R_Hellow-World_Multiple-Regression

Rob McCulloch

January 22, 2018

Simple Example of Multiple Regression in R

Let's do a "hello world" example of R.

We will use the used cars data (susedcars.csv) to relate the price of a used car to it's mileage and year.

We will:

- read in the data from a *.csv file into an R data.frame
- process the data by selecting a few columns and rescaling two of them
- do some simple summaries
- do some simple plots
- run the multiple regression
- plot y versus $\hat{y}$.
- make some predictions

Read in the data and get the variables we want

```
cd = read.csv("susedcars.csv")
names(cd)
```

```
## [1] "price"        "trim"         "isOneOwner"   "mileage"
## [5] "year"         "color"        "displacement"
```

```
cd = cd[,c(1,4,5)]
cd$price = cd$price/1000
cd$mileage = cd$mileage/1000
head(cd)
```

```
##    price mileage year
## 1 43.995  36.858 2008
## 2 44.995  46.883 2012
## 3 25.999 108.759 2007
## 4 33.880  35.187 2007
## 5 34.895  48.153 2007
## 6  5.995 121.748 2002
```

```
summary(cd)
```

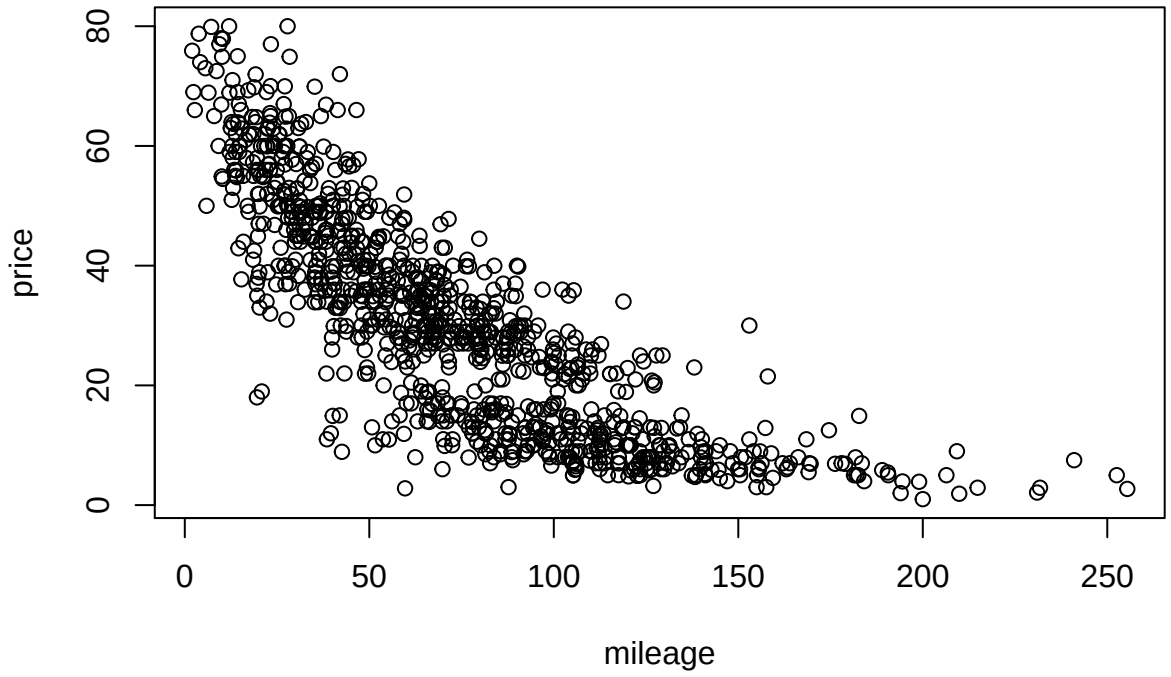
```
##      price           mileage             year
##  Min.   : 0.995   Min.   :  1.997   Min.   :1994
##  1st Qu.:12.995   1st Qu.: 40.133   1st Qu.:2004
##  Median :29.800   Median : 67.919   Median :2007
##  Mean   :30.583   Mean   : 73.652   Mean   :2007
##  3rd Qu.:43.992   3rd Qu.:100.138   3rd Qu.:2010
##  Max.   :79.995   Max.   :255.419   Max.   :2013
```

```
cor(cd)
```

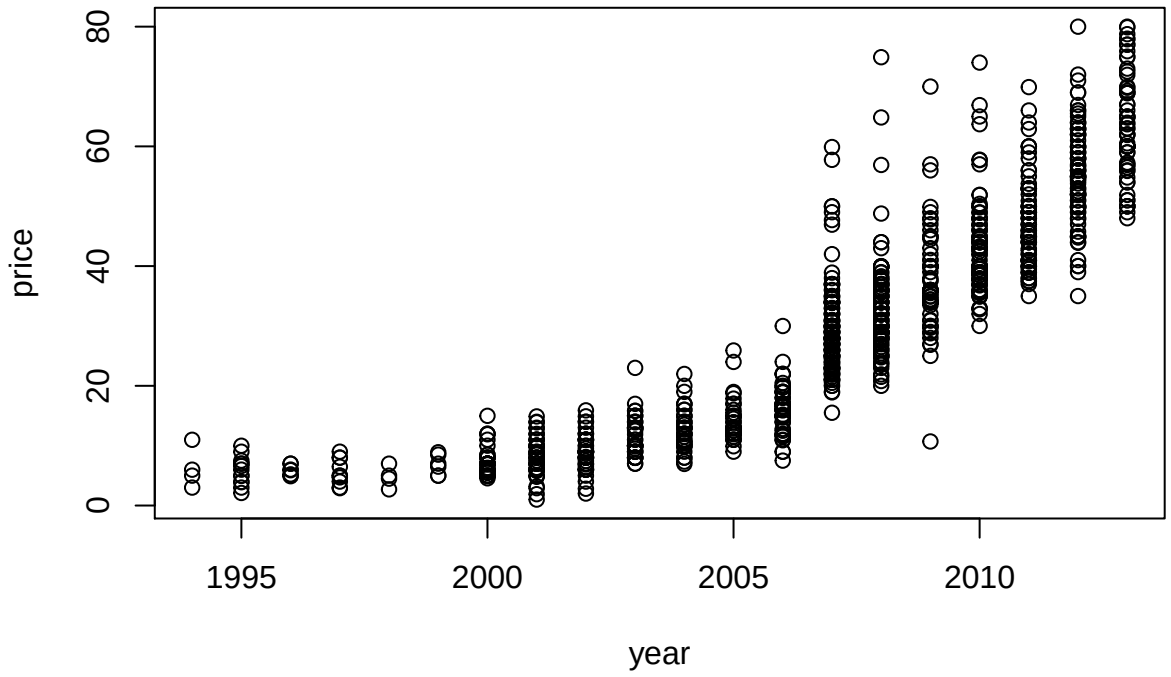
```
##               price    mileage       year
## price     1.0000000 -0.8152458  0.8805373
## mileage -0.8152458  1.0000000 -0.7447292
## year      0.8805373 -0.7447292  1.0000000
```

Plot y versus each x

```
plot(cd$mileage,cd$price,xlab="mileage",ylab="price")
```




```
plot(cd$year,cd$price,xlab="year",ylab="price")
```



Run the Regression of y=price on X=(mileage,year)

Ok, now we can run the regression.

```
lmmod = lm(price~mileage+year,cd)
summary(lmmod)
```

```
##
## Call:
## lm(formula = price ~ mileage + year, data = cd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.857  -4.855  -1.670   3.483  34.499
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.365e+03  1.716e+02  -31.27   <2e-16 ***
## mileage     -1.537e-01  8.339e-03  -18.43   <2e-16 ***
## year         2.694e+00  8.526e-02   31.60   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.544 on 997 degrees of freedom
## Multiple R-squared:  0.8325, Adjusted R-squared:  0.8321
## F-statistic:  2477 on 2 and 997 DF,  p-value: < 2.2e-16
```

```
cat("the coefficients are:",lmmod$coefficients,"\n")
```

```
## the coefficients are: -5365.49 -0.1537219 2.69435
```

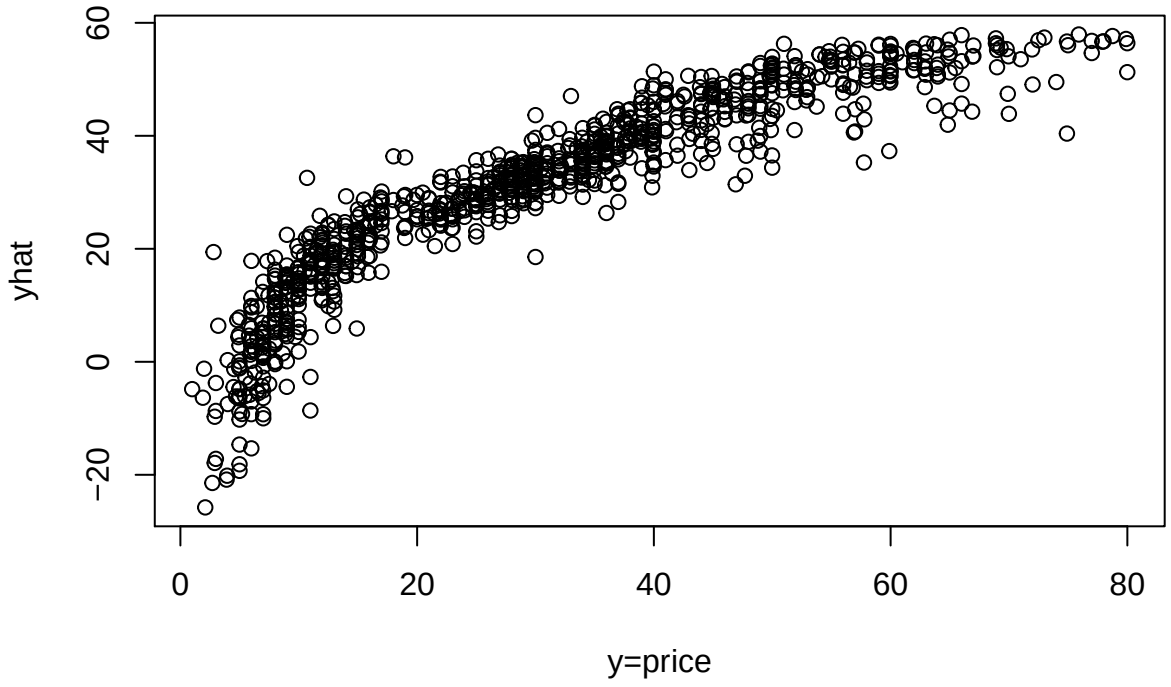
So, the fitted relationship is

$$price = -53695.49 - 0.154\,mileage + 2.7\,year$$

Get and Plot the Fits

We will pull off the fitted values and plot them versus y.

```
yhat = lmmod$fitted.values
plot(cd$price,yhat,xlab="y=price")
```



Clearly, it is really bad !!

Predictions

```
xpdf = data.frame(mileage=c(40,100),year=c(2010,2004))
ypred = predict(lmmod,xpdf)
cat("at x:")
```

```
## at x:
```

```
xpdf
```

```
##   mileage year
## 1      40 2010
## 2     100 2004
```

```
cat("predicted price is\n")
```

```
## predicted price is
```

```
ypred
```

```
##        1        2
## 44.00383 18.61442
```