

The Validation Set Approach

Contents

1	Load the Data	1
2	Leave-One-Out Cross-Validation	1
3	k-Fold Cross-Validation	2

1 Load the Data

```
library(ISLR2)
set.seed(1)
train <- sample(392, 196)

names(Auto)

## [1] "mpg"          "cylinders"    "displacement" "horsepower"   "weight"
## [6] "acceleration" "year"        "origin"       "name"

lm.fit <- lm(mpg ~ horsepower, data = Auto, subset = train)

attach(Auto)
# -train index below selects only the observations that are not in the training set.
mean((mpg - predict(lm.fit, Auto))[-train]^2)

## [1] 23.26601

lm.fit2 <- lm(mpg ~ poly(horsepower, 2), data = Auto,
              subset = train)
mean((mpg - predict(lm.fit2, Auto))[-train]^2)

## [1] 18.71646

lm.fit3 <- lm(mpg ~ poly(horsepower, 3), data = Auto,
              subset = train)
mean((mpg - predict(lm.fit3, Auto))[-train]^2)

## [1] 18.79401
```

2 Leave-One-Out Cross-Validation

`glm()` without passing in the family argument is the same as `lm()` :

```
glm.fit <- glm(mpg ~ horsepower, data = Auto)
coef(glm.fit)
```

```
## (Intercept) horsepower
## 39.9358610 -0.1578447
```

```

lm.fit <- lm(mpg ~ horsepower, data = Auto)
coef(lm.fit)

## (Intercept)  horsepower
##  39.9358610  -0.1578447

library(boot)
glm.fit <- glm(mpg ~ horsepower, data = Auto)
cv.err <- cv.glm(Auto, glm.fit)
cv.err$delta

## [1] 24.23151 24.23114

cv.error <- rep(0, 10)
for (i in 1:10) {
  glm.fit <- glm(mpg ~ poly(horsepower, i), data = Auto)
  cv.error[i] <- cv.glm(Auto, glm.fit)$delta[1]
}
cv.error

## [1] 24.23151 19.24821 19.33498 19.42443 19.03321 18.97864 18.83305 18.96115
## [9] 19.06863 19.49093

```

3 k-Fold Cross-Validation