Example Ch8: Multiple regression in R (Testing)

R Package: GLMsData Dataset: lungcap

A study of 654 youths in East Boston investigate the relationships between lung capacity (measured by forced expiratory volume in litres (FEV)) and several factors.

Dependent variable: FEV

Independent variables:

- Smoke status: 1 = smokers; 0 = non-smokers
- age
- height
- gender

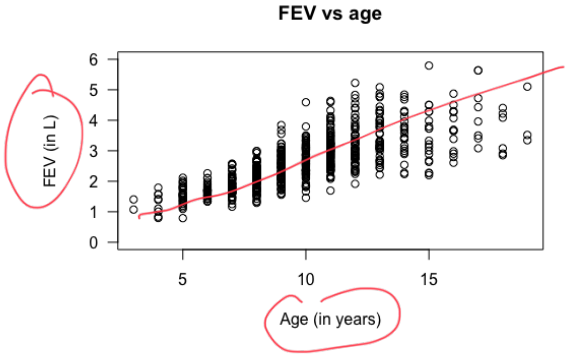
```
> library(GLMsData)
> data(lungcap)
> head(lungcap)
```

```
> head(lungcap)   #Show the first few lines of data
  Age   FEV Ht Gender Smoke
1   3 1.072 46      F     0
2   4 0.839 48      F     0
3   4 1.102 48      F     0
4   4 1.389 48      F     0
5   4 1.577 49      F     0
6   4 1.418 49      F     0
```

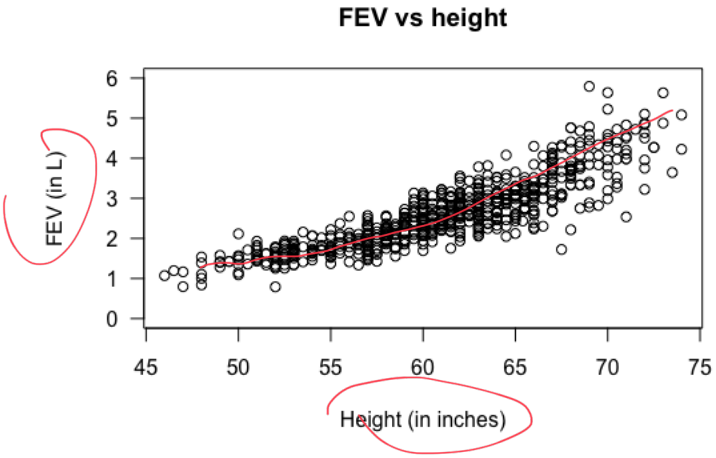
```
> summary(lungcap)
      Age              FEV              Ht        Gender       Smoke
 Min.   : 3.000   Min.   :0.791   Min.   :46.00   F:318   Min.   :0.00000
 1st Qu.: 8.000   1st Qu.:1.981   1st Qu.:57.00   M:336   1st Qu.:0.00000
 Median :10.000   Median :2.547   Median :61.50           Median :0.00000
 Mean   : 9.931   Mean   :2.637   Mean   :61.14           Mean   :0.09939
 3rd Qu.:12.000   3rd Qu.:3.119   3rd Qu.:65.50           3rd Qu.:0.00000
 Max.   :19.000   Max.   :5.793   Max.   :74.00           Max.   :1.00000
```

Plotting the data

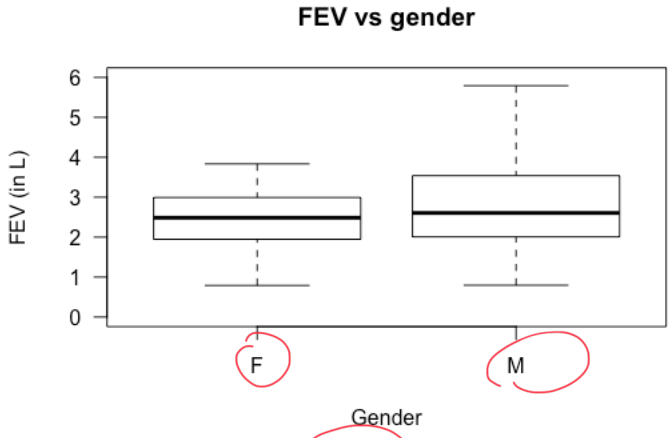
```
> plot(FEV ~ Age, data=lungcap, xlab="Age (in years)", ylab="FEV (in L)", main
= "FEV vs age", xlim=c(0,20), ylim=c(0,6), las=1)
```



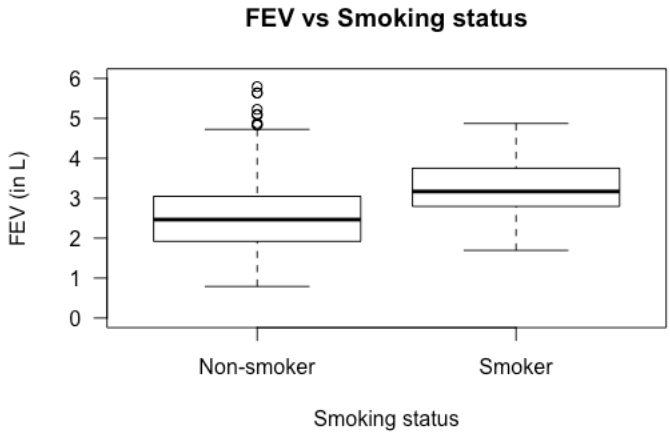
```
> plot(FEV ~ Ht, data=lungcap, main="FEV vs height", xlab="Height (in
inches)", ylab="FEV (in L)", las=1, ylim=c(0,6))
```



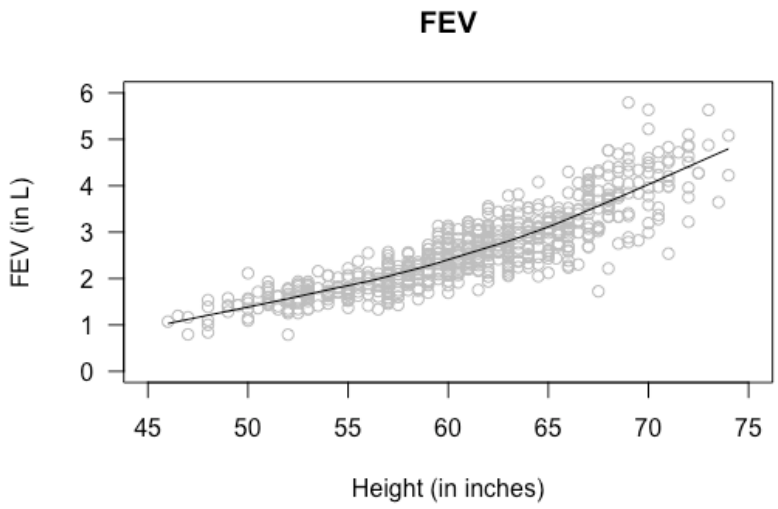

```
> plot(FEV ~ Gender, data=lungcap, main="FEV vs height", ylab="FEV (in L)",
las=1, ylim=c(0,6))
```



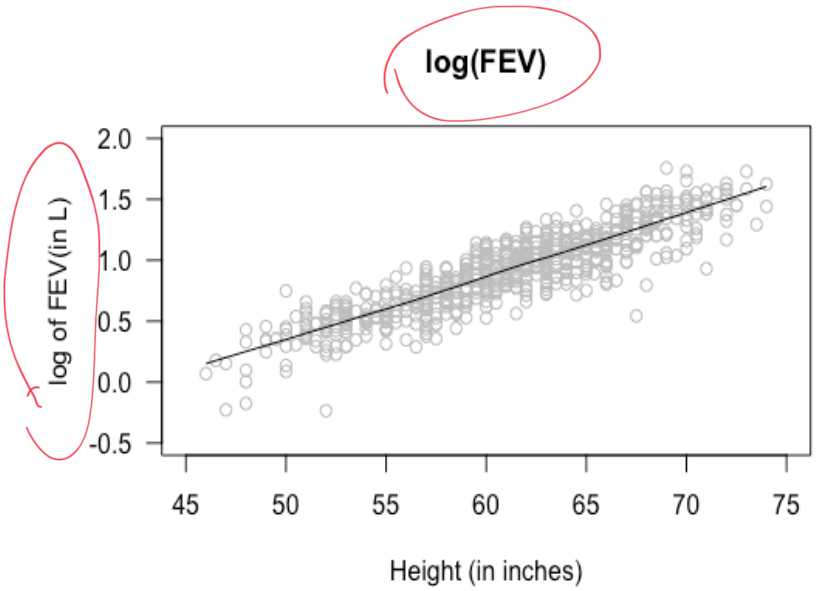

```
> lungcap$Smoke <- factor(lungcap$Smoke, levels=c(0,1), labels=c("Non-
smoker","Smoker"))  #change smoke from quantitative to factor
> plot(FEV~Smoke, data=lungcap, main ="FEV vs Smoking status", ylab = "FEV (in
L)", xlab="Smoking status", las = 1, ylim=c(0,6))
```



```
> scatter.smooth(lungcap$Ht, lungcap$FEV, las=1, col="grey",ylim=c(0,6),
xlim=c(45, 75), main="FEV", xlab="Height (in inches)", ylab="FEV)")
```




```
> scatter.smooth(lungcap$Ht, log(lungcap$FEV), las=1, col="grey",ylim=c(-
0.5,2), xlim=c(45, 75), main="log(FEV)", xlab="Height (in inches)",
ylab="FEV)")
```



Multiple regression

Model A: with independent variables Age, Ht, Gender, and Smoke (full model)

```
> reg1 <- lm(log(FEV)~Age+Ht+Gender+Smoke,data=lungcap)
> summary(reg1)

Call:
lm(formula = log(FEV) ~ Age + Ht + Gender + Smoke, data = lungcap)

Residuals:
     Min       1Q   Median       3Q      Max
-0.63278 -0.08657  0.01146  0.09540  0.40701

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.943998   0.078639 -24.721  < 2e-16 ***
Age          0.023387   0.003348   6.984  7.1e-12 ***
Ht           0.042796   0.001679  25.489  < 2e-16 ***
GenderM      0.029319   0.011719   2.502   0.0126 *
SmokeSmoker -0.046068   0.020910  -2.203   0.0279 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1455 on 649 degrees of freedom
Multiple R-squared:  0.8106,Adjusted R-squared:  0.8095
F-statistic: 694.6 on 4 and 649 DF,  p-value: < 2.2e-16

> anova(reg1)
Analysis of Variance Table

Response: log(FEV)
           Df Sum Sq Mean Sq   F value    Pr(>F)
Age         1 43.210  43.210 2041.9564 < 2.2e-16 ***
Ht          1 15.326  15.326  724.2665 < 2.2e-16 ***
Gender      1  0.153   0.153    7.2451  0.007293 **
Smoke       1  0.103   0.103    4.8537  0.027937 *
Residuals 649 13.734   0.021
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Overall utility of the model?
- What are the values of R squares and Adj R squares? Interpretation?
- Tests of usefulness of individual predictor variables?

Model B: with independent variables Age and Smoke (reduced model)

```
> reg2 <- lm(log(FEV)~Age+Smoke,data=lungcap)
> summary(reg2)

Call:
lm(formula = log(FEV) ~ Age + Smoke, data = lungcap)

Residuals:
     Min       1Q   Median       3Q      Max
-0.71124 -0.13458  0.00104  0.14909  0.60261

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.022939   0.030376   0.755  0.45041
Age          0.090768   0.003053  29.733  < 2e-16 ***
SmokeSmoker -0.089927   0.030118  -2.986  0.00293 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2108 on 651 degrees of freedom
Multiple R-squared:  0.6012,Adjusted R-squared:    0.6
F-statistic: 490.8 on 2 and 651 DF,  p-value: < 2.2e-16

> anova(reg2)
Analysis of Variance Table

Response: log(FEV)
           Df Sum Sq Mean Sq  F value    Pr(>F)
Age         1 43.210  43.210 972.6805 < 2.2e-16 ***
Smoke       1  0.396   0.396   8.9151  0.002934 **
Residuals 651 28.920   0.044
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Question: Is the full model better than the reduced model? (partial F test)

```
> anova(reg2,reg1)
Analysis of Variance Table

Model 1: log(FEV) ~ Age + Smoke
Model 2: log(FEV) ~ Age + Ht + Gender + Smoke
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    651 28.920
2    649 13.734  2    15.186 358.82 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Example Ch8: on confidence interval

R Package: GLMsData Dataset: lungcap

A study of 654 youths in East Boston investigate the relationships between lung capacity (measured by forced expiratory volume in litres (FEV)) and several factors.

Dependent variable: FEV

Independent variables:

- Smoke status: 1 = smokers; 0 = non-smokers
- age
- height
- gender

Multiple regression

Model reg1: with independent variables Age, Ht, Gender, and Smoke (full model)

```
> reg1 <- lm(log(FEV)~Age+Ht+Gender+Smoke,data=lungcap)
```

Confidence interval for betas

```
> confint(reg1, level=0.95)
                    2.5 %      97.5 %
(Intercept) -2.098414941 -1.789581413
Age          0.016812109  0.029962319
Ht           0.039498923  0.046092655
GenderM      0.006308481  0.052330236
SmokeSmoker -0.087127344 -0.005007728
```

Confidence interval for expected value of the dependent variable for given **x**

```
> predict(reg1,level=0.95,newdata=data.frame(Age=18, Ht=66, Gender="F",
Smoke="Smoker"), interval="confidence")
       fit      lwr      upr
1 1.255426 1.209268 1.301584
```

prediction interval for expected value of the dependent variable for given **x**

```
> predict(reg1,level=0.95,newdata=data.frame(Age=18, Ht=66, Gender="F",
Smoke="Smoker"), interval="prediction")
       fit      lwr      upr
1 1.255426 0.966075 1.544777
```

Confidence Ellipse for betas related to Age and Ht

```
> install.packages("ellipse")
Installing package into ‘/Users/siuhungcheung/Library/R/3.6/library’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/macosx/el-
capitan/contrib/3.6/ellipse_0.4.1.tgz'
Content type 'application/x-gzip' length 71606 bytes (69 KB)
==================================================
downloaded 69 KB


The downloaded binary packages are in
                                /var/folders/0m/hwl7r7_94f9ddhyn1sjywmg40000gn/T//
Rtmp4PkszD/downloaded_packages
> library(ellipse)

> plot(ellipse(reg1, c(2,3)), type = "l")
> points(coef(reg1)[2],coef(reg1)[3],pch=18)
> abline(v=confint(reg1)[2,],lty=2)
> abline(h=confint(reg1)[3,],lty=2)
```