The *R* library `MASS` includes the dataframe `Cars93`. It includes 26 variables, some of which are categorical. We consider predicting the `Price` of a car based on its performance showed by the city mileage of the car, `MPG.city`, and the number of `Airbags` included. Consider the following steps.

1. Verify `Airbags` is a categorical variable. How many observations are in each level?

2. Rename the levels of `Airbags` as `"DP","D",` and `"O"` for `Driver & Passenger, Driver only` and no airbag.

3. Fit a MLR model for `Price` with `MPG.city` and `Airbags` as predictors. What is the base level of Airbags?

4. Change the base level to no airbags, labeled as `"O"`.

5. Refit

6. Write the fitted equations for the mean `Price`.

7. Plot fitted vs observed prices (that is, yhat vs y). Identify outliers.

8. Find a 95% CI on the mean price of a car with 27.5 city mileage and Driver only airbags.

9. Find a 90% PI on the price of a car with 27.5 city mileage and Driver only airbags.

```
# 1cat1con.r       Price = f(MPG.city, Airbags)   3 levels(Airbags)

library(MASS)
library(PASWR2)
d0 = Cars93
str(d0)

table(d1$AirBags)
# DP  D  0
# 16 43 34

d1 = d0[,c(5,7,9)]
head(d1)
levels(d1$AirBags) # "Driver & Passenger" "Driver only"  "None"

# rename levels
levels(d1$AirBags)=c("DP","D","0")
levels(d1$AirBags)
# [1] "DP" "D"  "0"

# fit
#=====================================================================
m1=lm(Price~MPG.city*AirBags,d1)
summary(m1)

# Coefficients:
#                    Estimate Std. Error t value Pr(>|t|)
# (Intercept)         85.9565    14.3376   5.995 4.51e-08 ***
# MPG.city            -2.9438     0.7282  -4.043 0.000114 ***
# AirBagsD           -42.7046    15.0074  -2.846 0.005527 **
# AirBags0           -60.5693    14.9876  -4.041 0.000114 ***
# MPG.city:AirBagsD   1.9253     0.7551   2.550 0.012532 *
# MPG.city:AirBags0   2.4476     0.7481   3.272 0.001533 **
#
# Residual standard error: 6.51 on 87 degrees of freedom
# Multiple R-squared:  0.5704,    Adjusted R-squared:  0.5457
# F-statistic:  23.1 on 5 and 87 DF,  p-value: 1.085e-14

# AirBagsDP is the base level (it does not show up in Coeff Table)

# AirBagsD shows the additional intercept if cars has Driver only bag
# AirBags0 shows the additional intercept if cars has no airbag
# MPG.city:AirBagsD shows the additional slope if cars has Driver only bag
# change the base level
#=======================================================================
levels(d1$AirBags)
#[1] "DP" "D"  "0"

# one at a time
d1$AirBags=relevel(d1$AirBags,"D")
levels(d1$AirBags)   #[1] "D"  "DP" "0"
d1$AirBags=relevel(d1$AirBags,"0")
levels(d1$AirBags)   #[1] "0"  "D"  "DP"

# refit
m2=lm(Price˜MPG.city*AirBags,d1)
summary(m2)

# Coefficients:
#                     Estimate Std. Error t value Pr(>|t|)
# (Intercept)          25.3873     4.3658   5.815 9.83e-08 ***
# MPG.city             -0.4961     0.1714  -2.894 0.004809 **
# AirBagsD             17.8647     6.2221   2.871 0.005135 **
# AirBagsDP            60.5693    14.9876   4.041 0.000114 ***
# MPG.city:AirBagsD    -0.5224     0.2633  -1.984 0.050366 .
# MPG.city:AirBagsDP  -2.4476     0.7481  -3.272 0.001533 **

# Residual standard error: 6.51 on 87 degrees of freedom
# Multiple R-squared:  0.5704,    Adjusted R-squared:  0.5457
# F-statistic:  23.1 on 5 and 87 DF,  p-value: 1.085e-14


# fitted equations (Price as function of city mileage)
#=====================================================================

# No airbags   E[Y] =  25.3873 - 0.4961 MPG
# Driver only  E[Y] = (25.3873 + 17.8647) + (-0.4961-0.5224) MPG
# Two airbags  E[Y] = (25.3873 + 60.5693) + (-0.4961-2.4476) MPG
# plotting
#==========================================================================
predicted=m2$fitted
plot(predicted~Price,d1,pch=19,cex=0.6)
abline(0,1)
grid()

a = c(0,70)
plot(predicted~Price,d1,xlim=a,ylim=a,pch=19,cex=0.6)
text(predicted~d1$Price,labels=rownames(d0),pos=1,cex=0.5)
abline(0,1)
grid()

# outliers
d2 = data.frame(d1,predicted)
d2[c(42,48,59),]
   Price MPG.city AirBags  predicted
42  12.1       42       D  0.4736358
48  47.9       17       D 25.9369252
59  61.9       19      DP 30.0246286

# antioutliers
d2[c(11,51,80),]
   Price MPG.city AirBags predicted
11  40.1       16      DP 38.855981
51  34.3       17      DP 35.912197
80   8.4       33       0  9.014734

# MSE
anova(m2)
# Analysis of Variance Table
#                  Df Sum Sq Mean Sq F value    Pr(>F)
# MPG.city          1 3034.5 3034.49 71.5918 5.686e-13 ***
# AirBags           2 1311.2  655.59 15.4672 1.790e-06 ***
# MPG.city:AirBags  2  550.8  275.39  6.4972  0.002345 **
# Residuals        87 3687.6   42.39

b=mean((predicted-d1$Price)^2)
b
# [1] 39.65137
c = (93/87)*b
c
# [1] 42.38595
# 95% CI
#====================================================================
newval = data.frame(MPG.city = 27.5,AirBags="D")
predict(m2,newval,interval="conf")

#        fit      lwr      upr
# 1 15.24234 12.18772 18.29697

# Mean price of a 27.5 city mileage car with Driver airbag is 15242.34 USD
# Mean price of a 27.5 city mileage car with Driver airbag is in (12.18772 18.29697)

# Using fitted equation for Driver Airbag,
(25.3873 + 17.8647) +(-0.4961-0.5224)*27.5
# 15.24325

# 90% PI
#--------------------------------------------------------------
predict(m2,newval,interval="pred",level=0.90)
#      fit       lwr      upr
# 15.24234 4.120865 26.36382

# Price of a 27.5 city mileage car with Driver airbag is 15242.34 USD
# Price of a 27.5 city mileage car with Driver airbag is in (4.120865 26.36382)

# we need model with higher R2 to obtain more accurate prediction
```

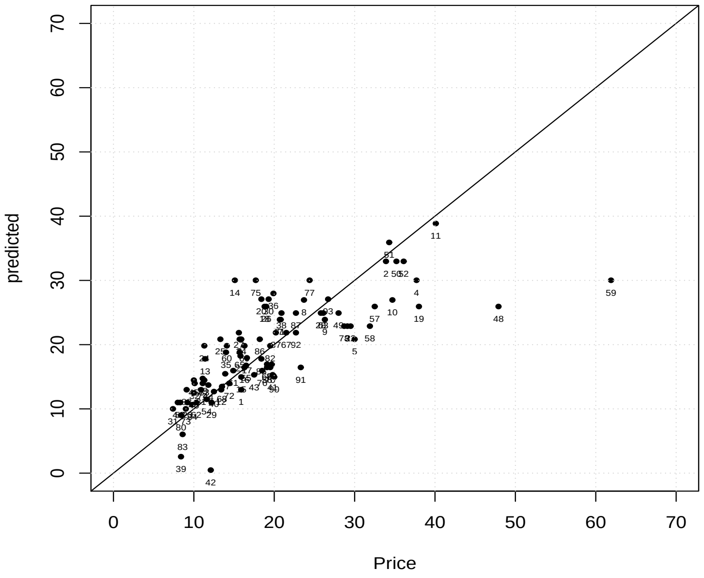

Figure 1: Predicted versus Observed prices