Logistic Regression, Resampling Methods

Ryan Zhang

October 29, 2015

What's Covered Last Time

- Best Subset Method
  - Consider all possible models
  - 2^p
  - Impossible to do when p is large
  - $2^{40} = 10^{11}$
- Stepwise Method
  - Reduce the searching space for models
  - p for first round, $p - 1$ for second round, etc,
  - $0.5p^2$
  - $0.5 \times 40^2 = 800$

What's Covered Last Time

- Model Selection Criteria
  - For interpretation purpose:
    - test results, t, F, etc.
  - In sample error(How well model fit the data we have)
    - R^2, Standard Error of Estimate, etc.
  - Out of sample error(How well model will like to perform on unseen data)
    - Adjust in sample error estimates: $adj.R^2, C_p, AIC, BIC$ etc.
    - Directly estimate: Validation
- Machine Learning folks always do validation or...

What's Covered Last Time

- Ridge and Lasso Regression
  - $\min_{\beta}(\Sigma(Y-\hat{Y})^2+\lambda\Sigma\beta^2)$
  - $\min_{\beta}(\Sigma(Y-\hat{Y})^2+\lambda\Sigma|\beta|)$
- Regression with Regularizers/ Regularization Term
- Shrinkage Methods
- Can perform feature/model selection for you
- How to choose $\lambda\ldots$
  - Of course validation would work
- λ is a `Hyper Parameter`, meaning:
  - it is not a parameter in your model
  - it controls the complexity of your model

Example Data Set for Today

- The red wine dataset
- P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.
- We are interested in predicting wine quality
- It is a classification problem today

Example Data Set for Today

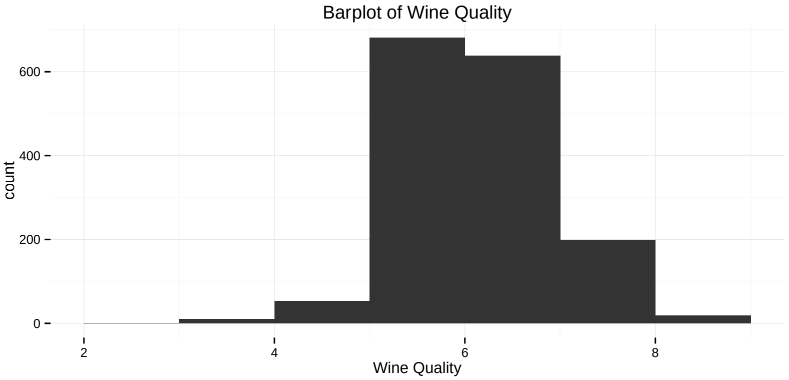
- Variables in the dataset

```r
str(wine)
```

```
## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.06
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.3
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

Example Data Set for Today

```
qplot(wine$quality, geom = "bar", binwidth = 1) +
    xlab("Wine Quality") +
    ggtitle("Barplot of Wine Quality")
```



Example Data Set for Today

- For simplicity we only do binary classification as a starting point
- Classify wines into ≤ 5 or ≥ 6

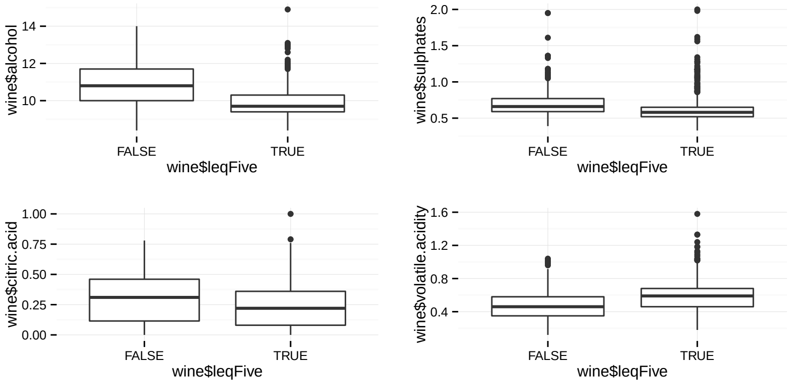
```r
wine$leqFive <- wine$quality <= 5
wine$quality <- NULL
table(wine$leqFive)
```

```
##
## FALSE  TRUE
##   855   744
```

Can We Just Use Linear Regression?

- Dependent variable only take value 0 or 1

```
p1 <- qplot(x = wine$leqFive , y = wine$alcohol, geom = "boxplot")
p2 <- qplot(x = wine$leqFive , y = wine$sulphates, geom = "boxplot")
p3 <- qplot(x = wine$leqFive , y = wine$citric.acid, geom = "boxplot")
p4 <- qplot(x = wine$leqFive , y = wine$volatile.acidity, geom = "boxplot")
grid.arrange(p1,p2,p3,p4, ncol = 2)
```



Can We Just Use Linear Regression?

- What if we use a 0.5 as threshold to turn the linear regression model predictions into 0 and 1?
- For example: if predicted value is > 0.5 we say it is high quality

```
model0 <- lm(leqFive~., wine)
pred0 <- predict(model0, wine)
pred0[1:20]
```

```
##         1         2         3         4         5         6         7
## 0.7398508 0.7339481 0.6960068 0.4804294 0.7398508 0.7271972 0.7670584
##         8         9        10        11        12        13        14
## 0.5736239 0.5835983 0.5409571 0.8111055 0.5409571 0.7245603 0.2728949
##        15        16        17        18        19        20
## 0.7724254 0.7578138 0.4191744 0.5851014 0.7678432 0.5851570
```

```
summary(pred0)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.5114  0.2607  0.4863  0.4653  0.6671  1.3070
```

```
table(pred0 > 0.5)
```

```
##
## FALSE  TRUE
##   826   773
```

How Accurate is This Approach In Sample?

- Not bad!
- This approach has a name LDA
    - Linear Discriminant Analysis
- It can do classification, but no interpretation at all.
- We often want a probability interpretation in addition to the class labels

```
table(pred0 > 0.5, wine$leqFive)
```

```
##
##          FALSE TRUE
##   FALSE    637  189
##   TRUE     218  555
```

```
accFromTable <- function(table) return(sum(diag(table))/sum(table)*100)
accFromTable(table(pred0 > 0.5, wine$leqFive))
```

```
## [1] 74.54659
```

Logistic Regression Formula

- Recall the definiation of Odds Ratio in elementary stats...

$$\frac{Pr(Y=1|X)}{1-Pr(Y=1|X)}$$

- We take the natural logarithm of the Odds Ratio

$$log(\frac{Pr(Y=1|X)}{1-Pr(Y=1|X)})$$

- What we get is called log odds or logit
- Instead of regress to the dependent variable, we regress to this logit

$$log(\frac{Pr(Y=1|X)}{1-Pr(Y=1|X)}) = \beta^T X$$

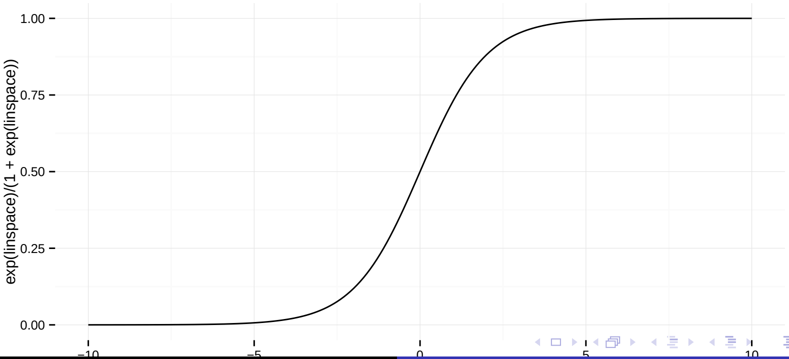
Logistic Regression Formula

$$log(\frac{Pr(Y=1|X)}{1-Pr(Y=1|X)}) = \beta^T X$$

+ If we solve for $Pr(Y=1|X)$, we can get:

$$Pr(Y=1|X) = \frac{e^{\beta^T X}}{1+e^{\beta^T X}}$$

```r
linspace = seq(-10,10,0.1)
qplot(x = linspace, y = exp(linspace)/(1+exp(linspace)), geom = "line")
```



Logistic Regression in R

- Much better interpretability

```
model1 <- glm(leqFive~., data = wine, family = binomial)
pred1 <- predict(model1, wine,type = "response")
pred1[1:20]
```

```
##         1         2         3         4         5         6         7
## 0.7840295 0.7776263 0.7365512 0.4678567 0.7840295 0.7709312 0.8034563
##         8         9        10        11        12        13        14
## 0.5878975 0.6007995 0.5305674 0.8381712 0.5305674 0.7591721 0.2290651
##        15        16        17        18        19        20
## 0.8074906 0.7946125 0.3848017 0.5906242 0.8029482 0.5962617
```

```
summary(pred1)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.003063 0.201500 0.474200 0.465300 0.710400 0.989000
```

```
accFromTable(table(pred1 > 0.5, wine$leqFive))
```

```
## [1] 74.42151
```

Model Selection for Logistic Regression

- Now... we are facing the same problem as with linear regression
- How many and what variables should we use in the logistic regression model?
- There is no adjusted estimatation for out of sample now
    - No $RSS(SSE)$ and σ^2 for classification
    - So, No $adj.R^2$ for logistic regression..
    - Only have contengency table

```
table(pred1 > 0.5, wine$leqFive)
```

```
##
##          FALSE TRUE
##   FALSE    641  195
##   TRUE     214  549
```

- One thing we can get from the table is accuracy
- Will talk about other information we get from that table next time

Which Model Rank Supreme?

```
for (i in 1:11)    print(formulas[[i]])
```

```
## [1] "leqFive~alcohol"
## [1] "leqFive~alcohol+volatile.acidity"
## [1] "leqFive~alcohol+volatile.acidity+total.sulfur.dioxide"
## [1] "leqFive~alcohol+volatile.acidity+total.sulfur.dioxide+sulphates"
## [1] "leqFive~alcohol+volatile.acidity+total.sulfur.dioxide+sulphates+free.su
## [1] "leqFive~alcohol+volatile.acidity+total.sulfur.dioxide+sulphates+free.su
## [1] "leqFive~alcohol+volatile.acidity+total.sulfur.dioxide+sulphates+free.su
## [1] "leqFive~alcohol+volatile.acidity+total.sulfur.dioxide+sulphates+free.su
## [1] "leqFive~alcohol+volatile.acidity+total.sulfur.dioxide+sulphates+free.su
## [1] "leqFive~alcohol+volatile.acidity+total.sulfur.dioxide+sulphates+free.su
## [1] "leqFive~alcohol+volatile.acidity+total.sulfur.dioxide+sulphates+free.su
```

Select Logitstic Regression Model

- As you may guessed, we will do validation to select logistic regression models
- We need to split the data into two parts: Training Set and Validation Set
- Validation Set can be viewed as a random sample from our data, and so is the Training Set
- And the data we have is in nature a sample of the population
- Take sample from a sample is called resample

```
set.seed(0363); n = nrow(wine)
split = sample(1:n, size = round(0.3*n), replace = F)
ValidationSet = wine[split, ]; TrainingSet = wine[-split,]
```

Naive Resampling May Produce Bias

```r
hqRatio <- function(v) return(round(table(v)[2]/sum(table(v)),3))
c(hqRatio(wine$leqFive),
  hqRatio(ValidationSet$leqFive),
  hqRatio(TrainingSet$leqFive))
```

```
##  TRUE  TRUE  TRUE
## 0.465 0.450 0.472
```

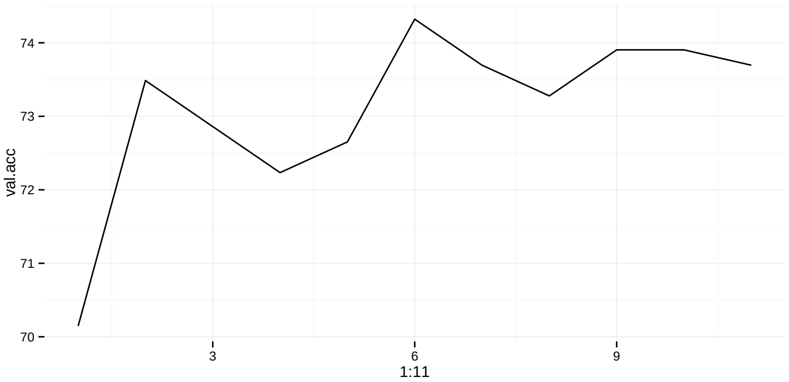
Sampling Reserve the Ratio

```
library(caTools); set.seed(1026)
split = sample.split(wine$leqFive,SplitRatio = 0.3)
ValidationSet = wine[split, ]; TrainingSet = wine[!split,]
c(hqRatio(wine$leqFive),
  hqRatio(ValidationSet$leqFive),
  hqRatio(TrainingSet$leqFive))
```

```
##  TRUE  TRUE  TRUE
## 0.465 0.466 0.465
```

Select Logistic Regression Model Using Validation

```r
val.acc <- vector()
for (i in 1:11){
    model <- glm(formulas[[i]], data = wine, family = binomial)
    pred <- predict(model,ValidationSet,type = "response")>0.5
    acc <- accFromTable(table(pred, ValidationSet$leqFive))
    val.acc <- c(val.acc, acc)}
qplot(x = 1:11, y = val.acc, geom = "line")
```



Is this 6 Variables Model Better?

```
formulas[[6]]
```

```
## [1] "leqFive~alcohol+volatile.acidity+total.sulfur.dioxide+sulphates+free.su
```

```
validatedModel <- glm(formulas[[6]], data = wine, family = binomial)
accFromTable(table(predict(validatedModel,wine, type = "response")>0.5,
                          wine$leqFive))
```

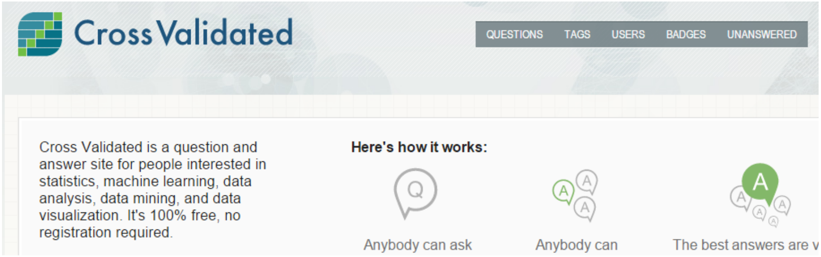
```
## [1] 74.98437
```

```
accFromTable(table(predict(model1,wine,type = "response")>0.5,
                          wine$leqFive))
```

```
## [1] 74.42151
```

- Slightly better when tested in sample
- More importantly, we are somewhat more confident that it may perform better on out of sample data than the full model as well.
- Because it is validated!

Cross Validation



Cross Validated is a question and answer site for people interested in statistics, machine learning, data analysis, data mining, and data visualization. It's 100% free, no registration required.

Here's how it works:

Anybody can ask

Anybody can

The best answers are v

http://stats.stackexchange.com/

link to cross validated

- Sometime you ask question, sometime you answer question
- Sometime this part of data used for training, sometime this part of data used for validtion

Why Resampling?

- Extract additional information about the model
- Be it prediction accuracy, the variance or bias
- We use resampling to get out of sample estimates of these things
- A single Validation set approach is already useful, cross validation further reduces possible bias in estimation
- The ability to generalize to future unseen cases is the main concern for all machine learning practitioner

Creating folds

- We will do a 5-fold cross-validation here

```r
library(caret); set.seed(1126)
```

```
## Loading required package: lattice
```

```r
folds <- createFolds(wine$leqFive, k = 5)
folds[[1]][1:20]
```

```
##  [1]   1   5  17  21  23  26  44  70  75  84  86  88  99 100 107 109 126
## [18] 133 137 141
```

```r
folds[[2]][1:20]
```

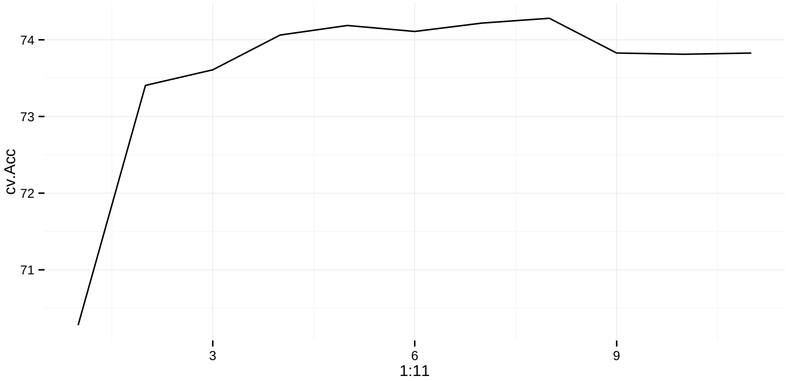
```
##  [1]  3  4  6 12 19 28 30 35 43 45 50 56 71 72 76 79 82 83 87 92
```

Cross Validation On Logistic Regression Model

```r
cv.Acc <- vector()
for (i in 1:11){
    Acc <- vector()
    for (k in 1:5){
        train <- wine[folds[[k]],]; val <- wine[-folds[[k]],]
        model <- glm(formulas[[i]], data = train, family = binomial)
        pred <- predict(model, val,type = "response") > 0.5
        Acc <- c(Acc, accFromTable(table(pred, val$leqFive)))
    }
    cv.Acc <- c(cv.Acc, mean(Acc))
}
```

Cross Validation On Logistic Regression Model

```r
qplot(x = 1:11, y = cv.Acc, geom = "line")
```



Is this 8 Variables Model Better?

```
formulas[[8]]
```

```
## [1] "leqFive~alcohol+volatile.acidity+total.sulfur.dioxide+sulphates+free.su
```

```
crossValidatedModel <- glm(formulas[[8]], data = wine, family = binomial)
accFromTable(table(predict(crossValidatedModel,wine,type = "response")>0.5,
                           wine$leqFive))
```

```
## [1] 74.42151
```

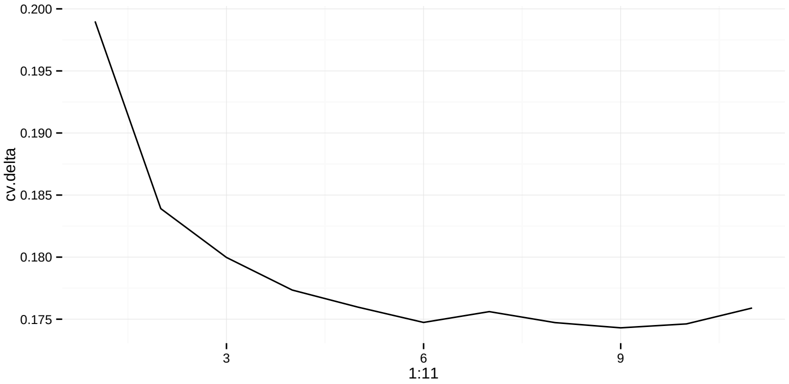
```
accFromTable(table(predict(model1,wine,type = "response")>0.5,
                           wine$leqFive))
```

```
## [1] 74.42151
```

- Same in sample accuracy, 3 variables less than model1(the full model)
- And, we have much more confident it may perform better on out of sample data than the full model. . .
- Because it is Cross-validated!

Doing Cross-Validation Using Packages

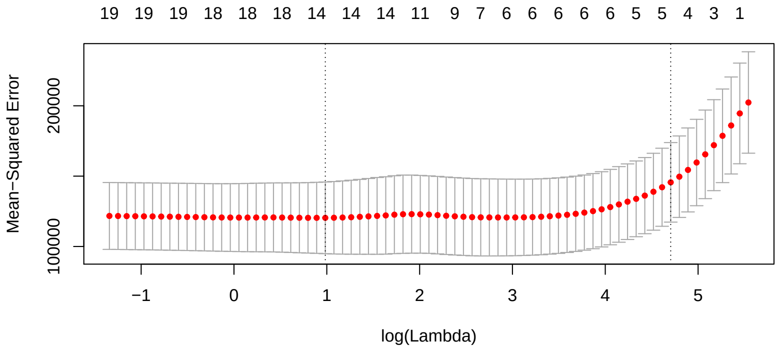
```r
library(boot); cv.delta <- vector(); set.seed(1126)
for (i in 1:11){
    fit <- glm(formulas[[i]], data = wine, family = binomial)
    cv.delta <- c(cv.delta, cv.glm(wine,fit, K = 10)$delta[2])}
qplot(x = 1:11, y = cv.delta, geom = "line")
```



Lasso Regression Revisited

- Selecting λ using cross-validation

```
library(glmnet); library(ISLR); Hitters=na.omit(Hitters);
y=Hitters$Salary; X=model.matrix(Salary~.-1,data=Hitters)
cv.lasso=cv.glmnet(X,y)
plot(cv.lasso)
```



Using Lasso as LDA to Select Features for Wine Classification?

- Six features chosen

```r
y=wine$leqFive; X=model.matrix(leqFive~.-1,wine)
cv.lasso=cv.glmnet(X,y)
coef(cv.lasso)
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                                1
## (Intercept)          1.882530663
## fixed.acidity        .
## volatile.acidity     0.491778846
## citric.acid          .
## residual.sugar       .
## chlorides            0.297214346
## free.sulfur.dioxide  .
## total.sulfur.dioxide 0.001789643
## density              .
## pH                   .
## sulphates           -0.350137013
## alcohol             -0.149239966
```

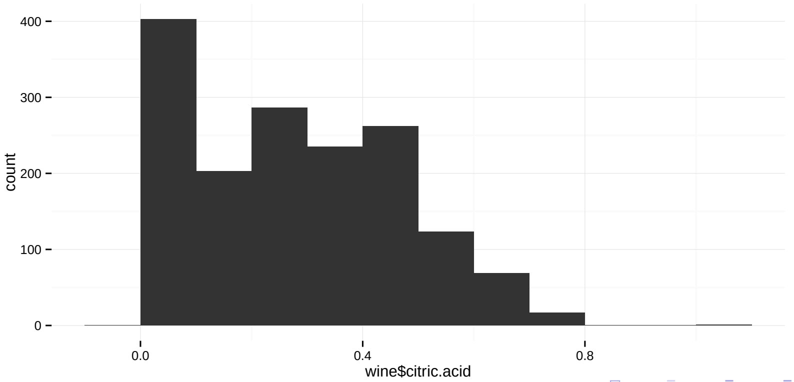
Bootstrap

- What is a Bootstrap ?

Bootstrap

- A resampling method that can be used to quantify the uncertainty associated with a given estimator
- What is the mean, se with crtric.acid ?
- We can't resample from the population, if we can, we can get the real sampling distribution
- We can however resample from the sample

```r
qplot(wine$citric.acid, binwidth =0.1)
```



Bootstrap

- Use bootstrap to estimate population mean and standard error

```
boot.mean <- vector()
set.seed(123)
for (i in 1:99999){
    bootstrapSample <- sample(wine$citric.acid, nrow(wine), replace = T)
    boot.mean <- c(boot.mean, mean(bootstrapSample))}
mean(boot.mean) - mean(wine$citric.acid)
```

```
## [1] 1.063669e-05
```

```
sd(boot.mean)
```

```
## [1] 0.004860836
```

```
sd(wine$citric.acid)/sqrt(nrow(wine))
```

```
## [1] 0.004871551
```

Bootstrap

- Use bootstrap to estimate mean and se of regression coefficients

```
boot.slope <- vector()
set.seed(123)
for (i in 1:999){
    bootstrapSample <- sample(1:nrow(wine), nrow(wine), replace = T)
    model <- glm(leqFive~., data = wine[bootstrapSample,], family = binomial)
    boot.slope <- c(boot.slope, model$coefficients['fixed.acidity'])
}
```

Bootstrap

```
c(mean(boot.slope),sd(boot.slope))
```

```
## [1] -0.1354663  0.1010405
```

```
summary(model1)$coefficients['fixed.acidity',]
```

```
##    Estimate  Std. Error     z value    Pr(>|z|)
## -0.13598034  0.09848346 -1.38074290  0.16735803
```

- Didn’t seem to be perticular useful?

Bootstrap

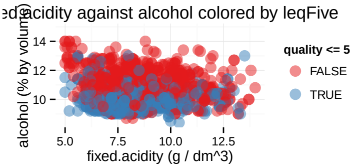
- Let's create a confidence interval of $adj.R^2$

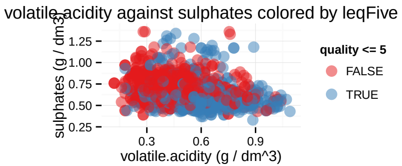
```r
boot.adj.r.square <- vector()
set.seed(123)
for (i in 1:999){
    bootstrapSample <- sample(1:nrow(Hitters), nrow(Hitters), replace = T)
    model <- lm(Salary~., data = Hitters[bootstrapSample,])
    boot.adj.r.square <- c(boot.adj.r.square, summary(model)$adj.r.square)
}
mean(boot.adj.r.square) + 1.96 * sd(boot.adj.r.square) * c(-1,1)
```

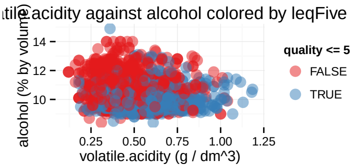
```
## [1] 0.4224755 0.6928918
```

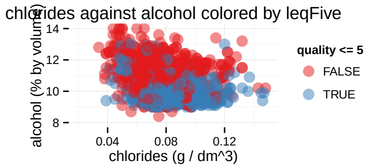
Some Plots to Look At

- We will go on talking about classification methods next week
- Here are a few plots to look and think about before next week


ed acidity against alcohol colored by leqFive
alcohol (% by volume)
fixed.acidity (g / dm^3)
quality <= 5
FALSE
TRUE


volatile.acidity against sulphates colored by leqFive
sulphates (g / dm3)
volatile.acidity (g / dm^3)
quality <= 5
FALSE
TRUE


tile acidity against alcohol colored by leqFive
alcohol (% by volume)
volatile.acidity (g / dm^3)
quality <= 5
FALSE
TRUE


chlorides against alcohol colored by leqFive
alcohol (% by volume)
chlorides (g / dm^3)
quality <= 5
FALSE
TRUE