NAFLD Research

Raghava Govil

4/13/2019

```
#Reading in the file
library(readxl)
data <- read_excel("Mexican Bariatric Surgery Database_april_2019.xlsx")
```

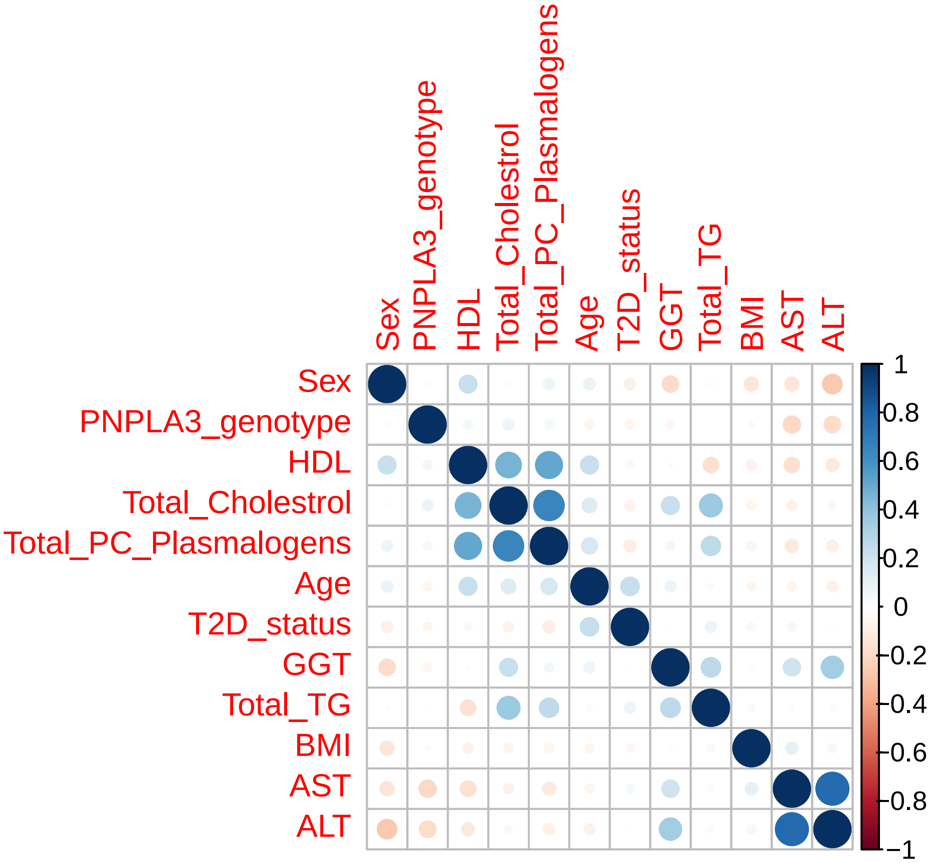
Classification based on Diagnostic Status

```
final <- data[,c(2, 3, 4, 19, 20, 21, 13, 15, 62, 74, 42, 43, 37)]#Subsetting the required variables
final <- final[-c(1, 2), ]#Removing first two rows
```

```
library(dplyr)
final <- final[complete.cases(final), ]#Removing NAs
colnames(final)[2] <- "Sex"
colnames(final)[7] <- "Total_Cholestrol"
colnames(final)[12] <- "PNPLA3_genotype"
final[, 1:12] <- final[, 1:12] %>% mutate_if(is.character, as.numeric)#Converting character variables t
final[, 4:10] <- lapply(final[, 4:10], log)#Taking the log of the required variables
final[final$Status_diagnostic == "Uncertain", 13] <- "NASH Bordeline"
final[final$Status_diagnostic == "Control", 13] <- "Steatosis"
final <- final[complete.cases(final), ]
#colSums(is.na(final))
```

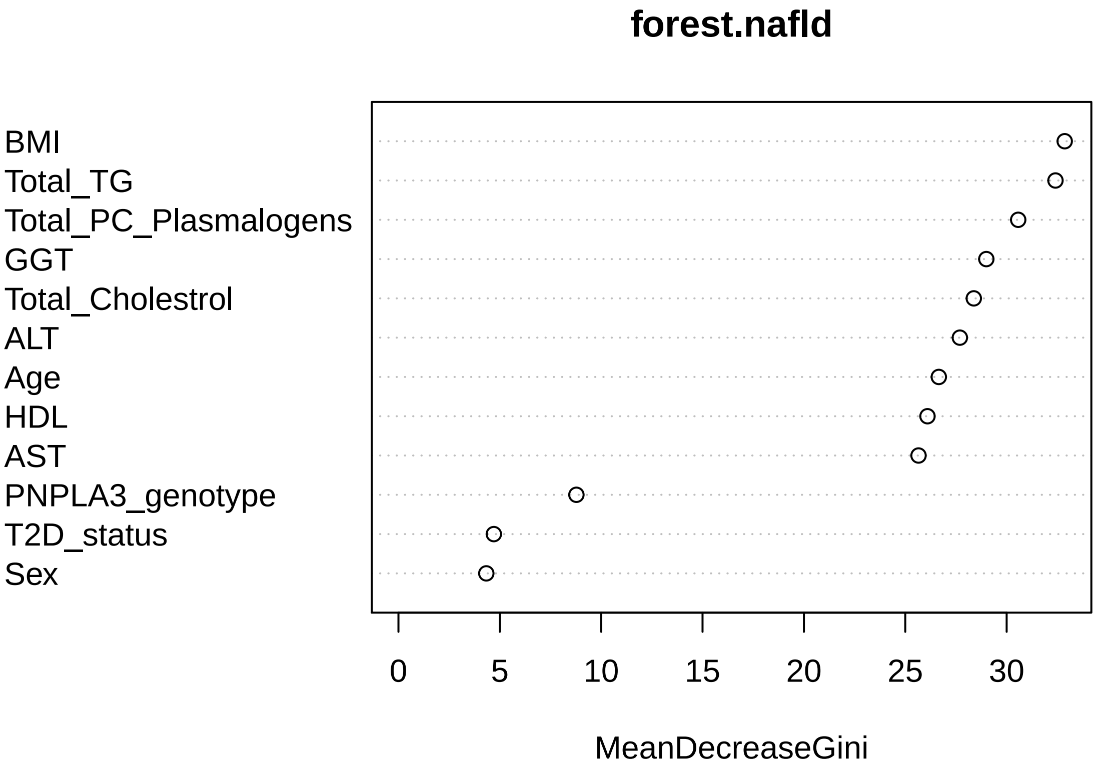
Correlation PLot

```
library(corrplot)
corrplot(cor(final[,1:12]), order = "hclust")
```



Random Forest Variable Importance Plot

```
library(randomForest)
forest.nafld <- randomForest(factor(Status_diagnostic) ~., data=final, type="classification")
varImpPlot(forest.nafld, type = 2)
```



Decision Tree Variable Importance Values

```
library(tree)
library(caret)
library(rpart)
tree.nafld = rpart(factor(Status_diagnostic) ~., data=final, method = "class")
tree.nafld$variable.importance
```

```
##                   ALT                   BMI                   GGT
##            14.2490539            11.8539714             8.9440747
## Total_PC_Plasmalogens                   Age              Total_TG
##             8.9194349             7.6288851             7.5527047
##      Total_Cholestrol                   AST                   HDL
##             7.5475523             7.1649942             2.6544081
##                   Sex
##             0.3381884
```

Linear Model for Steatosis Stage

```
final_steatosis <- data[,c(2, 3, 4, 19, 20, 21, 13, 15, 62, 74, 42, 43, 45:91, 38)]
final_steatosis <- subset(final_steatosis, select = -c(44, 51))
final_steatosis <- final_steatosis[-c(1, 2), ]
final_steatosis <- final_steatosis[, c(1:43, 58)]
```

```
#Data Cleaning
colnames(final_steatosis)[2] <- "Sex"
colnames(final_steatosis)[7] <- "Total_Cholestrol"
colnames(final_steatosis)[9] <- "Total_PC_Plasmalogens1"
```

```r
colnames(final_steatosis)[10] <- "Total_TG1"
colnames(final_steatosis)[12] <- "PNPLA3_genotype"
final_steatosis <- final_steatosis %>% mutate_if(is.character, as.numeric)
final_steatosis[, c(4:10, 13:43)] <- lapply(final_steatosis[, c(4:10, 13:43)], log)
#final1[is.na(final1)] <- 0
final_steatosis <- final_steatosis[complete.cases(final_steatosis), ]
#colSums(is.na(final_steatosis))
```

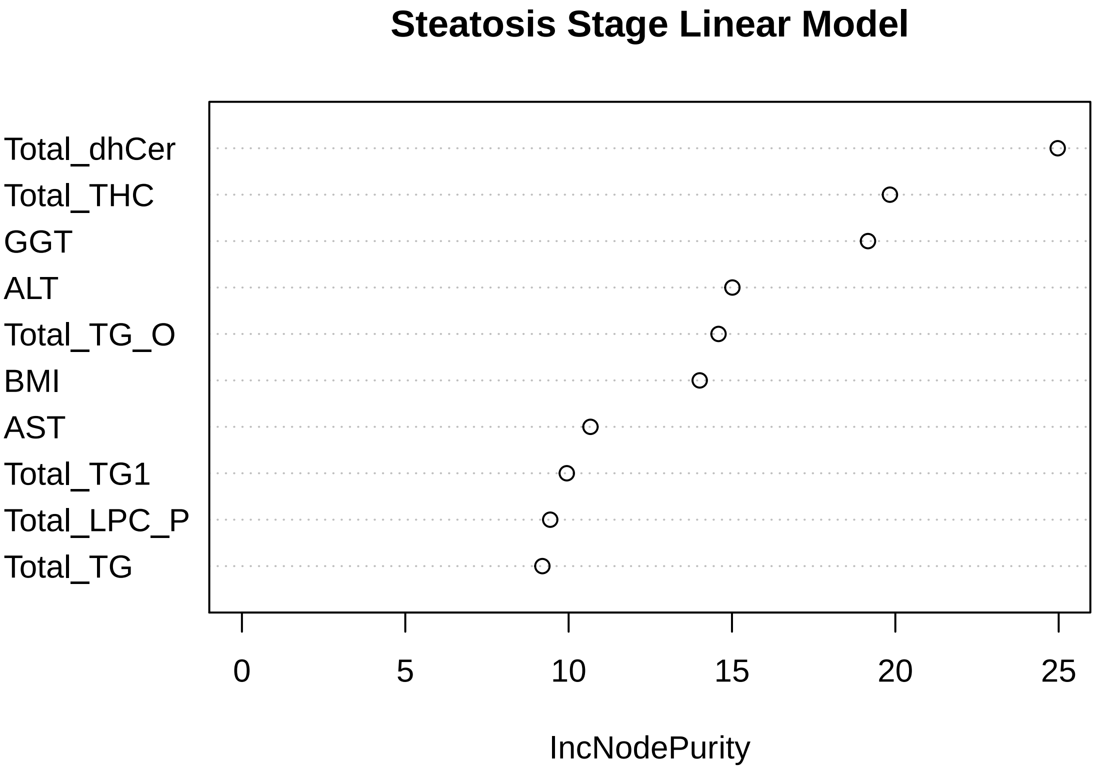
Linear Model (r-squared = 42%) - Asterisks near variable row indicate significance

```r
m1 <- lm(final_steatosis$Steatosis_stage~., data = final_steatosis)
summary(m1)
```

```
##
## Call:
## lm(formula = final_steatosis$Steatosis_stage ~ ., data = final_steatosis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.12901 -0.58644  0.00618  0.60805  1.80345
##
## Coefficients: (2 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -8.0955039 11.0211149  -0.735 0.463216
## Age                     0.0008428  0.0059808   0.141 0.888038
## Sex                    -0.3611075  0.1505957  -2.398 0.017130 *
## BMI                     0.0186927  0.0084519   2.212 0.027779 *
## AST                     0.0022339  0.1928906   0.012 0.990768
## ALT                     0.1950659  0.1709504   1.141 0.254793
## GGT                     0.1398195  0.1138016   1.229 0.220219
## Total_Cholestrol       -0.6456565  0.4177260  -1.546 0.123292
## HDL                     0.3423479  0.2823204   1.213 0.226272
## Total_PC_Plasmalogens1  0.5942670  0.8391395   0.708 0.479404
## Total_TG1               0.5765486  0.7125761   0.809 0.419125
## T2D_status             -0.0457039  0.1284321  -0.356 0.722206
## PNPLA3_genotype        -0.3421729  0.0737274  -4.641 5.28e-06 ***
## Total_acylcarnitine    -0.0519621  0.2069565  -0.251 0.801934
## Total_CE                0.6044555  0.6948257   0.870 0.385061
## Total_CE_other          0.2669584  0.2425863   1.100 0.272051
## Total_COH              -0.4868079  0.3994075  -1.219 0.223911
## Total_Cer              -0.3672050  0.3204522  -1.146 0.252793
## Total_DG               -0.3482021  0.3466977  -1.004 0.316061
## Total_DHC              -0.4730133  0.3552673  -1.331 0.184104
## Total_GM3              -0.3166637  0.3535906  -0.896 0.371235
## Total_LPC              -1.4428321  0.8135318  -1.774 0.077199 .
## Total_LPC_O             1.9428955  0.7124424   2.727 0.006783 **
## Total_LPC_P            -0.8458776  0.7401303  -1.143 0.254042
## Total_LPE               0.7964977  0.4682853   1.701 0.090048 .
## Total_LPE_P            -0.7866983  0.3953944  -1.990 0.047578 *
## Total_LPI               0.7179629  0.3554037   2.020 0.044298 *
## Total_MHC               0.2397650  0.2649940   0.905 0.366333
## Total_PC                0.8492824  1.0338852   0.821 0.412073
## Total_PC_O             -2.0376666  0.8658965  -2.353 0.019285 *
## Total_PC_Plasmalogens           NA         NA      NA       NA
## Total_PE                -0.4582462  0.3255745  -1.408 0.160361
## Total_PE_O               0.3761661  0.2771840   1.357 0.175816
## Total_PE_P               0.2027409  0.4181924   0.485 0.628186
## Total_PG                 0.0943421  0.2882692   0.327 0.743702
## Total_PI                 0.0053435  0.3729430   0.014 0.988578
## Total_PS                 0.1201544  0.1023585   1.174 0.241425
## Total_SM                 0.4710554  0.6511105   0.723 0.469984
## Total_sulfatide          0.3215628  0.5495892   0.585 0.558943
## Total_THC               -0.6027681  0.3855804  -1.563 0.119090
## Total_ubiquinone         0.3301698  0.2522832   1.309 0.191674
## Total_dhCer              0.5585873  0.1542615   3.621 0.000347 ***
## Total_TG                        NA         NA      NA       NA
## Total_TG_O               0.5607157  0.2507819   2.236 0.026129 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8408 on 287 degrees of freedom
## Multiple R-squared:  0.4212, Adjusted R-squared:  0.3385
## F-statistic: 5.093 on 41 and 287 DF,  p-value: < 2.2e-16
```

Random Forest Regression (Top 10)

```
library(randomForest)
m2 <- randomForest(final_steatosis$Steatosis_stage~., data = final_steatosis, type="regression")
varImpPlot(m2, type = 2, n.var = 10, main = "Steatosis Stage Linear Model")
```



Linear Model for Inflammation Stage

```
final_inflammation <- data[,c(2, 3, 4, 19, 20, 21, 13, 15, 62, 74, 42, 43, 45:91, 39)]
final_inflammation <- subset(final_inflammation, select = -c(44, 51))
final_inflammation <- final_inflammation[-c(1, 2), ]
final_inflammation <- final_inflammation[, c(1:43, 58)]
```

```
#Data Cleaning
colnames(final_inflammation)[2] <- "Sex"
colnames(final_inflammation)[7] <- "Total_Cholestrol"
colnames(final_inflammation)[9] <- "Total_PC_Plasmalogens1"
colnames(final_inflammation)[10] <- "Total_TG1"
colnames(final_inflammation)[12] <- "PNPLA3_genotype"
final_inflammation <- final_inflammation %>% mutate_if(is.character, as.numeric)
final_inflammation[, c(4:10, 13:43)] <- lapply(final_inflammation[, c(4:10, 13:43)], log)
#final1[is.na(final1)] <- 0
final_inflammation <- final_inflammation[complete.cases(final_inflammation), ]
#colSums(is.na(final_inflammation))
```

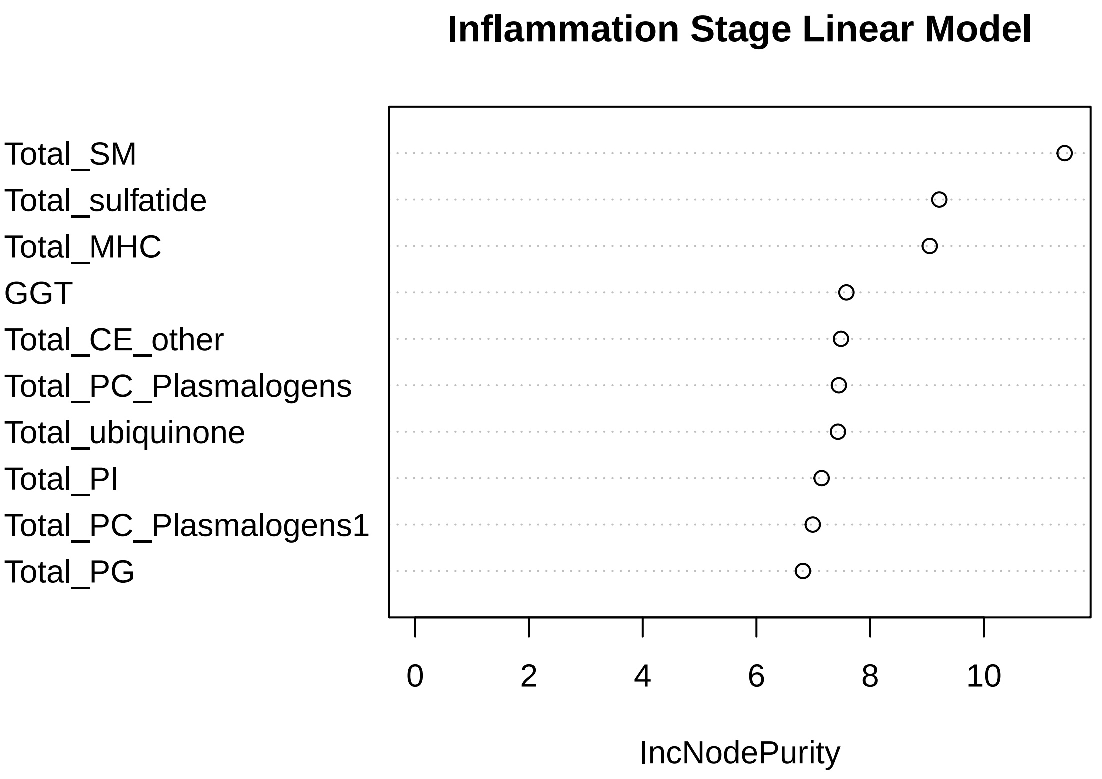
Linear Model (r-squared = 17%) - Asterisks near variable row indicate significance

```
m3 <- lm(final_inflammation$Inflammation_stage~., data = final_inflammation)
summary(m3)
```

```
##
## Call:
## lm(formula = final_inflammation$Inflammation_stage ~ ., data = final_inflammation)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6286 -0.6134 -0.1566  0.5242  2.3642
##
## Coefficients: (2 not defined because of singularities)
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            13.771856  11.109621   1.240   0.2161
## Age                     0.002849   0.006029   0.473   0.6369
## Sex                    -0.063018   0.151805  -0.415   0.6784
## BMI                     0.003159   0.008520   0.371   0.7111
## AST                     0.007371   0.194440   0.038   0.9698
## ALT                     0.065463   0.172323   0.380   0.7043
## GGT                     0.099226   0.114716   0.865   0.3878
## Total_Cholestrol       -0.345116   0.421081  -0.820   0.4131
## HDL                     0.722479   0.284588   2.539   0.0117 *
## Total_PC_Plasmalogens1  0.529149   0.845878   0.626   0.5321
## Total_TG1               0.181195   0.718298   0.252   0.8010
## T2D_status              0.083544   0.129463   0.645   0.5192
## PNPLA3_genotype        -0.016158   0.074319  -0.217   0.8280
## Total_acylcarnitine    -0.171471   0.208618  -0.822   0.4118
## Total_CE               -0.366082   0.700406  -0.523   0.6016
## Total_CE_other          0.479628   0.244534   1.961   0.0508 .
## Total_COH               0.364589   0.402615   0.906   0.3659
## Total_Cer              -0.589701   0.323026  -1.826   0.0690 .
## Total_DG                0.194770   0.349482   0.557   0.5777
## Total_DHC               0.400109   0.358120   1.117   0.2648
## Total_GM3               0.060693   0.356430   0.170   0.8649
## Total_LPC              -0.147405   0.820065  -0.180   0.8575
## Total_LPC_O             0.490302   0.718164   0.683   0.4953
## Total_LPC_P            -0.622940   0.746074  -0.835   0.4044
## Total_LPE               0.204669   0.472046   0.434   0.6649
## Total_LPE_P            -0.004544   0.398570  -0.011   0.9909
## Total_LPI               0.340942   0.358258   0.952   0.3421
## Total_MHC              -0.366966   0.267122  -1.374   0.1706
## Total_PC                0.181305   1.042188   0.174   0.8620
## Total_PC_O             -0.840323   0.872850  -0.963   0.3365
## Total_PC_Plasmalogens         NA         NA      NA       NA
## Total_PE               -0.320604   0.328189  -0.977   0.3294
## Total_PE_O              0.194828   0.279410   0.697   0.4862
## Total_PE_P             -0.252912   0.421551  -0.600   0.5490
## Total_PG               -0.044085   0.290584  -0.152   0.8795
## Total_PI                0.262704   0.375938   0.699   0.4852
## Total_PS                0.052794   0.103181   0.512   0.6093
## Total_SM               -1.376956   0.656339  -2.098   0.0368 *
## Total_sulfatide        -0.398750   0.554003  -0.720   0.4723
## Total_THC               0.168079   0.388677   0.432   0.6657
## Total_ubiquinone        0.107938   0.254309   0.424   0.6716
## Total_dhCer             0.068855   0.155500   0.443   0.6582
## Total_TG                      NA         NA      NA       NA
## Total_TG_O              0.286851   0.252796   1.135   0.2574
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8475 on 287 degrees of freedom
## Multiple R-squared:  0.1792, Adjusted R-squared:  0.06189
## F-statistic: 1.528 on 41 and 287 DF,  p-value: 0.02581
```

Random Forest Regression (Top 10)

```
library(randomForest)
m4 <- randomForest(final_inflammation$Inflammation_stage~., data = final_inflammation, type="regression
varImpPlot(m4, type = 2, n.var = 10, main = "Inflammation Stage Linear Model")
```



Linear Model for Ballooning Stage

```
final_ballooning <- data[,c(2, 3, 4, 19, 20, 21, 13, 15, 62, 74, 42, 43, 45:91, 40)]
final_ballooning <- subset(final_ballooning, select = -c(44, 51))
final_ballooning <- final_ballooning[-c(1, 2), ]
final_ballooning <- final_ballooning[, c(1:43, 58)]
```

```
#Data Cleaning
colnames(final_ballooning)[2] <- "Sex"
colnames(final_ballooning)[7] <- "Total_Cholestrol"
colnames(final_ballooning)[9] <- "Total_PC_Plasmalogens1"
colnames(final_ballooning)[10] <- "Total_TG1"
colnames(final_ballooning)[12] <- "PNPLA3_genotype"
final_ballooning <- final_ballooning %>% mutate_if(is.character, as.numeric)
final_ballooning[, c(4:10, 13:43)] <- lapply(final_ballooning[, c(4:10, 13:43)], log)
#final1[is.na(final1)] <- 0
final_ballooning <- final_ballooning[complete.cases(final_ballooning), ]
#colSums(is.na(final_ballooning))
```

Linear Model (r-squared = 32%) - Asterisks near variable row indicate significance

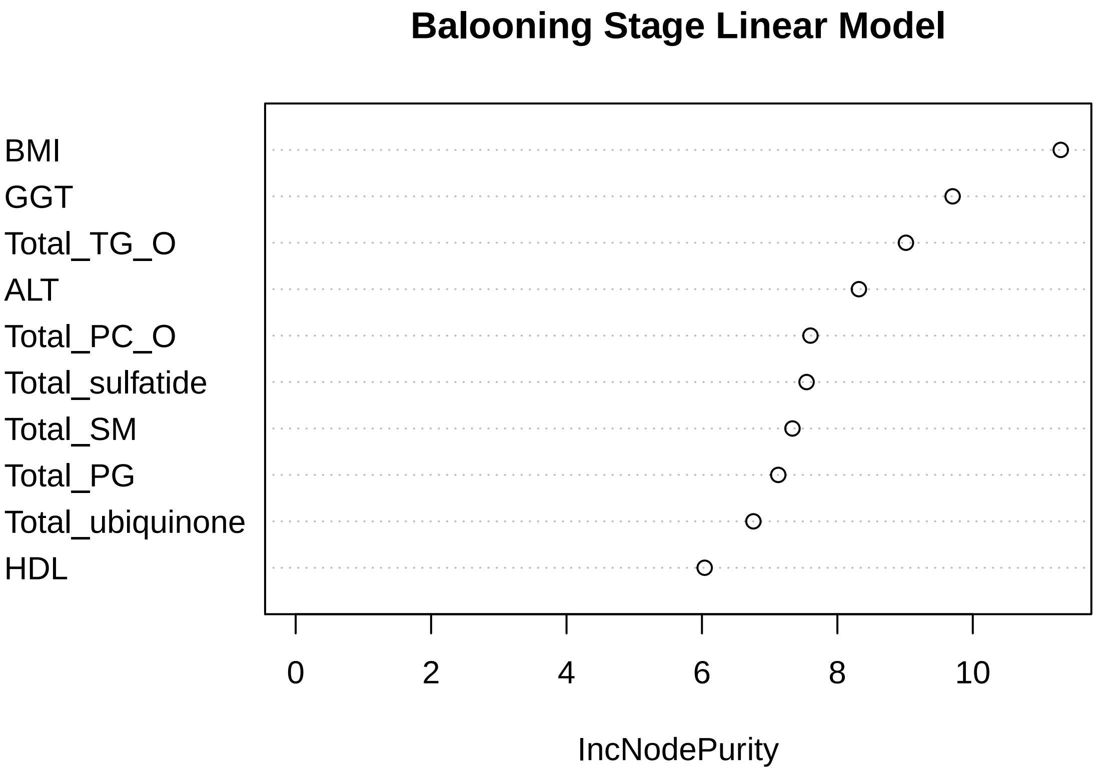
```
m5 <- lm(final_ballooning$Balloning_stage~., data = final_ballooning)
summary(m5)
```

```
##
## Call:
## lm(formula = final_ballooning$Balloning_stage ~ ., data = final_ballooning)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.55558 -0.48369 -0.08197  0.49963  1.71654
##
## Coefficients: (2 not defined because of singularities)
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            15.724714   9.233405   1.703 0.089646 .
## Age                     0.001107   0.005011   0.221 0.825266
## Sex                    -0.218376   0.126168  -1.731 0.084555 .
## BMI                     0.014786   0.007081   2.088 0.037665 *
## AST                    -0.167766   0.161602  -1.038 0.300078
## ALT                     0.276578   0.143221   1.931 0.054452 .
## GGT                     0.100040   0.095342   1.049 0.294935
## Total_Cholestrol       -0.384081   0.349968  -1.097 0.273353
## HDL                    -0.070939   0.236526  -0.300 0.764454
## Total_PC_Plasmalogens1  1.081264   0.703025   1.538 0.125146
## Total_TG1              -0.101067   0.596991  -0.169 0.865684
## T2D_status             -0.032230   0.107599  -0.300 0.764746
## PNPLA3_genotype        -0.118520   0.061768  -1.919 0.056003 .
## Total_acylcarnitine     0.076367   0.173387   0.440 0.659948
## Total_CE                0.071336   0.582120   0.123 0.902553
## Total_CE_other          0.199611   0.203237   0.982 0.326849
## Total_COH               0.045996   0.334621   0.137 0.890767
## Total_Cer               0.316161   0.268472   1.178 0.239921
## Total_DG               -0.605099   0.290461  -2.083 0.038114 *
## Total_DHC              -0.177582   0.297640  -0.597 0.551222
## Total_GM3               0.329988   0.296235   1.114 0.266237
## Total_LPC              -0.625924   0.681571  -0.918 0.359204
## Total_LPC_O             1.425503   0.596879   2.388 0.017575 *
## Total_LPC_P            -0.775885   0.620076  -1.251 0.211853
## Total_LPE               0.341642   0.392326   0.871 0.384585
## Total_LPE_P            -0.519410   0.331258  -1.568 0.117985
## Total_LPI               0.758880   0.297755   2.549 0.011334 *
## Total_MHC               0.130261   0.222010   0.587 0.557842
## Total_PC               -0.969237   0.866181  -1.119 0.264085
## Total_PC_O             -1.171810   0.725441  -1.615 0.107343
## Total_PC_Plasmalogens         NA         NA      NA       NA
## Total_PE               -0.037949   0.272764  -0.139 0.889445
## Total_PE_O              0.251109   0.232223   1.081 0.280459
## Total_PE_P             -0.099098   0.350358  -0.283 0.777498
## Total_PG                0.150365   0.241510   0.623 0.534038
## Total_PI                0.113322   0.312449   0.363 0.717104
## Total_PS               -0.010172   0.085755  -0.119 0.905661
## Total_SM               -0.909775   0.545495  -1.668 0.096447 .
## Total_sulfatide        -0.036143   0.460442  -0.078 0.937487
## Total_THC              -0.180465   0.323036  -0.559 0.576835
## Total_ubiquinone        0.231434   0.211361   1.095 0.274447
## Total_dhCer             0.165166   0.129239   1.278 0.202287
## Total_TG                      NA         NA      NA       NA
## Total_TG_O              0.805965   0.210103   3.836 0.000154 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7044 on 287 degrees of freedom
## Multiple R-squared:  0.3281, Adjusted R-squared:  0.2322
## F-statistic: 3.419 on 41 and 287 DF,  p-value: 6.582e-10
```

Random Forest Regression (Top 10)

```r
library(randomForest)
m6 <- randomForest(final_ballooning$Balloning_stage~., data = final_ballooning, type="regression")
varImpPlot(m6, type = 2, n.var = 10, main = "Balooning Stage Linear Model")
```



Linear Model for NAS Score

```r
final_NAS <- data[,c(2, 3, 4, 19, 20, 21, 13, 15, 62, 74, 42, 43, 45:91, 41)]
final_NAS <- subset(final_NAS, select = -c(44, 51))
final_NAS <- final_NAS[-c(1, 2), ]
final_NAS <- final_NAS[, c(1:43, 58)]
```

```r
#Data Cleaning
colnames(final_NAS)[2] <- "Sex"
colnames(final_NAS)[7] <- "Total_Cholestrol"
colnames(final_NAS)[9] <- "Total_PC_Plasmalogens1"
colnames(final_NAS)[10] <- "Total_TG1"
colnames(final_NAS)[12] <- "PNPLA3_genotype"
final_NAS <- final_NAS %>% mutate_if(is.character, as.numeric)
final_NAS[, c(4:10, 13:43)] <- lapply(final_NAS[, c(4:10, 13:43)], log)
#final1[is.na(final1)] <- 0
final_NAS <- final_NAS[complete.cases(final_NAS), ]
#colSums(is.na(final_NAS))
```

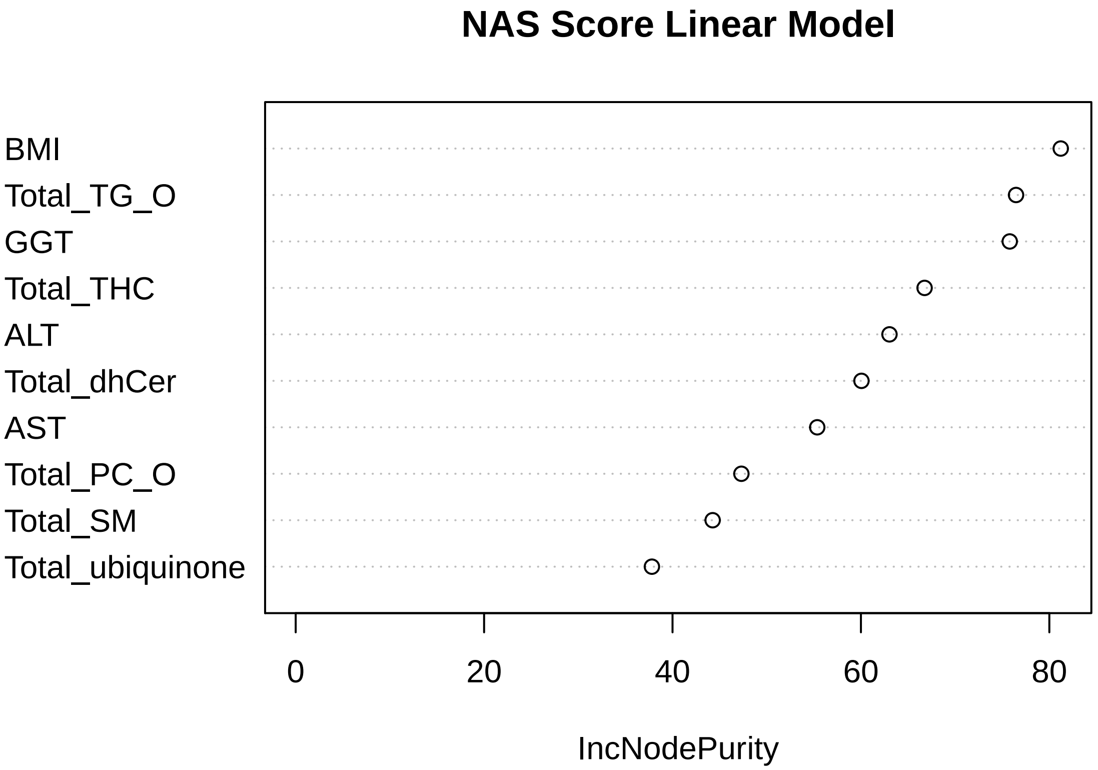
Linear Model (r-squared = 37%) - Asterisks near variable row indicate significance

```r
m7 <- lm(final_NAS$`NAS score`~., data = final_NAS)
summary(m7)
```

```
##
## Call:
## lm(formula = final_NAS$`NAS score` ~ ., data = final_NAS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9695 -1.2692 -0.1147  1.1407  4.2106
##
## Coefficients: (2 not defined because of singularities)
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            21.401066  23.577192   0.908  0.36480
## Age                     0.004799   0.012795   0.375  0.70786
## Sex                    -0.642501   0.322166  -1.994  0.04706 *
## BMI                     0.036638   0.018081   2.026  0.04366 *
## AST                    -0.158162   0.412646  -0.383  0.70179
## ALT                     0.537107   0.365710   1.469  0.14302
## GGT                     0.339086   0.243453   1.393  0.16475
## Total_Cholestrol       -1.374854   0.893631  -1.539  0.12503
## HDL                     0.993888   0.603961   1.646  0.10094
## Total_PC_Plasmalogens1  2.204680   1.795150   1.228  0.22040
## Total_TG1               0.656676   1.524396   0.431  0.66695
## T2D_status              0.005610   0.274752   0.020  0.98372
## PNPLA3_genotype        -0.476851   0.157723  -3.023  0.00273 **
## Total_acylcarnitine    -0.147067   0.442737  -0.332  0.74000
## Total_CE                0.309709   1.486423   0.208  0.83510
## Total_CE_other          0.946197   0.518959   1.823  0.06930 .
## Total_COH              -0.076223   0.854442  -0.089  0.92898
## Total_Cer              -0.640746   0.685535  -0.935  0.35075
## Total_DG               -0.758531   0.741682  -1.023  0.30730
## Total_DHC              -0.250487   0.760014  -0.330  0.74196
## Total_GM3               0.074018   0.756427   0.098  0.92212
## Total_LPC              -2.216161   1.740368  -1.273  0.20391
## Total_LPC_O             3.858700   1.524110   2.532  0.01188 *
## Total_LPC_P            -2.244702   1.583342  -1.418  0.15736
## Total_LPE               1.342809   1.001791   1.340  0.18117
## Total_LPE_P            -1.310653   0.845857  -1.549  0.12236
## Total_LPI               1.817785   0.760306   2.391  0.01745 *
## Total_MHC               0.003061   0.566895   0.005  0.99570
## Total_PC                0.061350   2.211765   0.028  0.97789
## Total_PC_O             -4.049800   1.852390  -2.186  0.02960 *
## Total_PC_Plasmalogens         NA         NA      NA       NA
## Total_PE               -0.816799   0.696493  -1.173  0.24188
## Total_PE_O              0.822103   0.592973   1.386  0.16670
## Total_PE_P             -0.149269   0.894628  -0.167  0.86761
## Total_PG                0.200623   0.616687   0.325  0.74517
## Total_PI                0.381370   0.797827   0.478  0.63301
## Total_PS                0.162776   0.218973   0.743  0.45787
## Total_SM               -1.815676   1.392904  -1.304  0.19344
## Total_sulfatide        -0.113331   1.175722  -0.096  0.92328
## Total_THC              -0.615154   0.824862  -0.746  0.45642
## Total_ubiquinone        0.669543   0.539703   1.241  0.21578   
## Total_dhCer             0.792608   0.330008   2.402  0.01695 * 
## Total_TG                      NA         NA      NA       NA   
## Total_TG_O              1.653532   0.536491   3.082  0.00226 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.799 on 287 degrees of freedom
## Multiple R-squared:  0.3761, Adjusted R-squared:  0.2869 
## F-statistic: 4.219 on 41 and 287 DF,  p-value: 2.154e-13
```

Random Forest Regression (Top 10)

```
library(randomForest)
m8 <- randomForest(final_NAS$`NAS score`~., data = final_NAS, type="regression")
varImpPlot(m8, type = 2, n.var = 10, main = "NAS Score Linear Model")
```


NAS Score Linear Model
BMI
Total_TG_O
GGT
Total_THC
ALT
Total_dhCer
AST
Total_PC_O
Total_SM
Total_ubiquinone
IncNodePurity