homework2stat632

Jessica Grover

2/18/2022

library(alr4)

## Loading required package: car

## Loading required package: carData

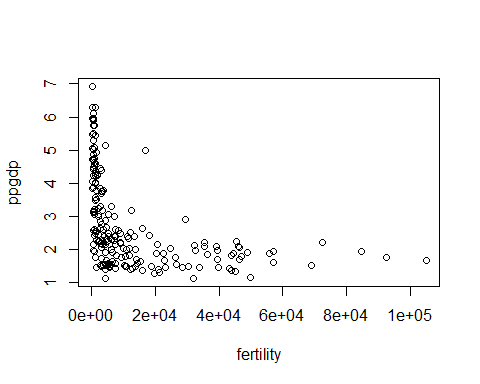
## Loading required package: effects

## lattice theme set by effectsTheme()  
## See ?effectsTheme for details.

UN11 <- read.csv("UN11.csv")  
dataset=alr4::UN11

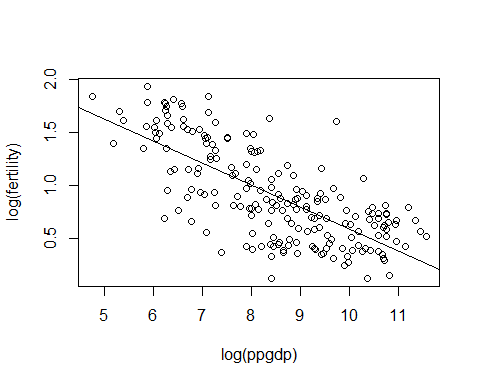
#A. Make a scatterplot with fertility on the y-axis and ppgdp on the x-axis. Explain why we should consider log transformations for this data.

plot(dataset$ppgdp, dataset$fertility,  
xlab = 'fertility' , ylab = 'ppgdp')

 Basically,log transformation will make the data linearly distributed. Here we can see it not linearly distributed.

#B. Make a scatterplot of log(fertility) versus log(ppgdp) and add the least squares regression line. Does the association appear to be reasonably linear?

lm1 <- lm(log(fertility) ~ log(ppgdp), data=UN11)  
plot(log(fertility) ~ log(ppgdp), data=UN11)  
abline(lm1)

 Yes, the association appear to be reasonably linear.

#C. Use the lm() function to fit a simple linear regression model with log(fertility) as the response variable, and log(ppgdp) as the explanatory variable. Use the summary() function to print the results.

model= lm(log(fertility)~log(ppgdp), data=UN11)  
summary(model)

##   
## Call:  
## lm(formula = log(fertility) ~ log(ppgdp), data = UN11)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.79828 -0.21639 0.02669 0.23424 0.95596   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.66551 0.12057 22.11 <2e-16 \*\*\*  
## log(ppgdp) -0.20715 0.01401 -14.79 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3071 on 197 degrees of freedom  
## Multiple R-squared: 0.526, Adjusted R-squared: 0.5236   
## F-statistic: 218.6 on 1 and 197 DF, p-value: < 2.2e-16

#D. Write down the equation for the least squares line.

model

##   
## Call:  
## lm(formula = log(fertility) ~ log(ppgdp), data = UN11)  
##   
## Coefficients:  
## (Intercept) log(ppgdp)   
## 2.6655 -0.2071

The equation for least square line is log(fertility) = 2.6655 + log(-0.2071)

#E. Interpret the slope of the regression model. A decrease of 20.71% is seen in fertility rate-number of children per women due to 1% increase of ppgdp- the gross national product per person in US dollar.

#F. For a locality not in the data with ppgdp = 1000, obtain a point prediction and a 95% prediction interval for log(fertility). If the interval (a, b) is a 95% prediction interval for log(fertility), then a 95% prediction interval for fertility is given by (exp(a), exp(b)). Use this results to get a 95% prediction interval for fertility.

new\_x<-data.frame(ppgdp=1000)  
predict1<-predict(lm1,newdata = new\_x,interval = "predict", level=0.95)  
predict1

## fit lwr upr  
## 1 1.234567 0.6258791 1.843256

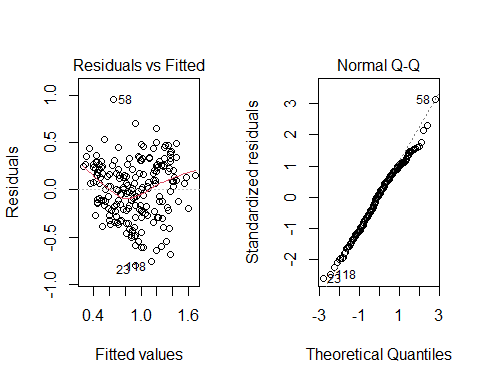
Exponentiating to express in the original scale of fertility

exp(predict1)

## fit lwr upr  
## 1 3.436891 1.869889 6.31707

#G. Make a plot of the standardized residuals versus fitted values, and a QQ plot of the standardized residuals. Comment on whether or not the assumptions for simple linear regression appear to be satisfied.

par(mfrow = c(1, 2))  
plot(lm1, 1:2)

 Here, the residual plot shows nonconstant variance. It means, the variability in the residuals tends to increase with the fitted values. The QQ plot also indicates that the residuals deviate from a normal distribution.

#H. Which countries are flagged as outliers? That is, which countries have standardized residuals outside the interval from -2 to 2. In your view, does it seem necessary to remove any of these points, and then refit the model?

The countries Guinea and Bosnia Herzegovina appear to be outside the interval from -2 to 2. These are the equatorial countries. It seems that there is no need to remove any of the points as n is large. Removing the points won’t make any difference.