Demo 2.1

Nguyen Xuan Binh

11/12/2021

2.1 Continuation of the homework.

a) Generate a scatter plot (CONSUMPTION, ILL). Add the estimated regression line to the figure.

Scatter plot (Figure 1):

```
smoking <- read.table("tobacco.txt",header=T,sep="\t")
model <- lm(ILL~CONSUMPTION,data=smoking)
countries <- c("Iceland","Norway","Sweden","Canada","Denmark",
"Austria","USA","Netherlands","Switzerland","Finland",
"England")

plot(smoking$CONSUMPTION,smoking$ILL, ylab="Cases in 1950",
xlab="CONSUMPTION in 1930", pch=16,
main="CONSUMPTION/ILL per 100 000 individuals")
abline(model,col="red")
text(smoking$CONSUMPTION, smoking$ILL, labels=countries, cex= 0.8,
pos=3)
```

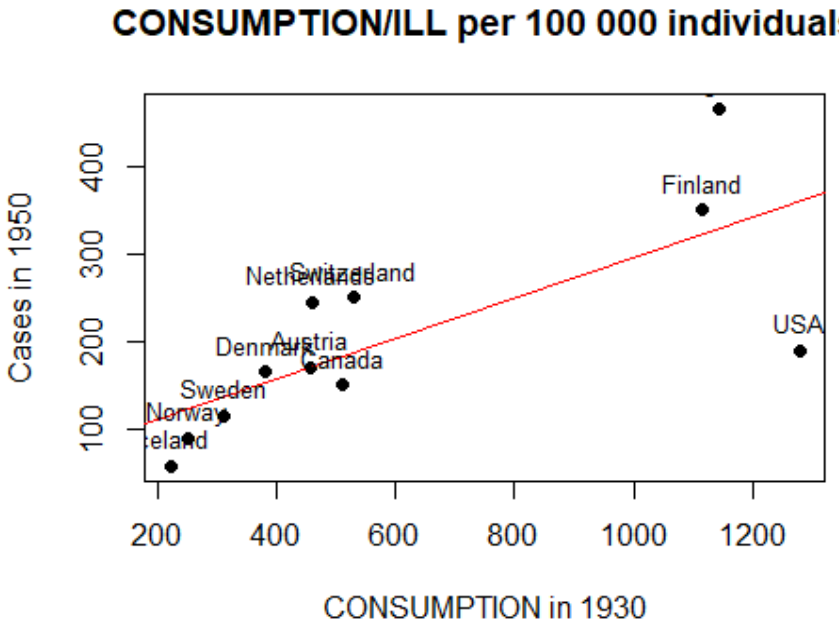

Figure 1: Scatter plot of CONSUMPTION and ILL.

Alternatively, you can use the function identify to label the observations

b) Determine the fitted values y^ and estimated residuals e from the corresponding model and assign them to variables FIT and RES, respectively.

The fitted values and the estimated residuals correspond to fitted.values and residuals from the estimated model, and they can be accessed by

```
FIT <- model$fit
RES <- model$res
FIT

##        1        2        3        4        5        6        7        8
## 116.1542 123.0277 136.7746 182.5977 152.8127 169.9963 359.0165 171.1419
##        9       10       11
## 187.1800 321.2125 328.0859

RES

##            1            2            3            4            5
6
##  -58.1542313  -33.0276915  -21.7746117  -32.5976793   12.1873146    0.0036
643
##            7            8            9           10           11
## -169.0164893   73.8580876   62.8200140   28.7875414  136.9140812
```

c) Generate scatter plots (ILL, FIT) and (FIT, RES).

Scatter plot (observed values, fitted values) (Figure 2). Plot the fitted values y^i against the observed values of the variable ILL.

```
plot(smoking$ILL,FIT, ylab="Fits",xlab="Sick",pch=16)
text(smoking$ILL,FIT, labels = ifelse(rownames(smoking)=="7",
countries, NA),pos=2)
```

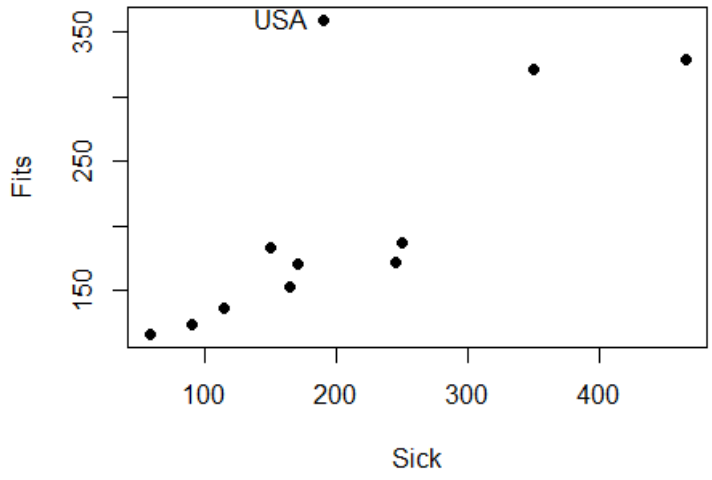

Figure 2: Scatter plot of the observed values and the fitted values.

The scatter plot illustrates the goodness of the model:
• The closer the points (yi; y^i); i = 1, 2,…,n are to the line with slope of 1, the better the model is.
• Outliers are usually visible. Note that, the squared Pearson correlation coefficient given by the points (yi; y^i); i = 1, 2,…, n is equal to the coefficient of determination:

[Cor(y; y^)]2 = R2

Scatter plot (fitted values, residuals) (Figure 3). Plot the residuals ei against the fitted values y^i.

```
plot(FIT,RES, xlab="Fits",ylab="Residuals",pch=16)
text(FIT,RES, labels = ifelse(rownames(smoking)=="7",
countries, NA),pos=3)
```

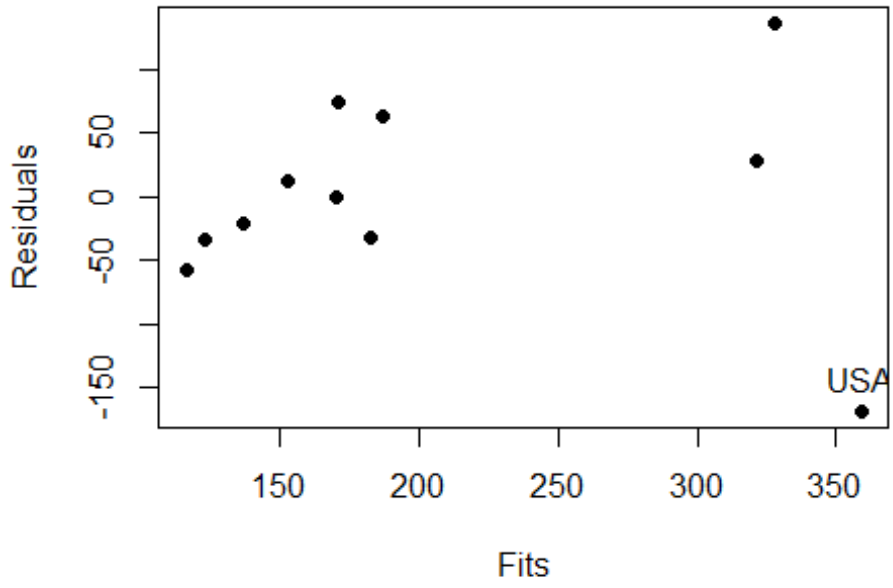

Figure 3: Scatter plot of the fitted values and the residuals.

The scatter plot illustrates the goodness of the model:

• The closer the points (^ yi; ei); i = 1, 2,…,n are to the line e = 0, the better the model is.

• Outliers are usually visible.

d) Study whether the observation 7=USA is an outlier by using the plots of part (c).

Especially, by the scatter plot (FIT,RES), the observation 7=USA looks like an outlier.

e) Study whether the observation 7=USA is an outlier by using Cook's distances.

Assign the Cook's distances to cooksd and plot the distances. See Figure 4.

```
cooksd <- cooks.distance(model)
x <-plot(cooksd,xaxt="n",xlab=" ",ylab="Cook's distances")
axis(side=1,at=1:11, labels=countries,las=2 )
```

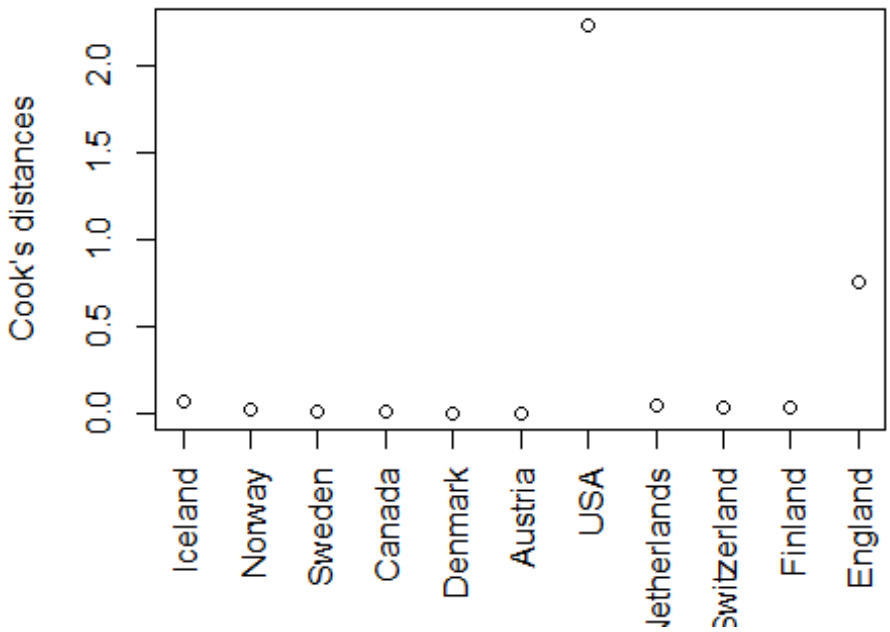

Figure 4: Cook's distances of the model.

f) Estimate the model without the observation USA. Compare the results with the homework assignment of the previous week.

Estimate the model without the observation 7=USA.

```
smoking2 <- smoking[-7,]
model2 <- lm(ILL~CONSUMPTION,data=smoking2)
summary(model2)

##
## Call:
## lm(formula = ILL ~ CONSUMPTION, data = smoking2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.353 -28.923  -7.861  35.321  66.919
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.55343   28.26713   0.479    0.644
## CONSUMPTION  0.35767    0.04547   7.867 4.93e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44.92 on 8 degrees of freedom
```

```
## Multiple R-squared:  0.8855, Adjusted R-squared:  0.8712
## F-statistic: 61.88 on 1 and 8 DF,  p-value: 4.928e-05
```

Compared to the first homework assignment, the estimate for the slope has increased from 0.23 to 0.36. This implies a stronger linear dependence between lung cancer cases and consumption of cigarettes among the remaining observations (countries).

Question: Can we remove the observation 7=USA? Answer: During the corresponding time period, tobacco was milder in the USA, when compared to the other countries of the study. Furthermore, the cigarettes sold in the USA had filters, whereas the cigarettes sold in the other countries did not have filters.

As we have found a contextual explanation, the observation USA can be regarded as an outlier and its removal from the data is justified. Remember that disregarding data without valid explanations is not allowed!