Mustangs Data Analysis

Gustavo Cepparo and Milica Cudina

Let's import the data set and poke around a bit.

```
library(boot)
mustangs<-read.csv("mustangs.csv")
names(mustangs)
```

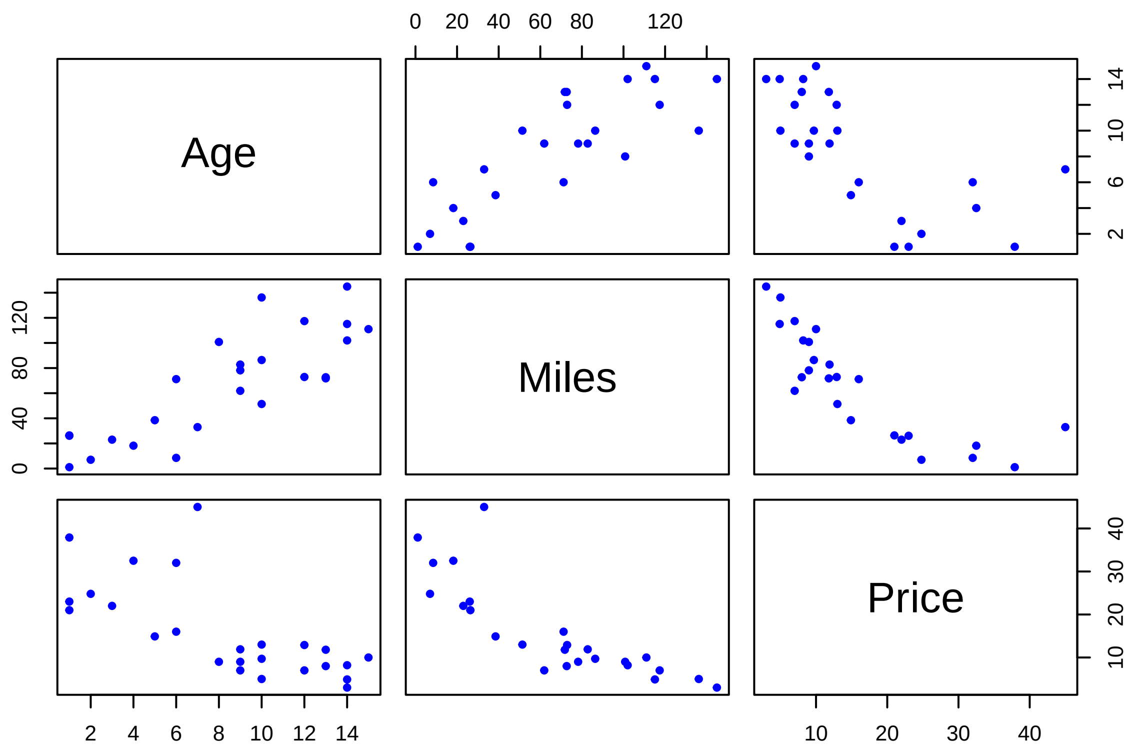
```
## [1] "Age"   "Miles" "Price"
```

```
dim(mustangs)
```

```
## [1] 25  3
```

Again, we undertake a rudimentary exploratory data analysis. It's natural to be interested in pairwise interactions. So, we create an array of scatterplots with which we can visually assess the shape of the dependence and the correlations of each pair of variables.

```
plot(mustangs,
     pch=20, col="blue")
```




```
cor(mustangs)
```

```
##             Age     Miles      Price
## Age   1.0000000 0.8249094 -0.7004497
```

```
## Miles  0.8249094  1.0000000 -0.8246164
## Price -0.7004497 -0.8246164  1.0000000
```

Let's create four models:

```r
attach(mustangs)
lm.fit.s=lm(Price~Miles)
summary(lm.fit.s)
```

```
##
## Call:
## lm(formula = Price ~ Miles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9515 -3.7189 -0.4785  3.3645 21.7251
##
## Coefficients:
##             Estimate Std. Error t value         Pr(>|t|)
## (Intercept)  30.4953     2.4415  12.491 0.00000000000989 ***
## Miles        -0.2188     0.0313  -6.991 0.00000039957781 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.422 on 23 degrees of freedom
## Multiple R-squared:   0.68,  Adjusted R-squared:  0.6661
## F-statistic: 48.87 on 1 and 23 DF,  p-value: 0.0000003996
```

```r
lm.fit.q=lm(Price~Miles+I(Miles^2))
summary(lm.fit.q)
```

```
##
## Call:
## lm(formula = Price ~ Miles + I(Miles^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4466 -3.8742 -0.0579  1.5680 22.3552
##
## Coefficients:
##               Estimate Std. Error t value       Pr(>|t|)
## (Intercept) 35.1198578  3.1808388  11.041 0.000000000193 ***
## Miles       -0.4283345  0.1046183  -4.094       0.000479 ***
## I(Miles^2)   0.0015243  0.0007308   2.086       0.048790 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6 on 22 degrees of freedom
## Multiple R-squared:  0.7328, Adjusted R-squared:  0.7085
## F-statistic: 30.17 on 2 and 22 DF,  p-value: 0.000000495
```

```r
lm.fit.m=lm(Price~Miles+Age)
summary(lm.fit.m)
```

```
##
## Call:
## lm(formula = Price ~ Miles + Age)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.784 -4.301 -0.601  3.599 21.982
##
## Coefficients:
##             Estimate Std. Error t value       Pr(>|t|)
## (Intercept)  30.8668     2.7875  11.073 0.000000000183 ***
## Miles        -0.2049     0.0565  -3.628        0.00149 **
## Age          -0.1551     0.5219  -0.297        0.76916
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.553 on 22 degrees of freedom
## Multiple R-squared:  0.6813, Adjusted R-squared:  0.6523
## F-statistic: 23.51 on 2 and 22 DF,  p-value: 0.000003449
```

```
lm.fit.all=lm(Price ~ Miles*Age)
summary(lm.fit.all)
```

```
##
## Call:
## lm(formula = Price ~ Miles * Age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4698 -4.0240 -0.6447  2.1690 22.8911
##
## Coefficients:
##              Estimate Std. Error t value     Pr(>|t|)
## (Intercept) 34.012975   4.114612   8.266 0.0000000485 ***
## Miles       -0.293682   0.102411  -2.868      0.00921 **
## Age         -0.630789   0.693886  -0.909      0.37363
## Miles:Age    0.009537   0.009188   1.038      0.31108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.542 on 21 degrees of freedom
## Multiple R-squared:  0.6968, Adjusted R-squared:  0.6535
## F-statistic: 16.09 on 3 and 21 DF,  p-value: 0.00001163
```

We can do a simple 1-cross-validation on the above models. We can define the training set in the same fashion for all:

```
set.seed(123)
n=length(Price)
train=sample(n, floor(n/2))
```

Now, we can simply calculate the MSEs on the validation models for all the models.

```
#simple linear regression
lm.fit.s<-lm(Price~Miles,data=mustangs,subset=train)
mean((Price-predict(lm.fit.s,mustangs))[-train]^2)
```

```
## [1] 55.37683
```

```r
#quadratic
lm.fit.q=lm(Price~Miles+I(Miles^2),data=mustangs,subset=train)
mean((Price-predict(lm.fit.q,mustangs))[-train]^2)
```

```
## [1] 78.26981
```

```r
#multiple linear regression
lm.fit.m=lm(Price~Miles+ Age,data=mustangs,subset=train)
mean((Price-predict(lm.fit.m,mustangs))[-train]^2)
```

```
## [1] 55.17432
```

```r
#multiple linear regression with interactions
lm.fit.mi=lm(Price~Miles*Age,data=mustangs,subset=train)
mean((Price-predict(lm.fit.mi,mustangs))[-train]^2)
```

```
## [1] 75.00728
```

With the above values in mind, we might develop a different belief about which model is preferable.

Finally, let's do LOOCV.

```r
#simple linear regression
glm.fit.s<-glm(Price~Miles,data=mustangs)
cv.err.s=cv.glm(mustangs,glm.fit.s)
cv.err.s$delta[1]
```

```
## [1] 44.29251
```

```r
#quadratic
glm.fit.q=glm(Price~Miles+I(Miles^2),data=mustangs)
cv.err.q=cv.glm(mustangs,glm.fit.q)
cv.err.q$delta[1]
```

```
## [1] 38.28141
```

```r
#multiple linear regression
glm.fit.m=glm(Price~Miles+ Age,data=mustangs)
cv.err.m=cv.glm(mustangs,glm.fit.m)
cv.err.m$delta[1]
```

```
## [1] 46.62496
```

```r
#multiple linear regression with interacions
glm.fit.mi=glm(Price~Miles*Age,data=mustangs)
cv.err.mi=cv.glm(mustangs,glm.fit.mi)
cv.err.mi$delta[1]
```

```
## [1] 47.05428
```