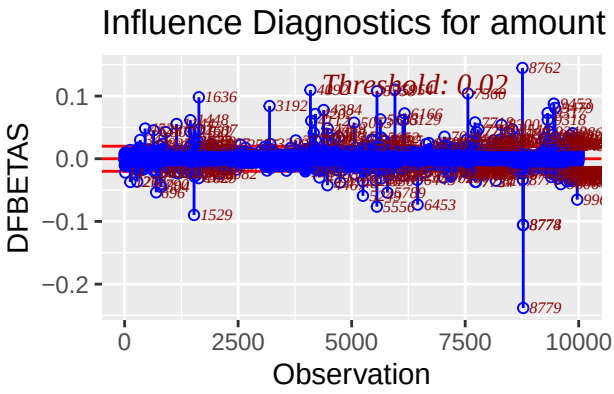

```
#ols plots
ols_plot_resid_lev(model.lm)
```

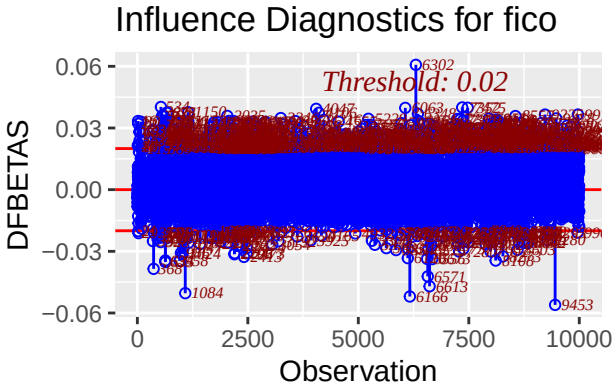



```
ols_plot_dfbetas(model.lm)
```

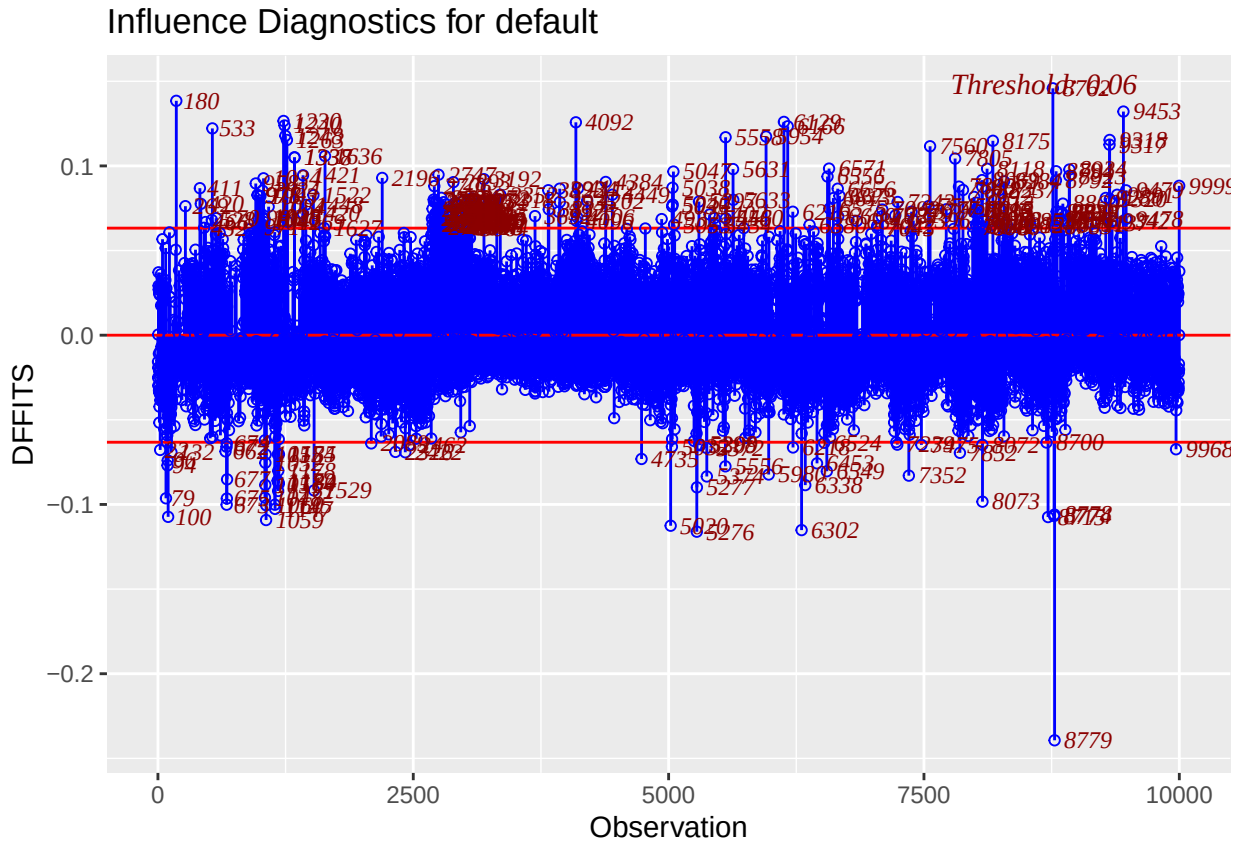

Influence Diagnostics for (Interc
Threshold: 0.02
DFBETAS
Observation
Influence Diagnostics for refinanc
Threshold: 0.02
DFBETAS
Observation
Influence Diagnostics for arm
Threshold: 0.02
DFBETAS
Observation
Influence Diagnostics for lien2
Threshold: 0.02
DFBETAS
Observation


Influence Diagnostics for term3(
Threshold: 0.0
DFBETAS
0.025
0.000
−0.025
−0.050
0
2500
5000
7500
10000
Observation
Influence Diagnostics for ltv
Threshold: 0.01
DFBETAS
0.05
0.00
−0.05
0
2500
5000
7500
10000
Observation
Influence Diagnostics for underw
Threshold: 0.02
DFBETAS
0.05
0.00
−0.05
0
2500
5000
7500
10000
Observation
Influence Diagnostics for rate
Threshold: 0.03
DFBETAS
0.05
0.00
−0.05
0
2500
5000
7500
10000
Observation






```
ols_plot_dffits(model.lm)
```

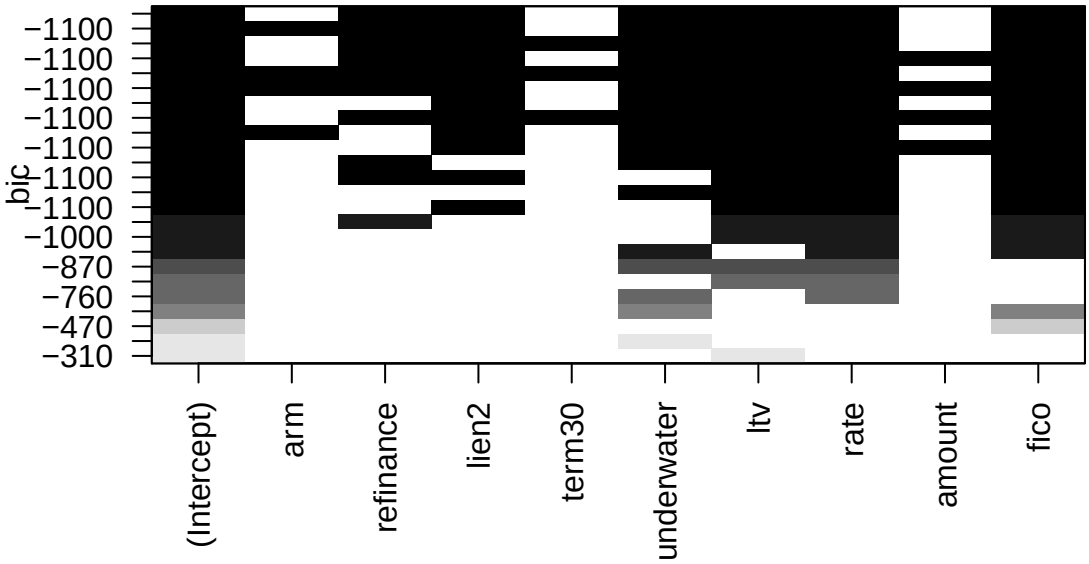


Based on the output in the table above, lm regression makes more sense If we set all the variables to zero, the probability of default equals 0.50, which is equivalent to a flip of a coin. This, intuitively, makes more sense, as it adds more randomness to the event and also because probabilities are positive values in the interval between (0,1). Therefore, we will be using the lm model to answer questions that follow.

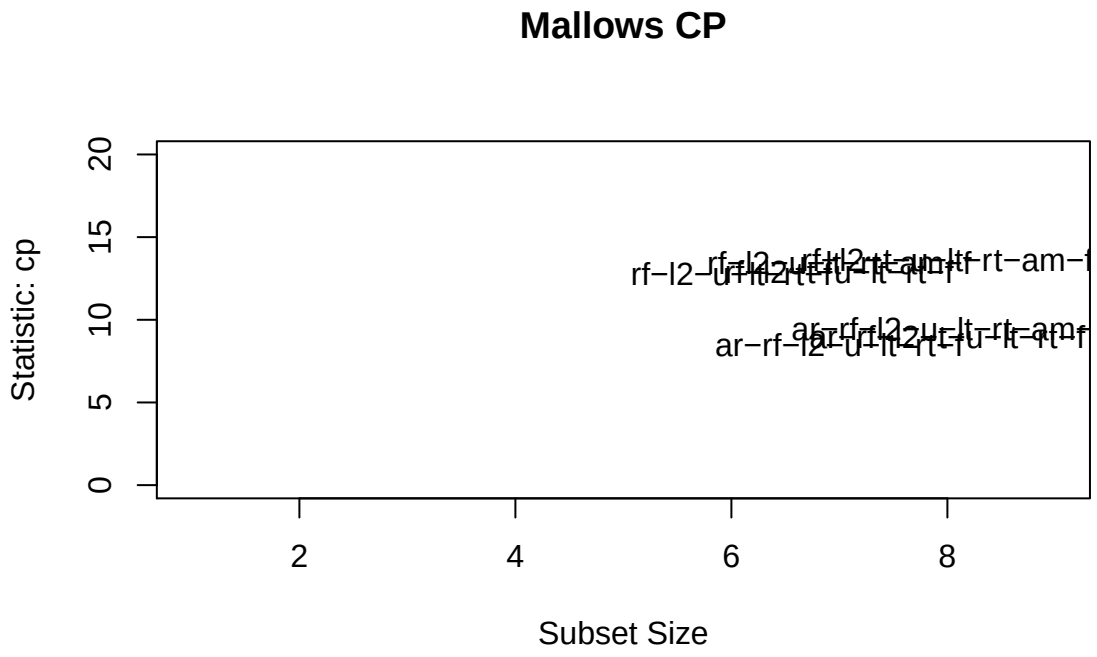
Question 4

Use Mallows Cp for identifying which terms you will keep in the model (based on part 3) and also use the Boruta algorithm for variable selection. Based on the two results, determine which subset of predictors you will keep.

```
lm.mallows.cp <- regsubsets(default~arm+refinance+lien2+term30+underwater+ltv+rate+amount+fico,
                            data =vegas5, method = c("exhaustive"), nbest = 3)

plot(lm.mallows.cp)
```

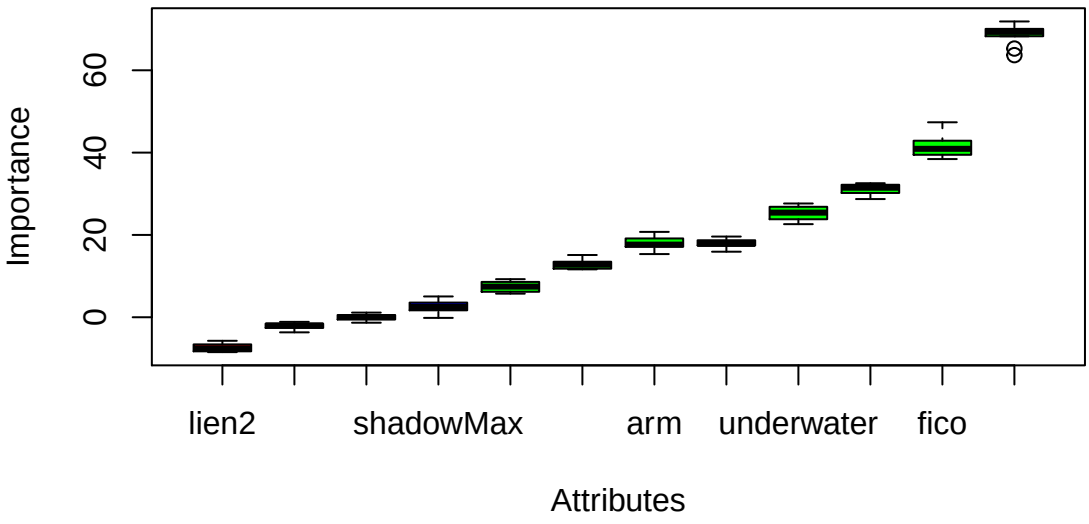



```
subsets(lm.mallows.cp, statistic = "cp", legend = F, main = "Mallows CP",
                   ylim = c(0,20))
```




```
##            Abbreviation
## arm                  ar
## refinance            rf
## lien2                l2
## term30                t
## underwater            u
## ltv                  lt
## rate                 rt
## amount               am
## fico                  f
```

```
Boruta.res <- Boruta(default~.,
                     data = vegas5,
                     doTrace = 3)

plot(Boruta.res)
```




```
attStats(Boruta.res)[order(-attStats(Boruta.res)$meanImp),]
```

```
##              meanImp medianImp    minImp    maxImp normHits  decision
## rate       68.651603 69.320352 63.674580 71.859346        1 Confirmed
## fico       41.355975 40.920879 38.426735 47.363959        1 Confirmed
## ltv        31.113518 31.472979 28.715295 32.559970        1 Confirmed
## underwater 25.353545 25.417128 22.617648 27.638791        1 Confirmed
## arm        17.994286 17.646905 15.349804 20.734043        1 Confirmed
## amount     17.830700 17.816142 15.923051 19.602818        1 Confirmed
## term30     12.964501 12.773108 11.632259 15.117017        1 Confirmed
## refinance   7.455327  7.445987  5.734732  9.251940        1 Confirmed
## lien2      -7.409504 -7.570815 -8.474821 -5.722263        0  Rejected
```

It looks like mallows CP prefers 7 predictor variables, they are: arm, refinance, lien2, underwater,ltv,rate,fico. In addition, based on Boruta's importance plot, lien2 is rejected as a predictor variable. Hence, we'll proceed to estimate a model without lien2.

Question 5

Test for multicollinearity using VIF on the model from (4) . Based on the test, remove any appropriate variables, and estimate a new regression model based on these findings.

```
updated.lm <- lm(default~arm+refinance+term30+underwater+ltv+rate+amount+fico,
              data = vegas5)

stargazer(model.lm, updated.lm,
          type = "text",
          column.labels = c("Original LM", "Updated LM"),
          model.names=FALSE, dep.var.labels.include=FALSE,
          dep.var.caption      =c("models"), digits=5,
          intercept.top = TRUE, intercept.bottom = FALSE)
```

```
##
## ================================================================================
##                                          models
##                        ------------------------------------------------
##                                Original LM                 Updated LM
##                                   (1)                         (2)
## --------------------------------------------------------------------------------
## Constant                       0.45142***                  0.59939***
##                                 (0.06942)                   (0.06624)
##
## arm                            -0.02389**                  -0.02542**
##                                 (0.01004)                   (0.01006)
##
## refinance                      -0.04835***                 -0.06033***
##                                 (0.00974)                   (0.00961)
##
## lien2                          0.18205***
##                                 (0.02615)
##
## term30                          -0.01744                   -0.03502**
##                                 (0.01426)                   (0.01407)
##
## underwater                     0.17818***                  0.15789***
##                                 (0.01937)                   (0.01919)
##
## ltv                            0.00499***                  0.00348***
##                                 (0.00042)                   (0.00036)
##
## rate                           0.04003***                  0.04318***
##                                 (0.00259)                   (0.00255)
##
## amount                          -0.00023                   -0.00041*
##                                 (0.00023)                   (0.00023)
##
## fico                           -0.00113***                 -0.00113***
##                                 (0.00008)                   (0.00008)
##
## --------------------------------------------------------------------------------
## Observations                     10,000                      10,000
## R2                               0.11530                     0.11100
## Adjusted R2                      0.11450                     0.11029
## Residual Std. Error       0.45432 (df = 9990)         0.45540 (df = 9991)
## F Statistic          144.65730*** (df = 9; 9990) 155.94000*** (df = 8; 9991)
```

```
## ================================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

```
vif(updated.lm)
```

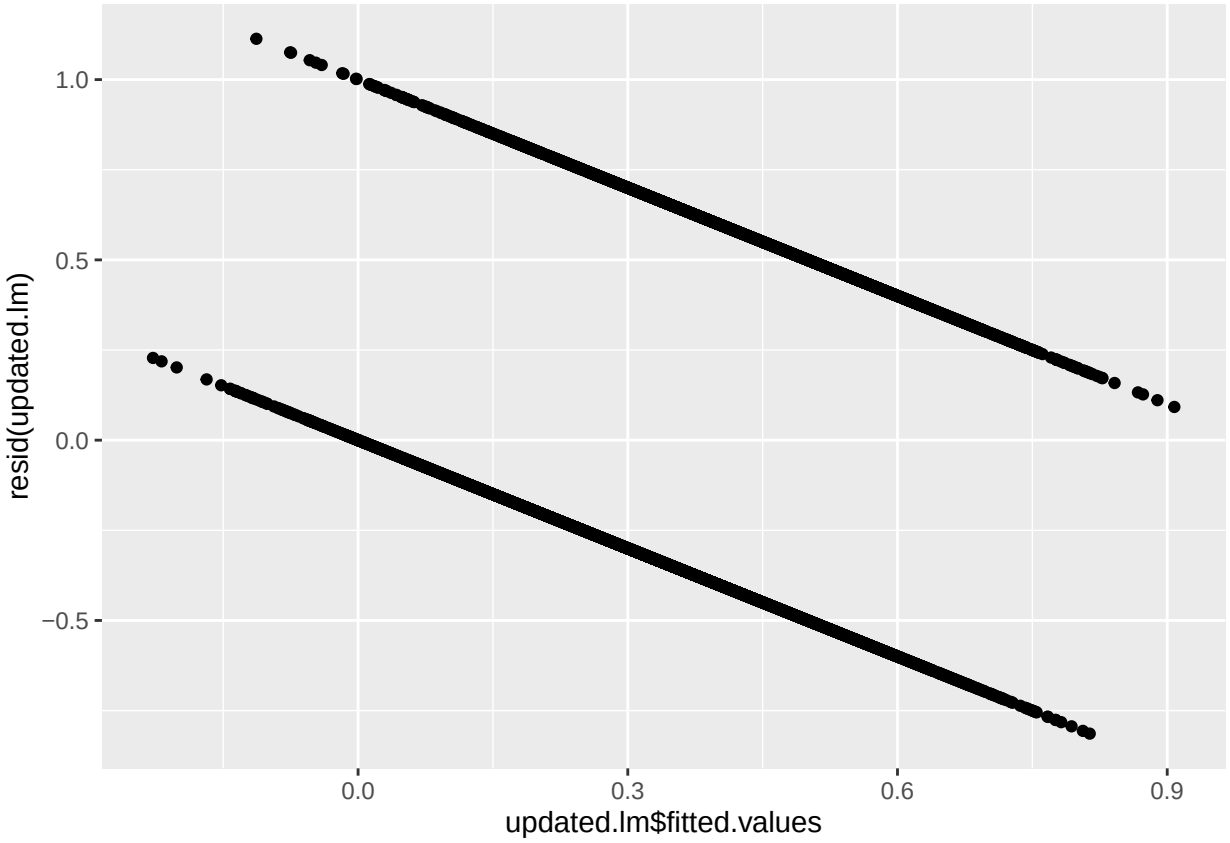
```
##        arm  refinance     term30 underwater        ltv       rate     amount
##   1.148701   1.103395   1.192906   2.590110   2.724746   1.336723   1.131893
##       fico
##   1.199913
```

It seems as though VIF's look fine and none are above 4 which we use as a threshold in this case. Therefore, we will not be removing any variables since there is not a reason to believe that there is strong multicollinearity. Comparing the updated model to the original, most of the variances have decreased

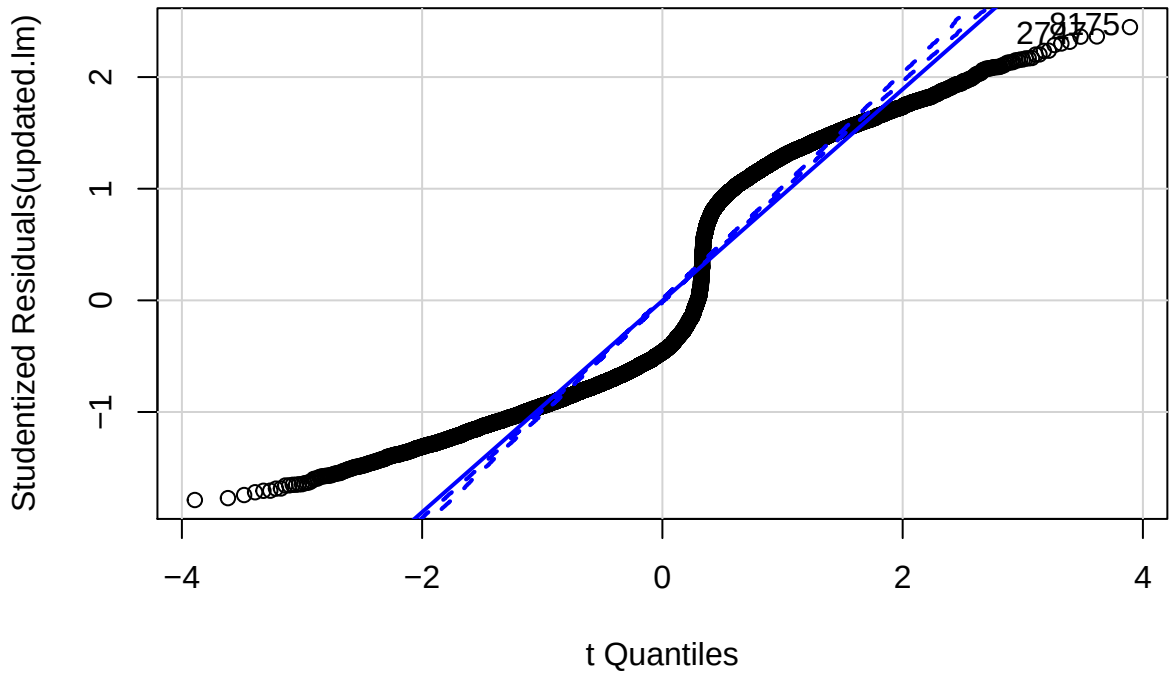
Question 6

For your model in part (5) plot the respective residuals vs. $\hat{y}$ and comment on your results.

```
ggplot() +
  geom_point(aes(x = updated.lm$fitted.values,
                 y = resid(updated.lm)))
```



```
qqPlot(updated.lm)
```




```
## [1] 2747 8175
```

Nothing too suspicious. QQplot reveals that our observations follow a logistic curve.

Question 7

For your model in part (5) perform a RESET test and comment on your results.

```
resettest(updated.lm, power = 2, type = "regressor")
```

```
##
##  RESET test
##
## data:  updated.lm
## RESET = 13.603, df1 = 8, df2 = 9983, p-value < 2.2e-16
```

Based on the results, there is strong evidence that we need to consider including higher-order terms.

Question 8

For your model in part (5) test for heteroskedasticity and comment on your results. If you identify heteroskedasticy, make sure to account for it before moving on to (9).

```
#accounting for heteroskedasticity

bptest(updated.lm)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  updated.lm
## BP = 520.51, df = 8, p-value < 2.2e-16
```

```
p <- fitted(updated.lm)
p[p<0.01] <- 0.01 #truncating probabilities that don't fall in (0,1) interval
p[p>0.99] <- 0.99
sigma.sq <- p*(1-p)
weight <- 1/sigma.sq

updated.lm.fgls <- lm(default~arm+refinance+term30+underwater+ltv+rate+amount+fico, #re-estimating the
              data = vegas5, weights = weight)
bptest(updated.lm.fgls)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  updated.lm.fgls
## BP = 520.51, df = 8, p-value < 2.2e-16
```

```
summary(updated.lm.fgls)
```

```
##
## Call:
## lm(formula = default ~ arm + refinance + term30 + underwater +
##     ltv + rate + amount + fico, data = vegas5, weights = weight)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9555 -0.7842 -0.6057  1.0013 10.1024
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.765e-01  5.681e-02  13.668  < 2e-16 ***
## arm         -1.150e-02  8.943e-03  -1.286   0.1984
## refinance   -7.447e-02  8.256e-03  -9.021  < 2e-16 ***
## term30      -1.804e-02  1.067e-02  -1.691   0.0908 .
## underwater   1.484e-01  1.414e-02  10.491  < 2e-16 ***
## ltv          1.879e-03  2.753e-04   6.826 9.22e-12 ***
## rate         3.463e-02  2.168e-03  15.976  < 2e-16 ***
## amount      -2.652e-04  1.086e-04  -2.441   0.0147 *
```

Table 2: Updated Linear Model-GLS with Boruta-selected Variables

term	estimate	std.error	statistic	p.value
(Intercept)	0.7765450	0.0568142	13.668145	0.0000000
arm	-0.0115033	0.0089431	-1.286271	0.1983784
refinance	-0.0744732	0.0082558	-9.020658	0.0000000
term30	-0.0180395	0.0106664	-1.691245	0.0908213
underwater	0.1483840	0.0141437	10.491163	0.0000000
ltv	0.0018792	0.0002753	6.826289	0.0000000
rate	0.0346328	0.0021678	15.975944	0.0000000
amount	-0.0002652	0.0001086	-2.441243	0.0146541
fico	-0.0011574	0.0000678	-17.068459	0.0000000

```
## fico        -1.157e-03  6.781e-05 -17.068  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.064 on 9991 degrees of freedom
## Multiple R-squared:  0.1799, Adjusted R-squared:  0.1792
## F-statistic:   274 on 8 and 9991 DF,  p-value: < 2.2e-16
```

```
kable(tidy(updated.lm.fgls), caption = "Updated Linear Model-GLS with Boruta-selected Variables")
```

```
#we believe we can do better in terms of coefficient significance
p1 <- fitted(updated.lm)
p1[p1 < 0.01 | p1 > 0.99] <- NA
sigsq <- p1*(1-p1)
w <- 1/sigsq

updated.lm.fgls.omit <- lm(default~arm+refinance+term30+underwater+ltv+rate+amount+fico, #re-estimating
              data = vegas5, weights = w)

summary(updated.lm.fgls.omit) #much better
```

```
##
## Call:
## lm(formula = default ~ arm + refinance + term30 + underwater +
##     ltv + rate + amount + fico, data = vegas5, weights = w)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0419 -0.7832 -0.5743  0.9915  8.1376
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.393e-01  6.401e-02  11.550  < 2e-16 ***
## arm         -1.773e-02  1.007e-02  -1.760  0.07842 .
## refinance   -7.181e-02  9.226e-03  -7.783 7.80e-15 ***
## term30      -3.347e-02  1.266e-02  -2.643  0.00823 **
## underwater   1.770e-01  1.634e-02  10.831  < 2e-16 ***
## ltv          2.030e-03  3.137e-04   6.470 1.02e-10 ***
```

Table 3: Linear GLS Model with Truncated Fitted Values

term	estimate	std.error	statistic	p.value
(Intercept)	0.7392803	0.0640088	11.549672	0.0000000
arm	-0.0177272	0.0100715	-1.760134	0.0784163
refinance	-0.0718051	0.0092259	-7.783002	0.0000000
term30	-0.0334738	0.0126650	-2.643020	0.0082301
underwater	0.1770240	0.0163444	10.830866	0.0000000
ltv	0.0020300	0.0003137	6.470433	0.0000000
rate	0.0401301	0.0024837	16.157192	0.0000000
amount	-0.0005120	0.0001600	-3.199236	0.0013823
fico	-0.0011729	0.0000735	-15.962670	0.0000000

```
## rate         4.013e-02  2.484e-03  16.157  < 2e-16 ***
## amount      -5.120e-04  1.600e-04  -3.199  0.00138 **
## fico        -1.173e-03  7.348e-05 -15.963  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.026 on 9807 degrees of freedom
##   (184 observations deleted due to missingness)
## Multiple R-squared:  0.1269, Adjusted R-squared:  0.1262
## F-statistic: 178.1 on 8 and 9807 DF,  p-value: < 2.2e-16
```

```
kable(tidy(updated.lm.fgls.omit),
      caption = "Linear GLS Model with Truncated Fitted Values") #this is the model we will keep
```

```
#linear hypothesis test indicates that we reject H0
linearHypothesis(updated.lm.fgls.omit,
                 hypothesis.matrix = c("term30",
                                       "arm",
                                       "amount"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## term30 = 0
## arm = 0
## amount = 0
##
## Model 1: restricted model
## Model 2: default ~ arm + refinance + term30 + underwater + ltv + rate +
##     amount + fico
##
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1   9810 10341
## 2   9807 10317  3    23.931 7.5829 4.611e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#here, linear hypothesis test shows that arm = 0
linearHypothesis(updated.lm.fgls.omit,
                 hypothesis.matrix = c("arm"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## arm = 0
##
## Model 1: restricted model
## Model 2: default ~ arm + refinance + term30 + underwater + ltv + rate +
##     amount + fico
##
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1   9808 10320
## 2   9807 10317  1    3.2591 3.0981 0.07842 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#to summarize

#original vs best fit
stargazer(model.lm, updated.lm.fgls,
          type = "text",
          column.labels = c("Original", "GLS Best Fit"),
          model.names=FALSE, dep.var.labels.include=FALSE,
          dep.var.caption      =c("Original VS Best Fit"), digits=5,
          intercept.top = TRUE, intercept.bottom = FALSE,
          out.header=FALSE, column.sep.width="1pt")
```

```
##
## ================================================================================
##                                      Original VS Best Fit
##                         ------------------------------------------------------
##                                  Original                    GLS Best Fit
##                                     (1)                          (2)
## --------------------------------------------------------------------------------
## Constant                         0.45142***                   0.77654***
##                                   (0.06942)                    (0.05681)
##
## arm                              -0.02389**                    -0.01150
##                                   (0.01004)                    (0.00894)
##
## refinance                        -0.04835***                  -0.07447***
##                                   (0.00974)                    (0.00826)
##
## lien2                            0.18205***
##                                   (0.02615)
##
## term30                            -0.01744                     -0.01804*
##                                   (0.01426)                    (0.01067)
##
## underwater                       0.17818***                   0.14838***
```

```
##                                (0.01937)                  (0.01414)
##
## ltv                           0.00499***                 0.00188***
##                                (0.00042)                  (0.00028)
##
## rate                          0.04003***                 0.03463***
##                                (0.00259)                  (0.00217)
##
## amount                         -0.00023                  -0.00027**
##                                (0.00023)                  (0.00011)
##
## fico                          -0.00113***                -0.00116***
##                                (0.00008)                  (0.00007)
##
## ------------------------------------------------------------------------------
## Observations                    10,000                     10,000
## R2                              0.11530                    0.17990
## Adjusted R2                     0.11450                    0.17925
## Residual Std. Error     0.45432 (df = 9990)         1.06422 (df = 9991)
## F Statistic         144.65730*** (df = 9; 9990) 273.96510*** (df = 8; 9991)
## ==============================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

```
#additional models considered
stargazer(model.glm, updated.lm, updated.lm.fgls.omit,
          type = "text",
          column.labels = c("Logistic", "Boruta", "GLS-Omit"),
          model.names=FALSE, dep.var.labels.include=FALSE,
          dep.var.caption      =c("Alternative models"), digits=5,
          intercept.top = TRUE, intercept.bottom = FALSE,
          out.header=FALSE, column.sep.width="1pt")
```

```
##
## ================================================================================================
##                                              Alternative models
##                        ------------------------------------------------------------------------
##                          Logistic                 Boruta                       GLS-Omit
##                            (1)                     (2)                          (3)
## ------------------------------------------------------------------------------------------------
## Constant                -0.83066**              0.59939***                   0.73928***
##                          (0.34954)               (0.06624)                    (0.06401)
##
## arm                     -0.14341***             -0.02542**                    -0.01773*
##                          (0.04835)               (0.01006)                    (0.01007)
##
## refinance               -0.24042***             -0.06033***                  -0.07181***
##                          (0.04761)               (0.00961)                    (0.00923)
##
## lien2                    1.27760***
##                          (0.15802)
##
## term30                    -0.06988               -0.03502**                   -0.03347***
##                          (0.07249)               (0.01407)                    (0.01266)
##
```

```
## underwater           1.24188***          0.15789***                  0.17702***
##                      (0.12132)           (0.01919)                   (0.01634)
##
## ltv                  0.02549***          0.00348***                  0.00203***
##                      (0.00219)           (0.00036)                   (0.00031)
##
## rate                 0.19066***          0.04318***                  0.04013***
##                      (0.01281)           (0.00255)                   (0.00248)
##
## amount                -0.00132           -0.00041*                   -0.00051***
##                      (0.00130)           (0.00023)                   (0.00016)
##
## fico                -0.00523***         -0.00113***                 -0.00117***
##                      (0.00037)           (0.00008)                   (0.00007)
##
## ------------------------------------------------------------------------------------
## Observations           10,000              10,000                      9,816
## R2                                         0.11100                    0.12687
## Adjusted R2                                0.11029                    0.12616
## Log Likelihood      -5,956.21100
## Akaike Inf. Crit.   11,932.42000
## Residual Std. Error                 0.45540 (df = 9991)         1.02566 (df = 9807)
## F Statistic                      155.94000*** (df = 8; 9991) 178.12710*** (df = 8; 9807)
## ====================================================================================
## Note:                                                      *p<0.1; **p<0.05; ***p<0.01
```

Based on the Breusch-Pagan Test, heteroskedasticity is present as expected. After removing the variables that are not statistically significant, we are confident that our estimates for the coefficients are unbiased. By applying Heteroskedasticity-Consistent standard errors and feasible generalized least squares, we have accounted for heteroskedasticity. And as $n \to \infty$ our standard errors, t-tests, interval estimates are valid in large samples.

Question 9

Estimate a model based on all your findings that also includes interaction terms (if appropriate) and if needed, any higher power terms. Comment on the performance of this model compared to your other models. Make sure to use AIC and BIC for model comparison.

```
lm.interaction <- lm(default~refinance+term30+underwater+ltv+rate*arm+amount+
                     fico, #re-estimating the model
                data = vegas5)

summary(lm.interaction) #seems a much better model compared to others
```

```
##
## Call:
## lm(formula = default ~ refinance + term30 + underwater + ltv +
##     rate * arm + amount + fico, data = vegas5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8252 -0.3717 -0.2039  0.4858  1.0728
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.546e-01  6.787e-02   9.645  < 2e-16 ***
## refinance   -6.065e-02  9.601e-03  -6.318 2.77e-10 ***
## term30      -3.932e-02  1.411e-02  -2.788 0.005318 **
## underwater   1.554e-01  1.919e-02   8.094 6.42e-16 ***
## ltv          3.297e-03  3.605e-04   9.145  < 2e-16 ***
## rate         3.553e-02  3.286e-03  10.815  < 2e-16 ***
## arm         -1.282e-01  2.961e-02  -4.330 1.51e-05 ***
## amount      -4.369e-04  2.341e-04  -1.867 0.061998 .
## fico        -1.116e-03  7.556e-05 -14.766  < 2e-16 ***
## rate:arm     1.838e-02  4.980e-03   3.691 0.000225 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4551 on 9990 degrees of freedom
## Multiple R-squared:  0.1122, Adjusted R-squared:  0.1114
## F-statistic: 140.3 on 9 and 9990 DF,  p-value: < 2.2e-16
```

```
b1 = BIC(model.lm)
b2 = BIC(model.glm)
b3 = BIC(updated.lm)
b4 = BIC(updated.lm.fgls)
b5 = BIC(updated.lm.fgls.omit)
b6 = BIC(lm.interaction)

c1 = AIC(model.lm)
c2 = AIC(model.glm)
c3 = AIC(updated.lm)
c4 = AIC(updated.lm.fgls)
c5 = AIC(updated.lm.fgls.omit)
c6 = AIC(lm.interaction)

v1 = c(b1,b2,b3,b4,b5,b6)
v2 = c(c1,c2,c3,c4,c5,c6)
df = cbind(v1,v2)

colnames(df) = c("BIC", "AIC")
rownames(df) = c("OLS", "Logistic",
                           "Boruta",
                           "GLS",
                           "GLS-Omit",
                           "Interaction")
kable(df)
```

	BIC	AIC
OLS	12691.07	12611.75
Logistic	12004.53	11932.42
Boruta	12730.26	12658.15
GLS	13207.13	13135.02
GLS-Omit	12786.55	12714.63
Interaction	12725.84	12646.53

We have finally attained a better model where include an interaction term as well. It seems as though there

is no need for any higher power terms, the model has been constructed with all terms being statistically significant, by adding higher power terms, we lose that significance.

Question 10

Evaluate your model performance (from 9) using cross-validation, and also by dividing your data into the traditional 2/3 training and 1/3 testing samples, to evaluate your out-of-sample performance. Comment on your results.

```
set.seed(123)
row.number <- sample(1:nrow(vegas5), 0.75*nrow(vegas5))
training_set <- vegas5[row.number,]
test_set <- vegas5[-row.number,]
dim(training_set)
```

```
## [1] 7500   10
```

```
#calculating RMSE
sqrt(mean((training_set$default-predict(lm.interaction,training_set))^2))
```

```
## [1] 0.4569231
```

```
#the predicted value is off by 0.45747

sqrt(mean((test_set$default-predict(lm.interaction,test_set))^2))
```

```
## [1] 0.4487122
```

```
#on average the predicted value is off by 0.45011

fit <- lm(default~refinance+term30+underwater+ltv+rate*arm+amount+
                    fico,
              data = vegas5,
          x = TRUE,
          y = TRUE)

cv.lm(fit, k = 10)
```

```
## Mean absolute error        :  0.415739
## Sample standard deviation  :  0.004543515
##
## Mean squared error         :  0.2072976
## Sample standard deviation  :  0.005419535
##
## Root mean squared error    :  0.4552644
## Sample standard deviation  :  0.005955342
```

```
#RMSE 0.4550
```

Question 11

Provide a short (1 paragraph) summary of your overall conclusions/findings.

Based on simple linear regression models we created, there are several variables which initially seem relevant to the default rate turned out to be statistically insignificant. They include the adjustable rate mortgage (arm), 30-year vs. 15-year mortgage, amount. Through removing outlier data points and transformation we eventually created models with higher order. We also omitted terms that seem to be irrelevant to the default rate, and finally we were able to include interaction terms. One interesting finding is that when we include interactions, term30 and arm became statistically significant, therefore for completeness, we decided to retain term30 and arm as predictor variables. And our conclusions is the following:

Lien2 and amount are not statistically significant as predictor variables;

Refinancing loans have a slightly lower chance of default;

30-year mortgages have slightly lower change of default;

Underwater mortgages have a higher default rate;

Adjustable rate mortgages have higher default rate;

LTV is positively correlated with default rate, specifically 1% increase in LTV ratio is correlated to 0.0033% increase in default rate; The mortgage rate is positively correlated with default rate, 1% increase in the mortgage rate can lead to 0.0355% increase in default rate;

Although the amount of the loan does not seem to have significant correlation with the default rate, we did find a positive correlation between default rate and the interaction between amount and mortgage rate. 1% increase in the product of amount and rate is expected to increase 0.0184% increase in default rate;

Finally, there is also an expected negative correlation between FICO scores and the default rate, for an extra point increase in FICO is translated to 0.00112% decrease in default rate.