Untitled

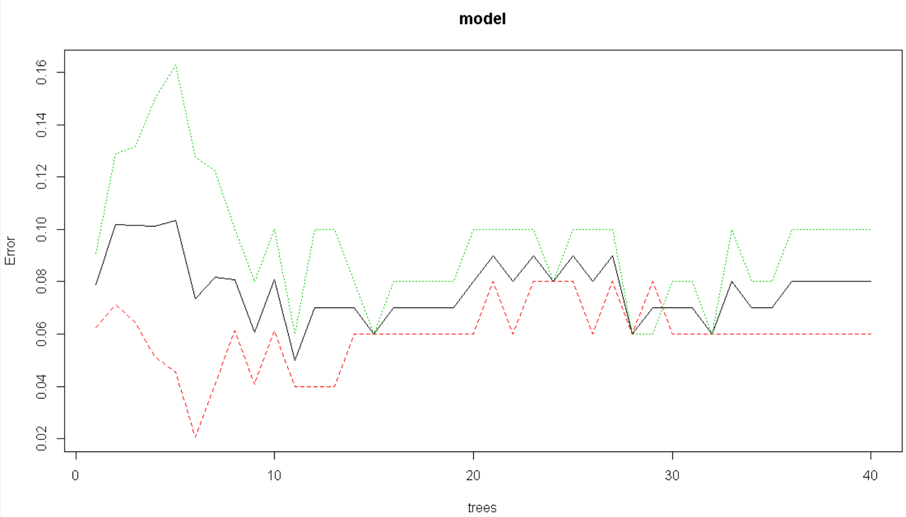
Random Forest

Data Description (from R):

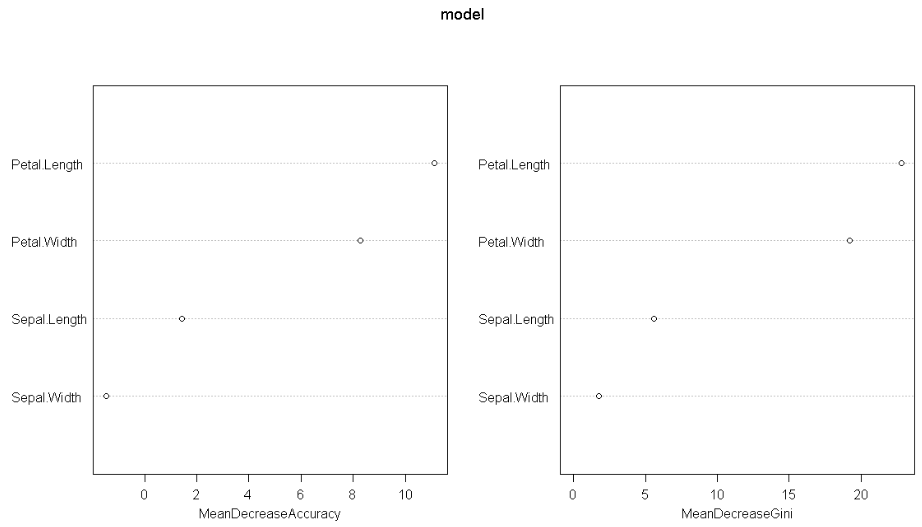
This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica. Out of these species, setosa was ignored to convert it into binary classification problem.

Variable Description:
Iris data set comprises of the following variables"
Independent variables: Sepal.Length, Sepal.Width, Petal.Length, Petal.Width
Dependent variable: Species

Error Rate vs Number of Trees:



Variable Importance:



Code:

```
data(iris)
iris <- iris[iris$Species!="setosa", ]
iris$Species <- as.factor(as.character(iris$Species))
library(randomForest)
library(SDMTools)
set.seed(123)
model <- randomForest(formula = Species ~ ., data=iris, ntree=40, importance=T)
plot(model)
prediction <- predict(model, iris)
actual <- rep(1, nrow(iris))
actual[iris$Species=="virginica"] <- 0
pred <- rep(1, nrow(iris))
pred[prediction=="virginica"] <- 0
confusion.matrix(obs = actual, pred = pred)
varImpPlot(model)
```


Classification Results:

Key: 1="versicolor", 0="virginica"

```
   obs
pred  0  1
  0 50  0
  1  0 50
```

Discussion:

Random forest model builds collection of decision trees, each decision tree built using randomly selected set of variables as predictors. The results (votes or probabilities) of all trees are averaged to reduce the variance of the prediction. Hence it can be used for variable selection.

In case of iris data, the classification accuracy in training data is 100%, which is an improvement to decision tree model (96%). As in decision tree model, order of predictor importance is Petal.Length > Petal.Width > (Sepal.Length > Sepal.Width) - variables in bracket were not used by decision tree algorithm. Least error rate is observed when number of trees is more than 1. Hence growing a forest is more advantageous.

Feb 8th, 2016

MORE YOU MIGHT LIKE

Classification Tree

Logistic Regression - University

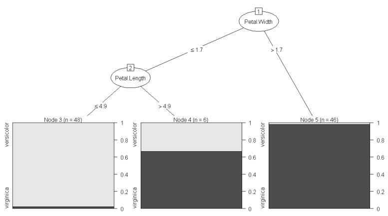
Multiple Regression - Motor Trend Car

Linear Regress Model -

Data Description (from R):
This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica. Out of these species, setosa was ignored to convert it into binary classification problem.

Variable Description:
Iris data set comprises of the following variables"
Independent variables: Sepal.Length, Sepal.Width, Petal.Length, Petal.Width
Dependent variable: Species

Classification Tree:



Code (written in R v 3.2.3):
data(iris)
iris <- iris[iris$Species!="setosa",]
library(C50)
set.seed(1)
model <- C5.0(formula = Species ~ ., data=iris)
plot(model)
print(summary(model))

Discussion:
Classification tree was built in order to check for non-linearity of relationship between outcome variable (species) and predictors. (Maximum) Information gain criteria is computed in each step for generating the binary splits in the tree.

The training set has highest entropy (-0.5*log_2(0.5) - 0.5*log_2(0.5) = 1). Hence the first iteration will always be an improvement, irrespective of how close the probabilities are to 0.5. However, the first iteration yielded good classification. Out of 46 cases in the right leaf, 1 is versicolor and 45 are virginica for Petal.Width >1.7. Out of 54 cases on the left leaf, 5 are virginica and 49 are versicolor for Petal.Width <= 1.7.

Admission Dataset

Data Description:

The data was provided in UCLA's website. It contains data on 400 applications to the university with the objective to study the likelihood of getting admission in graduate program at UCLA. **Hypothesis:** Likelihood of admission depends on GRE score, GPA and rank in undergraduate program.

Variable Descriptions:

admit Binary - whether or not the candidate got admission
gre Candidate's GRE score (average of section scores)
gpa Candidate's undergraduation GPA taken with common denominator of 4
rank Candidate's undergraduation rank

Program and Output (Analysis performed in R version 3.2.3):

Road Tests Dataset

Data Description (from R documentation):

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

Variable Descriptions (from R documentation):

The dataset is in data frame form with 11 variables described below:

mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (1000 lbs)
qsec	1/4 mile time
vs	V/S
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors

Out of these, only mpg (outcome), hp, wt and gear (independent) were chosen for analysis based on correlation.

Program and Output (Analysis performed in R version 3.2.3):

Dataset

Data Descriptio

This dataset is or
to be used for line
demonstrates sin
It was used by st
study the height c
height of parent (
mother and father
complications). T
observations = 92
concept known a
mediocrity'.

Variable Descrip

There are two va
variable - height c
with mean = 68.0
standard deviatio
and independent
parent (continuou
68.30819 units ar
= 1.787333 units.
categorical variab

Program and Ou
performed in R 3

```
library(UsingR
data(galton)
mean(galton$ch
sd(galton$chil
mean(galton$pa
sd(galton$pare
galton$parent=
nt)
galton$child=s
)
model=lm(child
data=galton)
summary(model)
```

The second iteration led to more information gain in the left leaf: Out of 48 cases in the left leaf, 1 is virginica and 47 are versicolor for Petal.Length <= 4.9 and Petal.Width <= 1.7. Out of 6 cases in the right leaf, 4 are virginica and 2 are versicolor for Petal.Length > 4.9 and Petal.Width <= 1.7. No more information gain is possible even when using other predictors.

Results:

Classification tree model was built to predict the specie based on petal length, petal width, sepal length and sepal width. The tree did not use sepal length and sepal width under information gain maximization setting. The model is non-linear. If the majority class is taken as predicted specie for each leaf in the training set, 4 cases out of 100 cases were misclassified. Hence the classification accuracy of the model is 96%.

```
> library(aod)
> # Loading the data
> data <-
read.csv("http://www.ats.ucla.e
du/stat/data/binary.csv")
> # Data summary
> summary(data)
 #    admit              gre
        gpa              rank

 #Min.   :0.0000   Min.
:220.0   Min.   :2.260   Min.
:1.000
 #1st Qu.:0.0000   1st
Qu.:520.0   1st Qu.:3.130   1st
Qu.:2.000
 #Median :0.0000   Median
:580.0   Median :3.395   Median
:2.000
 #Mean   :0.3175   Mean
:587.7   Mean   :3.390   Mean
:2.485
 #3rd Qu.:1.0000   3rd
Qu.:660.0   3rd Qu.:3.670   3rd
Qu.:3.000
# Max.   :1.0000   Max.
:800.0   Max.   :4.000   Max.
:4.000
> # Standard deviations
> sapply(data, sd)
#      admit         gre
  gpa        rank
#  0.4660867 115.5165364
0.3805668   0.9444602
> # Converting rank to factor
variable as it has very few
values. Combining 1,2 as factor
'0' and 3,4 as factor '1'
> data$rank[data$rank==1 |
data$rank==2] <- 0
> data$rank[data$rank>2] <- 1
> data$rank <-
as.factor(data$rank)
> # Model building (ignoring
rank to check if it is a
confounder)
> data$admit <-
as.factor(data$admit)
> data$gre <- scale(data$gre)
> data$gpa <- scale(data$gpa)
> model1 <- glm(admit ~ gre +
gpa, data=data,
family=binomial)
> summary(model1)
#
#Call:
#glm(formula = admit ~ gre +
gpa, family = binomial, data =
data)
#
#Deviance Residuals:
#    Min       1Q   Median
  3Q      Max
#-1.2730  -0.8988  -0.7206
1.3013   2.0620
#
#Coefficients:
```

```
> data(mtcars)
> sapply(mtcars, class)
      mpg       cyl      disp
     hp      drat        wt
 qsec         vs        am
 gear      carb
"numeric" "numeric" "numeric"
"numeric" "numeric" "numeric"
"numeric" "numeric" "numeric"
"numeric" "numeric"
>
indep=data.frame(cbind(mtcars$h
p, mtcars$wt, mtcars$gear))
> colnames(indep)=c("hp", "wt",
"gear")
> mean(indep$hp)
[1] 146.6875
> mean(indep$wt)
[1] 3.21725
> mean(indep$gear)
[1] 3.6875
> mean(mtcars$mpg)
[1] 20.09062
> sd(indep$hp)
[1] 68.56287
> sd(indep$wt)
[1] 0.9784574
> sd(indep$gear)
[1] 0.7378041
> sd(mtcars$mpg)
[1] 6.026948
>
indep=data.frame(scale(indep))
> mtcars$mpg=scale(mtcars$mpg)
> cor(indep)
              hp         wt
  gear
hp    1.0000000  0.6587479
-0.1257043
wt    0.6587479  1.0000000
-0.5832870
gear -0.1257043 -0.5832870
 1.0000000
>
model1=lm(mtcars$mpg~indep$hp-
1)
> summary(model1)

Call:
lm(formula = mtcars$mpg ~
indep$hp - 1)

Residuals:
    Min      1Q  Median      3Q
    Max
-0.9477 -0.3505 -0.1469  0.2625
 1.3665

Coefficients:
         Estimate Std. Error t
value Pr(>|t|)
indep$hp  -0.7762     0.1132
 -6.854 1.11e-07 ***
---
Signif. codes:  0 ‘***’ 0.001
‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1
```

```
Call:
lm(formula = (
1, data = galt

Residuals:
     Min
  3Q      Max
-3.09976 -0.54
 0.64889  2.35

Coefficients:
       Estimat
Error t-value
parent  0.4587
15.72   <2e-16
---
Signif. codes:
‘**’ 0.01 ‘*’
1

Residual stand
on 927 degrees
Multiple R-squ
Adjusted R-squ
F-statistic: 2
DF,  p-value:
```

Interpretation of

The independent 'parent', which is parent. The deper named 'child', wh child.

The first concept significance of the stat)< 2.2e-16, w (chosen as cutoff tests). Therefore, is significant. It ex variance in the ou

There are 928 ob which 1 degree o used for building of freedom = 927 ANOVA. The cor removed as it will scaling (in absen variables).

The independent p-value(t-stat) [tw 16, which is less the variable 'pare coefficient is estin

The regression e scaled(child) = 0. + error not explai

Summary of Res

The regression e through (average average child hei child of very tall p

```
#              Estimate Std.
Error z value Pr(>|z|)
#(Intercept)  -0.8098
0.1120  -7.233 4.74e-13 ***
#gre            0.3108
0.1222   2.544   0.0109 *
#gpa            0.2872
0.1216   2.361   0.0182 *
#---
#Signif. codes:  0 '***' 0.001
'**' 0.01 '*' 0.05 '.' 0.1 ' '
1
#
#(Dispersion parameter for
binomial family taken to be 1)
#
#    Null deviance: 499.98  on
399  degrees of freedom
#Residual deviance: 480.34  on
397  degrees of freedom
#AIC: 486.34
#
#Number of Fisher Scoring
iterations: 4
#
> model2 <- glm(admit ~ gre +
gpa + rank, data=data,
family=binomial)
> summary(model2)
#
#Call:
#glm(formula = admit ~ gre +
gpa + rank, family = binomial,
data = data)
#
#Deviance Residuals:
#    Min       1Q   Median
  3Q      Max
#-1.4290  -0.8902  -0.6552
1.1937   2.1122
#
#Coefficients:
#              Estimate Std.
Error z value Pr(>|z|)
#(Intercept)  -0.4109
0.1447  -2.839  0.00453 **
#gre            0.2633
0.1254   2.101  0.03568 *
#gpa            0.3288
0.1247   2.638  0.00834 **
#rank1         -0.9383
0.2329  -4.028 5.62e-05 ***
#---
#Signif. codes:  0 '***' 0.001
'**' 0.01 '*' 0.05 '.' 0.1 ' '
1
#
#(Dispersion parameter for
binomial family taken to be 1)
#
#    Null deviance: 499.98  on
399  degrees of freedom
#Residual deviance: 463.37  on
396  degrees of freedom
#AIC: 471.37
#
#Number of Fisher Scoring
```

```
Residual standard error: 0.6305
on 31 degrees of freedom
Multiple R-squared:  0.6024,
Adjusted R-squared:  0.5896
F-statistic: 46.98 on 1 and 31
DF,  p-value: 1.11e-07

>
model2=lm(mtcars$mpg~indep$hp+i
ndep$wt-1)
> summary(model2)

Call:
lm(formula = mtcars$mpg ~
indep$hp + indep$wt - 1)

Residuals:
    Min      1Q  Median      3Q
    Max
-0.6539 -0.2655 -0.0302  0.1742
 0.9713

Coefficients:
         Estimate Std. Error t
value Pr(>|t|)
indep$hp  -0.3615     0.1010
 -3.579   0.0012 **
indep$wt  -0.6296     0.1010
 -6.233 7.27e-07 ***
---
Signif. codes:  0 '***' 0.001
'**' 0.01 '*' 0.05 '.' 0.1 ' '
1

Residual standard error: 0.4231
on 30 degrees of freedom
Multiple R-squared:  0.8268,
Adjusted R-squared:  0.8152
F-statistic:  71.6 on 2 and 30
DF,  p-value: 3.791e-12

>
model3=lm(mtcars$mpg~indep$hp+i
ndep$wt+indep$gear-1)
> summary(model3)

Call:
lm(formula = mtcars$mpg ~
indep$hp + indep$wt +
indep$gear -
    1)

Residuals:
     Min       1Q   Median
  3Q      Max
-0.55936 -0.31554 -0.05714
 0.16399  1.00640

Coefficients:
           Estimate Std. Error
t value Pr(>|t|)
indep$hp    -0.4185     0.1106
 -3.785 0.000715 ***
indep$wt    -0.5192     0.1350
 -3.844 0.000609 ***
indep$gear   0.1249     0.1024
```

be tall, but not as
Similarly child of
expected to be sh
as the parent. Thi
as regression to r
behavior).

```
iterations: 3
#
> # Confidence intervals using
standard errors of log
likelihoods
> confint.default(model2)
#                  2.5 %
97.5 %
#(Intercept) -0.69459942
-0.1272177
#gre          0.01761952
 0.5090337
#gpa          0.08450421
 0.5731589
#rank1       -1.39488233
-0.4817514
> # Wald test for significance
of effect of rank
> wald.test(b=coef(model2),
Sigma=vcov(model2), Terms=4)
#Wald test:
#----------
#
#Chi-squared test:
#X2 = 16.2, df = 1, P(> X2) =
5.6e-05
```

Interpretation of Output:

All quantitative variable have been normalized ((X-mean)/sd). Rank group ‘0′ is treated as base group and the log likelihood coefficients are reported with respect to group ‘0′.

Let us consider p-value > 0.05 for not rejecting the null hypothesis in model 1. Intercept is significant and it’s estimate is -0.8098. gre is significant and it’s coefficient is estimated to be 0.3108. gpa is significant and it’s coefficient is estimated to be 0.2872. However, there is evidence of ‘rank’ being a confounder to mode model 1.

Analysis of model 2 suggests that intercept is significant with estimate = -0.4109 (CI= -0.695 to -0.127, p-value=0.00453). gre is significant with coefficient estimate = 0.2633 (unit increase in scaled gre increases log likelihood by 0.2633, CI=0.01762 to 0.509, p-value=0.0357). gpa is significant with coefficient estimate = 0.3288 (unit increase in scaled gre increases log likelihood by 0.3288, CI=0.0845 to 0.5732, p-value=0.008). rank1 is significant with coefficient estimate = -0.9383 (moving from rank group 0 to rank group 1 decrease log likelihood by 0.9383, CI= -1.395 to -0.482, p-value=5.62e-05)

Confounding:

```
  1.219 0.232593
---
Signif. codes:  0 ‘***’ 0.001
‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 0.4197
on 29 degrees of freedom
Multiple R-squared:  0.8352,
Adjusted R-squared:  0.8182
F-statistic:    49 on 3 and 29
DF,  p-value: 1.788e-11

> plot(model3)
```

Interpretation of Output:

All variables - dependent and independent - have been scaled. Intercept has been explicitly removed from linear model call as the estimate will pass through (X-avg, Y-avg) for quantitative X irrespective of the number of independent variables in X. Independent variables are correlated, but the magnitude of correlations is not very high. However, this introduces the possibility of confounding.

The first concept to be checked is the significance of the model. p-value(F-stat)< 1.788e-11, which is less than 0.05 (chosen as cutoff for most significance tests). Therefore, the regression model is significant. It explains 83.52% of the variance in the outcome variable. There are 32 observations out of which 3 have been used for model building. Hence degrees of freedom for error in ANOVA = 32-3 = 29.

The independent variable ‘hp’ has p-value(t-stat) [two sided test] = 0.000715, which is less than 0.05. Therefore, the variable ‘hp’ is significant. The coefficient is estimated to be -0.4185. The independent variable ‘wt’ has p-value(t-stat) [two sided test] = 0.000609, which is less than 0.05. Therefore, the variable ‘wt’ is significant. The coefficient is estimated to be -0.5192. The independent variable ‘gear’ has p-value(t-stat) [two sided test] = 0.232593, which is greater than 0.05. Therefore, the variable **‘gear’ is insignificant**. The coefficient is estimated to be 0.1249.

Confounding:

Adding rank to model 1 proves that rank is a significant predictor. Wald test of significance shows that rank is a significant independent variable to explain the dependent variable. Hence confounding is present in model 1.

Discussion:

Let us consider a candidate with rank group = 0, scaled gre score = 1 and scaled gpa = 1. Let p be the probability of getting admission as predicted by the model

log_e(p/(1-p)) = 0.1812496 => p = 1.198714341/2.198714341
= 0.545188758 = probability of getting admission as predicted by the model

There is clear association between the predictor and outcome variables. Hence the hypothesis has been validated. Higher gpa and higher gre score are desirable. Worse rank group is not desirable.

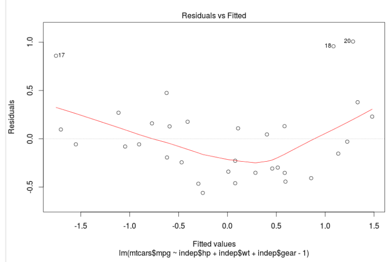
Show more

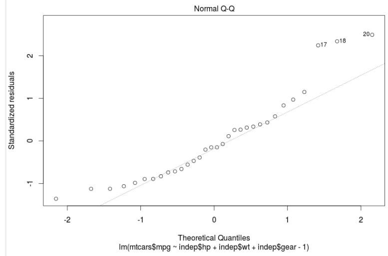
```
>
model4=lm(mtcars$mpg~indep$gear
-1)
> summary(model4)

Call:
lm(formula = mtcars$mpg ~
indep$gear - 1)

Residuals:
     Min       1Q   Median
  3Q      Max
-1.69904 -0.46347 -0.03401
 0.35272  2.08784

Coefficients:
           Estimate Std. Error
t value Pr(>|t|)
indep$gear   0.4803     0.1575
  3.049  0.00467 **
---
Signif. codes:  0 ‘***’ 0.001
‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 0.8771
on 31 degrees of freedom
Multiple R-squared:  0.2307,
Adjusted R-squared:  0.2059
F-statistic: 9.295 on 1 and 31
DF,  p-value: 0.004672
```

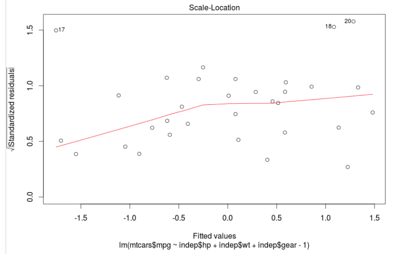
Discussion:

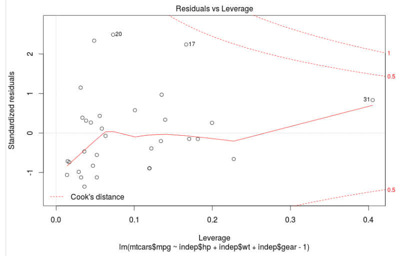
mpg(est) = - 0.4185*hp - 0.5192*wt + 0.1249*gear
p-values = (0.000715,   0.000609, 0.232593)
There is clear (significant) association between dependent variable and independent variables.
There is evidence of confounding as ‘gear’ is correlated with both independent (hp and wt) and dependent (mpg) variables and is insignificant (p-value > 0.05).

Diagnostic Plots:









Points 17, 18 and 20 deviate significantly from model behavior ("extreme point w.r.t y"). Additionally, point 31 has unusually high leverage with respect to predictors ("extreme point w.r.t x").