Read the white wine data. What covariates are most strongly associated with wine quality?

```
dat <- read.csv('winequality-white.csv', sep=';')


inds <- sample(1:nrow(dat), round(0.2*nrow(dat),0))
train <- dat[-inds,]
test  <- dat[inds,]

write.csv(train, file='whiteWineTrain.csv')
write.csv(test, file='whiteWineTest.csv')
```

red wine

```
dat2 <- read.csv('winequality-red.csv', sep=';')
inds2 <- sample(1:nrow(dat2), round(0.2*nrow(dat2),0))
train2 <- dat2[-inds2,]
test2  <- dat2[inds2,]
write.csv(train2, file='redWineTrain.csv')
write.csv(test2, file='redWineTest.csv')
```

```
mod <- lm(quality ~ . ,data=dat)
summary(mod)
```

```
##
## Call:
## lm(formula = quality ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8348 -0.4934 -0.0379  0.4637  3.1143
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.502e+02  1.880e+01   7.987 1.71e-15 ***
## fixed.acidity         6.552e-02  2.087e-02   3.139  0.00171 **
## volatile.acidity     -1.863e+00  1.138e-01 -16.373  < 2e-16 ***
## citric.acid           2.209e-02  9.577e-02   0.231  0.81759
## residual.sugar        8.148e-02  7.527e-03  10.825  < 2e-16 ***
## chlorides            -2.473e-01  5.465e-01  -0.452  0.65097
## free.sulfur.dioxide   3.733e-03  8.441e-04   4.422 9.99e-06 ***
## total.sulfur.dioxide -2.857e-04  3.781e-04  -0.756  0.44979
## density              -1.503e+02  1.907e+01  -7.879 4.04e-15 ***
## pH                    6.863e-01  1.054e-01   6.513 8.10e-11 ***
## sulphates             6.315e-01  1.004e-01   6.291 3.44e-10 ***
## alcohol               1.935e-01  2.422e-02   7.988 1.70e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7514 on 4886 degrees of freedom
## Multiple R-squared:  0.2819, Adjusted R-squared:  0.2803
## F-statistic: 174.3 on 11 and 4886 DF,  p-value: < 2.2e-16
```

Now let's focus on prediction. The point isn't to find what variables are associated, but to predict quality. Go HAM.

```
inds <- sample(1:nrow(dat), round(0.2*nrow(dat),0))
train <- dat[-inds,]
test  <- dat[inds,]
```

```
library(gbm)
```

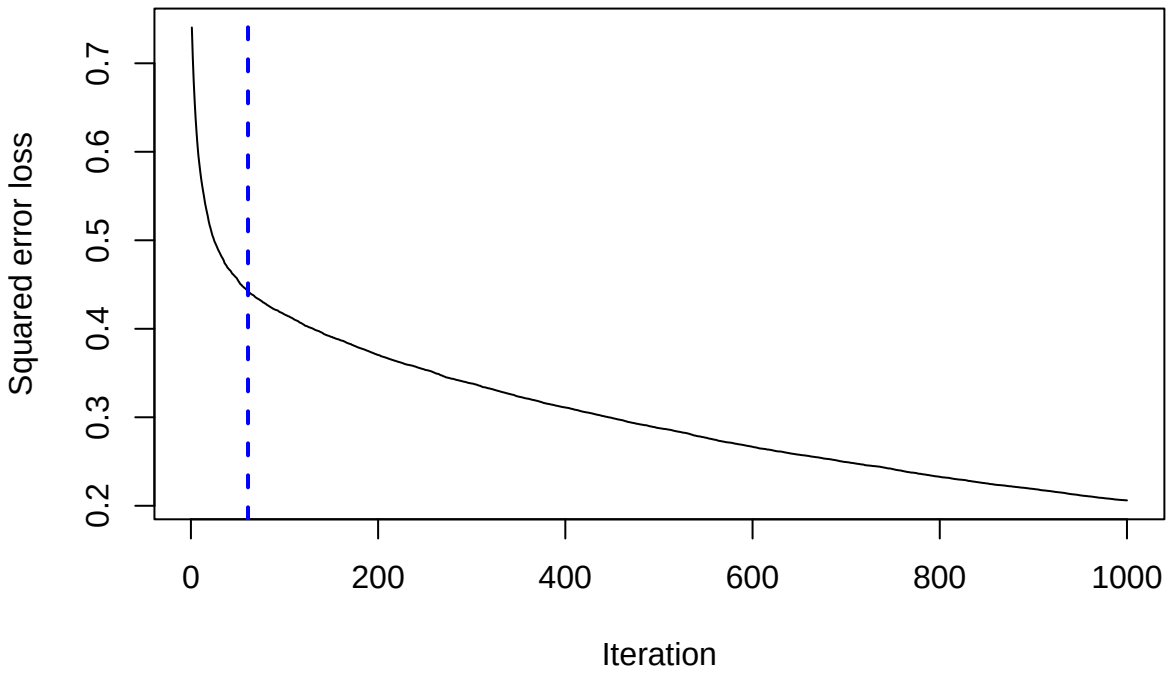
```
## Loaded gbm 2.1.8
```

```
mod <- gbm::gbm(quality ~. , data=train, n.trees=1000, interaction.depth=4, cv.folds=4)
```

```
## Distribution not specified, assuming gaussian ...
```

```
preds <- predict(mod, newdata=test, type='response', n.trees=gbm.perf(mod,method='OOB'))
```

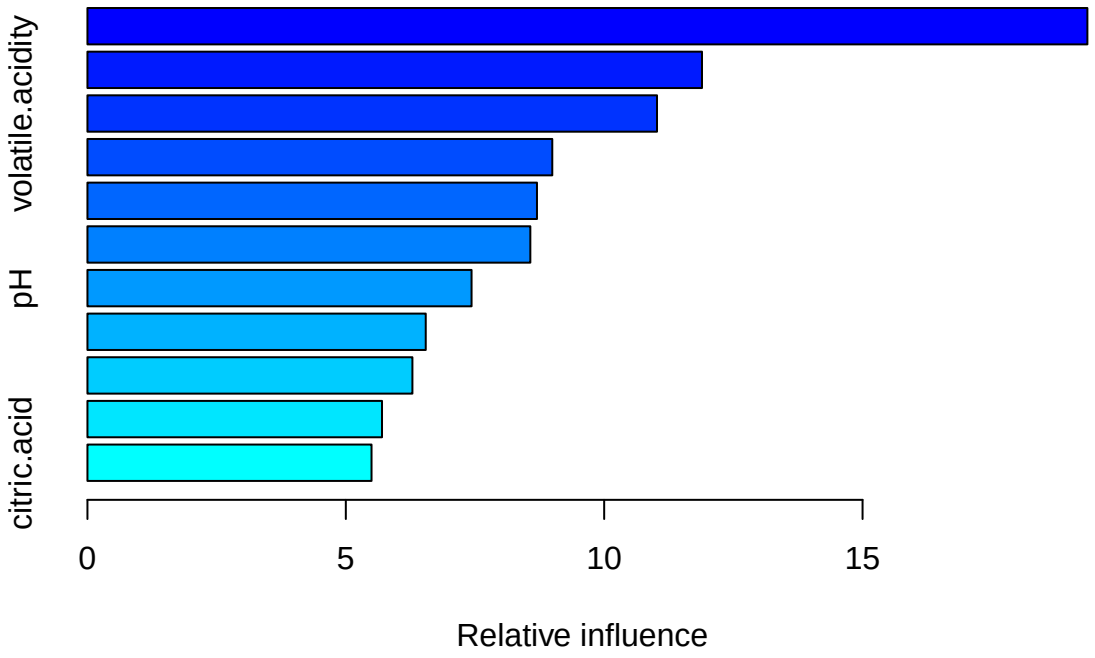
```
## OOB generally underestimates the optimal number of iterations although predictive performance is reas
```




```
sqrt(mean((preds-test$quality)**2))
```

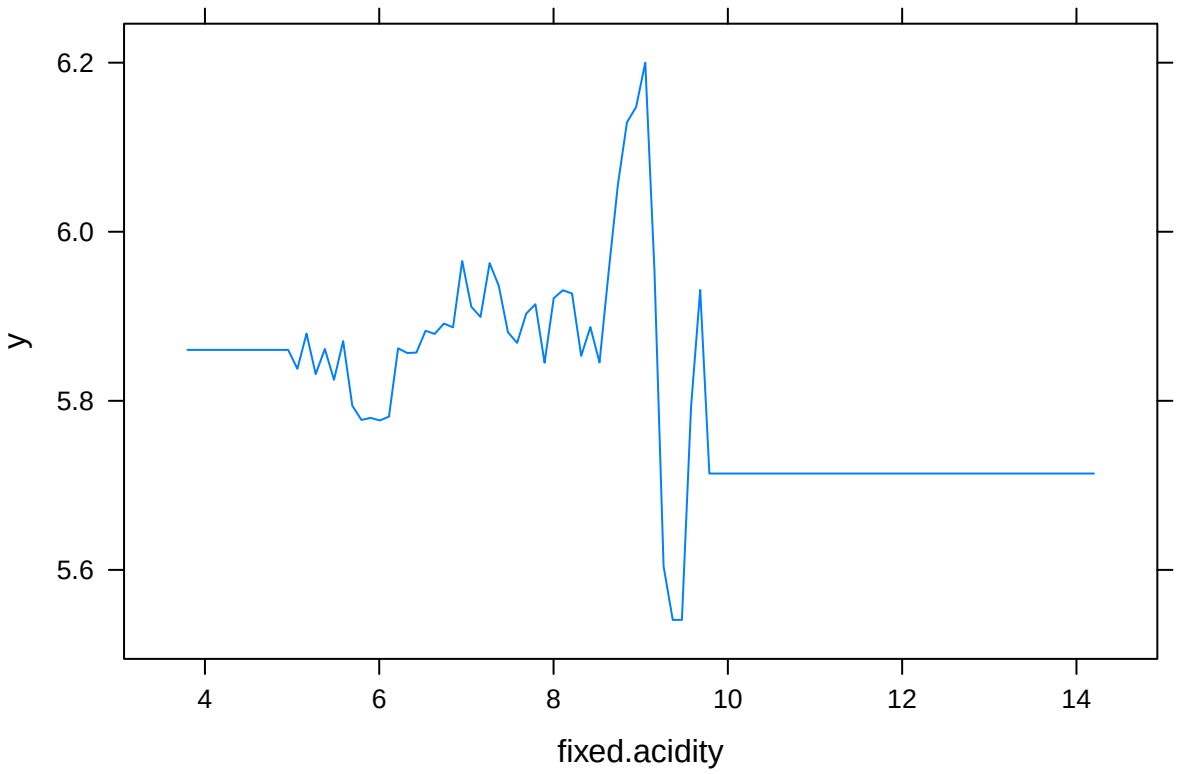
```
## [1] 0.7239701
```

```
summary(mod)
```




```
##                                     var   rel.inf
## alcohol                         alcohol 19.351172
## free.sulfur.dioxide   free.sulfur.dioxide 11.892857
## volatile.acidity         volatile.acidity 11.022296
## residual.sugar             residual.sugar  8.995838
## density                           density  8.699360
## total.sulfur.dioxide total.sulfur.dioxide  8.571099
## pH                                     pH  7.433500
## fixed.acidity               fixed.acidity  6.546916
## sulphates                       sulphates  6.289027
## chlorides                       chlorides  5.700389
## citric.acid                   citric.acid  5.497546
```

```
plot(mod)
```




```
rmse <- function(preds, actual){
    sqrt(mean((preds-actual)**2))
}
```

How uncertain are you in your predictions?

How consistent are your variable importance estimates?