

## REGRESSION ANALYSIS

### Laboratory Guide: Life Expectancy dataset

1. Consider that the average life expectancy of 38 countries is given, together with the average number of people per physician and average number of people per TV (the data is in file **LifeExp-data.txt**, just copy the data from this file to the R). Start by looking for evidences that the male (female) life expectancy in years can be explained by the average number of people per physician and the average number of people per TV. Fit two regression models (male and female) to the logarithm of the dataset.

```
head(data)
dim(data)
summary(data[,2:5])
pairs(data[,2:5])
library(psych)
pairs.panels(data[,2:5],smooth =FALSE,ellipses=FALSE,lm=TRUE)
pairs.panels(data[,2:5],smooth =FALSE,ellipses=FALSE,lm=FALSE)
var(data[,2:5])
cor(data[,2:5])

### log transformation to stabilized the variance

summary(log(data[,2:5]))
pairs(log(data[,2:5]))
pairs.panels((log(data[,2:5])),smooth =FALSE,ellipses=FALSE,lm=TRUE)
pairs.panels((log(data[,2:5])),smooth =FALSE,ellipses=FALSE,lm=FALSE)
var(log(data[,2:5]))
cor(log(data[,2:5]))

data<-as.data.frame(data)

regM<-lm(log(LifeExp.Male)~log(People.per.TV)+log(People.per.Dr),data=data)

regF<-lm((LifeExp.Female)~log(People.per.TV)+log(People.per.Dr),data=data)
```

#### Solution:

$$\log(\widehat{LifeExp.Male}) = 4.638 - 0.049\log(People.per.TV) - 0.039\log(People.per.Dr),$$
$$\log(\widehat{LifeExp.Female}) = 84.474 - 2.587\log(People.per.TV) - 1.860\log(People.per.Dr).$$

- Test the significance of regression using  $\alpha = 0.01$ . Find the p-value for this test and use it to draw your conclusions.

```
summary(regM)
summary(regF)
```

**Solution:** Let us consider the linear multiple regression model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

where  $Y$  is the  $\log(\text{LifeExp.Male})$  or the  $\log(\text{LifeExp.Female})$ ,  $x_1$  is  $\log(\text{People.per.TV})$ , and  $x_2$  is  $\log(\text{People.per.Dr})$ .

Testing the significance of regression is equivalent to the following hypothesis:

$$H_0 : \beta_1 = \beta_2 = 0 \text{ versus } H_1 : \beta_1 \neq 0 \vee \beta_2 \neq 0.$$

The p-values are  $2.682 \times 10^{-13}$  and  $2.754 \times 10^{-11}$ , for males and females. Since both p-values are smaller than 0.01, we reject the null hypothesis ( $H_0$ ) in both cases. This means that at least one of the explanatory variables ( $\log(\text{People.per.TV})$  or  $\log(\text{People.per.Dr})$ ) is important to explain life expectancy in males and females.

- Test the contribution of each variable to the model using the t-test with  $\alpha = 0.05$ . Find the p-value for these tests and use it to draw your conclusions.

```
summary(regM)
summary(regF)
```

**Solution:** Testing

$$H_0 : \beta_1 = 0 \text{ versus } H_1 : \beta_1 \neq 0,$$

leads to the p-values  $8.530 \times 10^{-6}$  and  $5.690 \times 10^{-5}$ , for males and females. Since both p-values are smaller than 0.05, we reject the null hypothesis ( $H_0$ ) in both cases.

Testing

$$H_0 : \beta_2 = 0 \text{ versus } H_1 : \beta_2 \neq 0,$$

leads to the p-values 0.00215 and 0.0135, for males and females. Since both p-values are smaller than 0.05, we reject the null hypothesis ( $H_0$ ) in both cases. This means that both explanatory variables ( $\log(\text{People.per.TV})$  and  $\log(\text{People.per.Dr})$ ) are important to explain life expectancy in males and females.

- Find the amount that the regressor *People.per.Dr* ( $x_2$ ) increases the regression sum of squares.

```
anova(regM)
anova(regF)
```

**Solution:**  $SSR(\beta_2|\beta_1) = 0.03796$  for males and  $SSR(\beta_2|\beta_1) = 84.93$  for females.

5. Use the results of part (4) to conduct an F-test for  $H_0 : \beta_2 = 0$  versus  $H_0 : \beta_2 \neq 0$  using  $\alpha = 0.05$ . What is the p-value for this test? What conclusions can you draw?

```
anova(regM)
anova(regF)
```

**Solution:** The p-values are 0.0021 and 0.0135, for males and females. Both p-values are smaller than 0.05, thus we reject the null hypothesis, at a 5% significance level.

6. Find a 99% confidence interval for  $\beta_1$  (regression coefficient associate with the variable *People.per.TV*).

```
confint(regM,level =0.99)
confint(regF,level =0.99)
```

**Solution:** For Males:

$$CI_{99\%}(\beta_1) = [-0.074, -0.023].$$

For Females:

$$CI_{99\%}(\beta_1) = [-4.126, -1.048].$$

7. Find a 95% confidence interval for the mean life expectancy for Spain. Find a 95% prediction interval for Spain

```
Spain=as.data.frame(data["Spain",])
p_conf<-predict(regM,interval="confidence",newdata=Spain,0.95)
p_pred<- predict(regM,interval="prediction",newdata=Spain,0.95)
```

**Solution:** For males,

$$CI_{95\%}(E(Y|\mathbf{x}_{Spain})) = [4.341; 4.401].$$

In the original scale, we can write:

$$CI_{95\%}(\{E(e^Y|e^{\mathbf{x}_{Spain}})\}) \simeq [76.777; 81.507].$$

$$PI_{95\%}(Y|\mathbf{x}_{Spain}) = [4.248; 4.494].$$

In the original scale, we can write:

$$PI_{95\%}(\{e^Y|e^{\mathbf{x}_{Spain}}\}) \simeq [69.946; 89.467].$$

8. What is the percentage of variability explained by the model?

```
summary(regM)
summary(regF)
```

**Solution:** For males:  $r^2 = 0.8087$  and  $r_{Adj}^2 = 0.7978$ .  
For females:  $r^2 = 0.7508$  and  $r_{Adj}^2 = 0.7365$ .

9. Construct a normal probability plot of the residuals. What conclusion can you draw from this plot? Confirm your conclusion using other graphics.

```
### Normal assumption - Residuals (e_i) Males

resM=log(data[,4])-regM$fitted.values

regM$residuals

par(mar=c(1,1,1,1))
par(mfrow=c(2,2))
plot(regM$residuals,lwd=2,ylim=c(-2.2,2.2))
title("Residuals Males")
abline(h=2,col="green")
abline(h=-2,col="green")
abline(h=0,col="black")

hist(regM$residuals)
boxplot(regM$residuals,main="Residuals Males Boxplot")

## Q-Q plot Normal

par(mfrow=c(1,2))

qqnorm(regM$residuals)
qqline(regM$residuals,col="red",lwd=2)

library(car)

qqPlot(regM$residuals,distribution="norm",envelope=FALSE,lwd=1,main="Residuals Males")

### Normal assumption - Standardized residuals (d_i=e_i/sqrt(MSE)) - Males

mseM=anova(regM)[3,3]
dM<-regM$residuals/sqrt(mseM)

par(mfrow=c(2,2))
hist(dM,prob=TRUE)
boxplot(dM)
plot(dM,lwd=2,ylim=c(-2.2,2.2))
abline(h=2,col="green")
abline(h=-2,col="green")
abline(h=0,col="black")

qqPlot(dM,distribution="norm",envelope=FALSE,lwd=1,main="Standardized Residuals Males")

### Normal assumption - Residuals Females

library(stats)
par(mar=c(1,1,1,1))
par(mfrow=c(2,2))
plot(regF$residuals,lwd=2,ylim=c(-2.2,2.2))
title("Residuals Females")
abline(h=2,col="green")
abline(h=-2,col="green")
abline(h=0,col="black")

hist(regF$residuals)
```

```

boxplot(regF$residuals,main="Residuals Females Boxplot")

## Q-Q plot Normal

par(mfrow=c(1,2))

qqnorm(regF$residuals)
qqline(regF$residuals,col="red",lwd=2)

library(car)

qqPlot(regF$residuals,distribution="norm",envelope=FALSE,lwd=1,main="Residuals Femles")

### Normal assumption - Standardized residuals ( $d_i=e_i/\sqrt{\text{MSE}}$ ) - Females

mseF=anova(regF)[3,3]
dF<-regF$residuals/sqrt(mseF)
par(mfrow=c(2,2))

hist(dF,prob=TRUE)
boxplot(dF)
plot(dF,lwd=2,ylim=c(-2.2,2.2))
abline(h=2,col="green")
abline(h=-2,col="green")
abline(h=0,col="black")

qqPlot(dF,distribution="norm",envelope=FALSE,lwd=1,main="Standardized Residuals Females")

```

**Solution:** For Males, the graphs does not question the assumption that Life Expectancy is a normal random variable.

10. Plot the residuals versus the fitted values ( $\hat{y}$ ) and versus each regressor ( $x_i$ ).

```

### Residuals Plots

library(car)
windows()
plot.new()
residualPlots(regM,quadratic=FALSE, main="Residuals Male")
dev.off()

windows()
plot.new()
residualPlots(regF,quadratic=FALSE, main="ResidualsFemale")
dev.off()

```

**Solution:** For Males/Females, the graphs does not question the model adequacy.

11. Are there any leverage and influential points in these data? Calculate and plot the h-values and the Cook's distance to answer this question.

```
#### Males regression #####
### Hat's values - leverage points
p=3
n=38
hM=hatvalues(regM)
hMlev=hM[hM>2*p/n]

### Cook's distances - influential observations

cM=cooks.distance(regM)
cMinfl=cM[cM>4/(n-p)]
cMinfl_R=cM[cM>4*mean(cM)] # R rule
### Influential plots
influenceIndexPlot(regM)
influencePlot(regM)

#### Females regression #####
### Hat's values - leverage points

hF=hatvalues(regF)
hFlev=hF[hF>2*p/n]

### Cook's distances - influential observations

cF=cooks.distance(regF)
cFinfl=cF[cF>4/(n-p)]
cFinfl_R=cF[cF>4*mean(cM)] # R rule
### Influential plots
influenceIndexPlot(regF)
influencePlot(regF)
```

**Solution for Males Regression:** For Males, the possible leverage countries are North Korea, Burma (Ethiopia and Sudan). The two countries with highest cook's distance are Sudan and North Korea, so the more possible influential countries are North Korea and Sudan.

12. The investigator suspects that women and men does not have the same mean life expectancy. Define a dummy variable for the gender and fit a multiple linear regression to these data combining the values for *LifeExp.Male* and *LifeExp.Female* to construct your new response variable.

```
### Regression model with dummy variable gender

dum1=c(rep(0,38),rep(1,38))

datal=log(data)
Dr=c(datal$People.per.Dr,datal$People.per.Dr)
TV=c(datal$People.per.TV,datal$People.per.TV)
LE=c(datal$LifeExp.Male,datal$LifeExp.Female)

dataAll=cbind(LE,TV,Dr,dum1)

regAll=lm(LE~TV+Dr+dum1)
```

```
summary(regAll)
anova(regAll)
```

**Solution:**

$\log(\widehat{LifeExp.Male}) = 4.597 - 0.0449\log(People.per.TV) - 0.0348\log(People.per.Dr)$ ,  
 $\log(\widehat{LifeExp.Female}) = 4.524 - 0.0449\log(People.per.TV) - 0.0348\log(People.per.Dr)$ ,  
where  $4.524 = 4.597 - 0.0729$ . The decreasing in the estimated mean value for  $\log(LifeExp.Female)$  comparing to the estimated mean value for  $\log(LifeExp.Male)$  is 0.0729.

Adjusting a regression model with gender as a dummy variable has the advantage to model life expectancy in a more general way, estimating the effect of gender in the mean life expectancy.