Practical on linear regression

Daniela Zugna

These are data from a study of lung function among 654 children aged 3 and 19 years. Expiratory volume in 1 second was measured by spirometer, and both age and height of the children were recorded. The aim of this practical is to evaluate how expiratory volume changes on average according to other measured variables.

Variables:
id: identity number
age: numeric
fev:expiratory volume in 1 second (l)
height: numeric (inches)
sex: string ("male", "female")
smoke: string ("current smoker", "non-current smoker")

1. Open the the fev.csv dataset.

```
rm(list=ls())
data<-read.table("",header=T, sep="")
dim(data)
names(data)
head(data)
```

2. Get basic descriptives for the entire data set.

```
summary(data)
str(data)
```

   - Comment the nature and the distribution of each variable.

3. Plot an histogram of fever with an overlapping normal density curve.

```
hist(data$fev,c(0,1,2,3,4,5,6),freq=F,main="Histogram with normal density",
xlab="Fever",ylim=c(0,0.6))
m<-mean(data$fev)
std<-sqrt(var(data$fev))
```

```
x<-seq(min(data$fev), max(data$fev), length = 40)
y<-dnorm(x, mean=m, sd=std)
lines(x,y,col="red", lwd=2)
```

4. Graph a scatter plot of fever *vs* age and and include a linear fit.

```
plot(data$age,data$fev,xlab="Age",ylab="Fever")
abline(lm(data$fev~data$age))
```

5. Fit a linear regression model for fever *vs* age and examine the regression coefficients. Compute the 95% confidence intervals for the regression coefficients.

```
mylinear.age<-lm(fev~age,data=data)
summary(mylinear.age)
```

   - On the basis of the estimated values of the model coefficients, how do you interpret the intercept?
   - On the basis of the estimated values of the model coefficients, how does fever vary for a unit change of age?
   - On the basis of the 95% confidence interval for the regression coefficient of age, can you conclude that fever depends linearly on age? Why?

6. After generating a variable indicating the age centered at its mean value (cage) re-fit a linear regression model for fever *vs* cage and examine the regression coefficients.

```
data$cage<-data$age-mean(data$age)
mylinear.age<-lm(fev~cage,data=data)
summary(mylinear.age)
confint(mylinear.age)
```

   - On the basis of the estimated values of the model coefficients, how do you interpret the intercept?
   - On the basis of the estimated values of the model coefficients, how does fever vary for a unit change of age?
   - On the basis of the 95% confidence interval for the regression coefficient of age, can you conclude that fever depends linearly on age? Why?

7. Graph a scatter plot of fever *vs* height and and include a linear fit.

```
plot(data$height,data$fev,xlab="Height",ylab="Fever")
abline(lm(data$fev~data$height))
```

8. After generating a variable indicating the height centered at its mean value (cheight) fit a linear regression model for fever *vs* cheight and examine the regression coefficients.

```
data$cheight<-data$height-mean(data$height)
mylinear.height<-lm(fev~cheight,data=data)
summary(mylinear.height)
confint(mylinear.height)
```

   - On the basis of the estimated values of the model coefficients, how do you interpret the intercept?
   - On the basis of the estimated values of the model coefficients, how does fever vary for a unit change of height?
   - On the basis of the 95% confidence interval for the regression coefficient of age, can you conclude that fever depends linearly on height? Why?

9. Fit a linear regression model for fever *vs* sex and examine the regression coefficients. Compute the 95% confidence intervals for the regression coefficients.

```
mylinear.sex<-lm(fev~sex,data=data)
summary(mylinear.sex)
confint(mylinear.sex)
```

   - On the basis of the estimated values of the model coefficients, how do you interpret the intercept?
   - On the basis of the estimated values of the model coefficients, how does fever vary according to sex?
   - On the basis of the 95% confidence interval for the regression coefficient of sex, can you conclude that fever depends on sex? Why?

10. Fit a linear regression model for fever *vs* smoke and examine the regression coefficients. Compute the 95% confidence intervals for the regression coefficients.

```
mylinear.smoke<-lm(fev~smoke,data=data)
summary(mylinear.smoke)
confint(mylinear.smoke)
```

   - On the basis of the estimated values of the model coefficients, how do you interpret the intercept?
   - On the basis of the estimated values of the model coefficients, how does fever vary according to smoke?

- On the basis of the 95% confidence interval for the regression coefficient of smoke, can you conclude that fever depends on smoke? Why?

11. Fit a linear regression model for fever with all the variables and examine the regression coefficients. Compute the 95% confidence intervals for the regression coefficients. Describe which conclusions you would draw from this study (which variables affect fever, and which ones might be the most relevant).

```
mylinear<-lm(fev~cage+cheight+sex+smoke,data=data)
summary(mylinear)
confint(mylinear)
```

12. Perform the residual's analysis, by checking they are normally distributed.

```
qqnorm(residuals(mylinear))
shapiro.test(residuals(mylinear))
```

- What would you expect from qqplot if the normality assumption were satisfied? What is your conclusion looking at the plot?
- Is there evidence of normality by Shapiro test? Why?

13. Predict the average fever with its confidence interval for four subjects with the following characteristics:
    1. age=mean of age, height=mean of height, female and no smoker
    2. age=mean of age, height=mean of height, male and no smoker
    3. age=mean of age, height=mean of height, female and smoker
    4. age=mean of age, height=mean of height, male and smoker

```
newdata1<- with(data, data.frame(age=mean(age),height=mean(height),
sex=0:1,smoke=0:1))
newdata1
newdata2 <- cbind(newdata1, predict(mylinear, newdata = newdata1,se.fit = TRUE))
newdata2
pred.w.clim <- predict(lm(y ~ x), newdata1, interval = "confidence")
```