Predictive Modeling

Chapter 2: A short tour of the predictive modeling process

STA 6543

The University of Texas at San Antonio

UTSA The University of Texas at San Antonio™

A Short Tour of the Predictive Modeling Process

Chapter 2

Case study: predicting fuel economy

- The fueleconomy.gov web site lists different estimates of fuel economy for passenger cars and trucks.
- For each vehicle, various characteristics are recorded such as the engine displacement or number of cylinders. Along with these values, laboratory measurements are made for the city and highway miles per gallon (MPG) of the car.
- Task: we would build a model on as many vehicle characteristics (X) as possible in order to find the most predictive model for predicting fuel efficiency (Y).

$$Y = f(X) + \epsilon,$$

$y = X\beta + \varepsilon$ ↑ Linear regression model

UTSA The University of Texas at San Antonio™

Case study: predicting fuel economy

- For illustrative purposes, we focus on a single predictor, engine displacement (EngDispl), and a single response, unadjusted highway MPG for 2010–2011 model year cars.

```
> library(AppliedPredictiveModeling)
> data(FuelEconomy)
> names(cars2010)
 [1] "EngDispl"           "NumCyl"             "Transmission"       "FE"
 [5] "AirAspirationMethod" "NumGears"           "TransLockup"        "TransCreeperGear"
 [9] "DriveDesc"          "IntakeValvePerCyl"  "ExhaustValvesPerCyl" "CarlineClassDesc"
[13] "VarValveTiming"     "VarValveLift"
```

- The common terminology would be that the 2010 data are used as the model "training set" and the 2011 values are the "test" or "validation" set.

build predictive models → prediction for the 'test' data set

The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249

Case study: predicting fuel economy

```
install.packages(AppliedPredictiveModeling) #install this package into R
library(AppliedPredictiveModeling)
data(FuelEconomy)

## Format data for plotting against engine displacement
## Sort by engine displacement
cars2010 <- cars2010[order(cars2010$EngDispl),]
cars2011 <- cars2011[order(cars2011$EngDispl),]

## Combine data into one data frame
cars2010a <- cars2010
cars2010a$Year <- "2010 Model Year"
cars2011a <- cars2011
cars2011a$Year <- "2011 Model Year"

plotData <- rbind(cars2010a, cars2011a)

library(lattice)
xyplot(FE ~ EngDispl|Year, plotData, xlab = "Engine Displacement", ylab = "Fuel Efficiency (MPG)",
    between = list(x = 1.2))
```

training data

test data

introduce a new predictor

year : categorical random variable having 2 levels

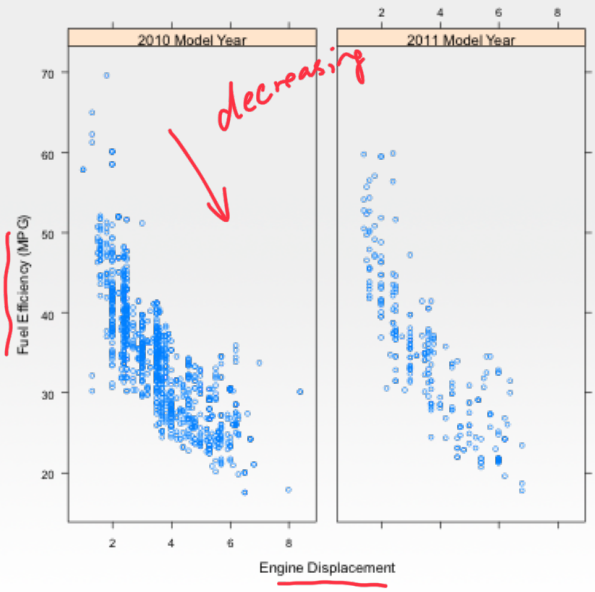
'2010 model year'

'2011 model year'

given

UTSA The University of Texas at San Antonio™

The relationship between engine displacement and fuel efficiency of all 2010 model year vehicles and new 2011 car lines



$$FE = \beta_0 + \beta_1 \underset{\text{Eng. disp}}{X} + \epsilon$$

The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249

Model 1: Fit a single linear model and conduct 10-fold CV to estimate the error

```
library(caret)
set.seed(1)
lm1Fit <- train(FE ~ EngDispl,
                data = cars2010,
                method = "lm",
                trControl = trainControl(method= "cv"))
lm1Fit
summary(lm1Fit)
#Fitted linear regression line
#efficiency = 50.5632 - 4.5209*displacement

#Quality of fit diagnostics
par(mfrow=c(1,2))
plot(cars2010$EngDispl, cars2010$FE,xlab = "Engine Displacement", ylab = "Fuel Efficiency (MPG)")
lines(cars2010$EngDispl, fitted(lm1Fit), col=2, lwd=2)

Observed =cars2010$FE
Predicted= fitted(lm1Fit)
plot(Observed, Predicted, ylim=c(12, 70))
```

'train'

$\equiv$ lm(FE ~ EngDispl, data = cars2010)

$\hat{\beta}_0$ $\hat{\beta}_1$

UTSA The University of Texas at San Antonio™

Linear model summary

```
> summary(lm1Fit)

Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-14.486  -3.192  -0.365   2.671  27.215

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  50.5632     0.3985  126.89   <2e-16 ***
EngDispl     -4.5209     0.1065  -42.46   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.624 on 1105 degrees of freedom
Multiple R-squared:   0.62,     Adjusted R-squared:  0.6196
F-statistic:  1803 on 1 and 1105 DF,  p-value: < 2.2e-16
```

$\hat{\beta}_0$ $\hat{\beta}_1$

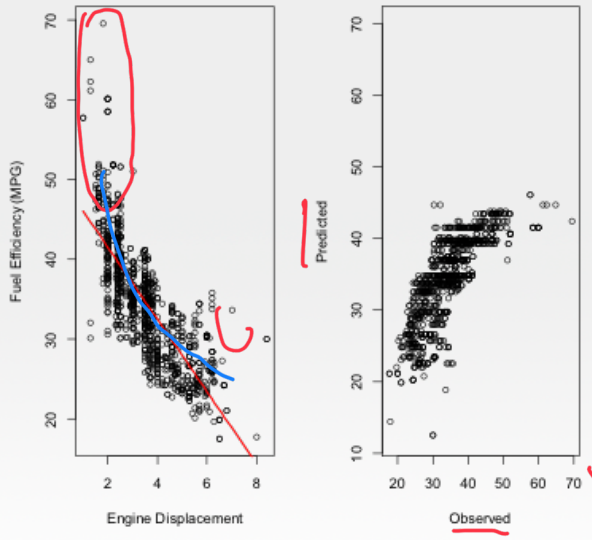
p-value < 0.05 = α

$H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$

We reject H_0 and conclude that the data provides sufficient evidence that there exists a strong relationship between x and Y

$R^2 = 0.62$

Fit a single linear model and conduct 10-fold CV to estimate the error



UTSA The University of Texas at San Antonio™

Model 2: Fit a quadratic model and conduct 10-fold CV to estimate the error

```
## Create squared terms
displacement =cars2010$EngDispl
cars2010$displacement2 =cars2010$EngDispl^2 #quadratic term
cars2011$displacement2 =cars2011$EngDispl^2

set.seed(1)
lm2Fit <- train(FE ~ EngDispl + displacement2,
                data = cars2010,
                method = "lm",
                trControl = trainControl(method= "cv"))
lm2Fit
summary(lm2Fit)

par(mfrow=c(1,2))
#Quality of fit diagnostics
plot(cars2010$EngDispl, cars2010$FE,xlab = "Engine Displacement", ylab = "Fuel Efficiency (MPG)")
lines(displacement, fitted(lm2Fit), col=2, lwd=2)

Observed =cars2010$FE
Predicted= fitted(lm2Fit)
plot(Observed, Predicted, ylim=c(12, 70))
```


The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249

Quadratic model summary

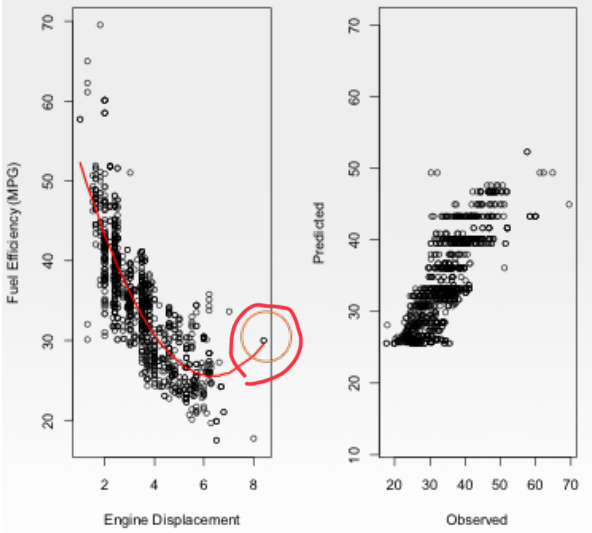
```
> summary(lm2Fit)

Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-19.1862  -2.5977  -0.3028   2.6089  24.7385

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    63.23343    0.94761   66.73   <2e-16 ***
EngDispl      -11.86738    0.51627  -22.99   <2e-16 ***
displacement2   0.93510    0.06453   14.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.241 on 1104 degrees of freedom
Multiple R-squared:  0.6807,    Adjusted R-squared:  0.6801
F-statistic:  1177 on 2 and 1104 DF,  p-value: < 2.2e-16
```

x^2

$R^2 = 0.6807$

UTSA The University of Texas at San Antonio™

Fit a quadratic model and conduct 10-fold CV to estimate the error



Nearby boundary?

As Engine displacement increases, fuel effiency will also increases, which makes no sense!

over-fitting issue

The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249

Model 3: Fit a MARS model (via the earth package)

Multivariate Adaptive Regression spline (MARS)

```
library(earth)
set.seed(1)
marsFit <- train(FE ~ EngDispl,
                 data = cars2010,
                 method = "earth",
                 tuneLength = 15,
                 trControl = trainControl(method= "cv"))
marsFit
summary(marsFit)
plot(marsFit)

par(mfrow=c(1,2))
#Quality of fit diagnostics
plot(cars2010$EngDispl, cars2010$FE,xlab = "Engine Displacement", ylab = "Fuel Efficiency (MPG)")
lines(displacement,fitted(marsFit), col=2, lwd=2)

Observed =cars2010$FE
Predicted= fitted(marsFit)
plot(Observed, Predicted, ylim=c(12, 70))
```

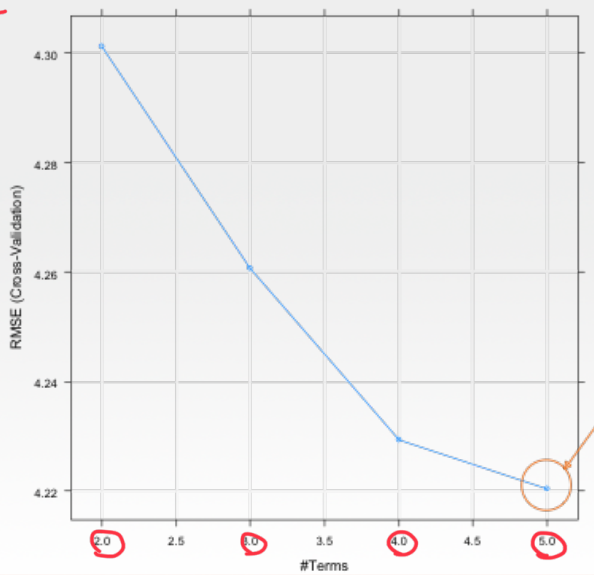
instead of 'Lm'

Mars model has a tuning parameter

UTSA The University of Texas at San Antonio™

Fit a MARS model (via the earth package)

Tuning parameter is # of terms in the model.



Minimum RMSE

when # of terms = 5.

The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249

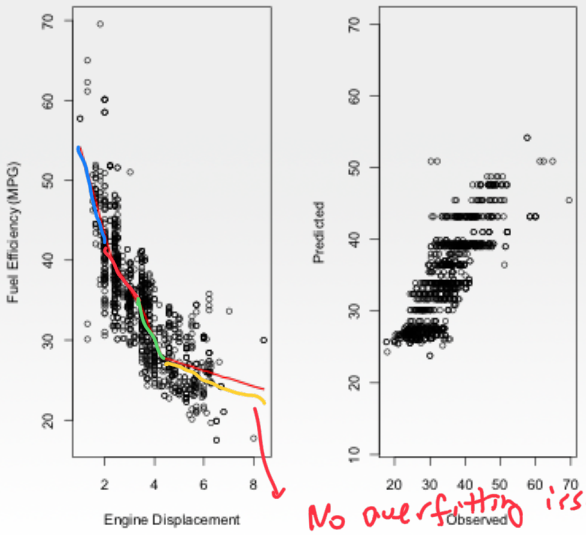
MARS model summary

```
> summary(marsFit)
Call: earth(x=c(1,1,1.3,1.3,1...), y=c(57.8,57.8,30...), keepxy=TRUE, degree=1,
            nprune=5)

                coefficients
(Intercept)        18.049001
h(EngDispl-2.3)     5.942387
h(EngDispl-3.5)    -2.552332
h(4.3-EngDispl)    10.925240
h(EngDispl-4.3)    -4.378082

Selected 5 of 5 terms, and 1 of 1 predictors
Termination condition: RSq changed by less than 0.001 at 5 terms
Importance: EngDispl
Number of terms at each degree of interaction: 1 4 (additive model)
GCV 17.97703    RSS 19578.3    GRSq 0.6805294    RSq 0.6851344
```

R^2

UTSA The University of Texas at San Antonio™

Fit a MARS model (via the earth package)



No overfitting is one nearby boundary.

The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249

Prediction performance

```
## Predict the test set data
cars2011$lm1  <- predict(lm1Fit,  cars2011)
cars2011$lm2  <- predict(lm2Fit,  cars2011)
cars2011$mars <- predict(marsFit, cars2011)

## Get test set performance values via caret's postResample function

postResample(pred = cars2011$lm1,  obs = cars2011$FE)
postResample(pred = cars2011$lm2,  obs = cars2011$FE)
postResample(pred = cars2011$mars, obs = cars2011$FE)
```

test data

test

UTSA The University of Texas at San Antonio™

Prediction performance

Mean Absolute Error

	Model	RMSE	R2	MAE
lm1	Simple linear model	5.1625309	0.7018642	3.8699447
lm2	Quadratic linear model	4.7162853	0.7486074	3.6638836
mars	A MARS model	**4.6731291**	**0.7506505**	**3.6101626**

- According to your selected criteria, which predictive should be preferred for predicting fuel efficiency? Why?

the smaller the RMSE (MAE), the better the performance for prediction

the larger the R^2, the better the model for prediction!

The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249

Introduction to R

Exercise 1