MGEC11 Assignment 2

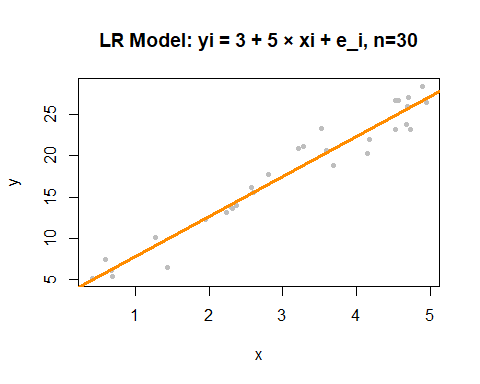
Xinrui Wang,1004741078

2021/6/24

## Q1

1. yi = 3 + 5 × xi + epsilon\_i where xi is U(0, 5)

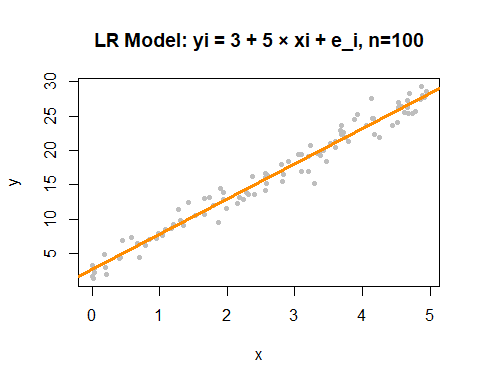
sim\_1 = function(sample\_size = 30) {  
 x = runif(n=sample\_size)\*5  
 y = 3 + 5 \* x + rnorm(n = sample\_size, mean = 0, sd = sqrt(2))  
 data.frame(x, y)  
}  
  
set.seed(42)  
sim\_data\_1 = sim\_1()  
plot(y ~ x, data = sim\_data\_1, col = "grey", pch = 20,  
 main = "LR Model: yi = 3 + 5 × xi + e\_i, n=30")  
fit\_1 = lm(y ~ x, data = sim\_data\_1)  
abline(fit\_1, col = "darkorange", lwd = 3)



b) sim for n=100,150,500

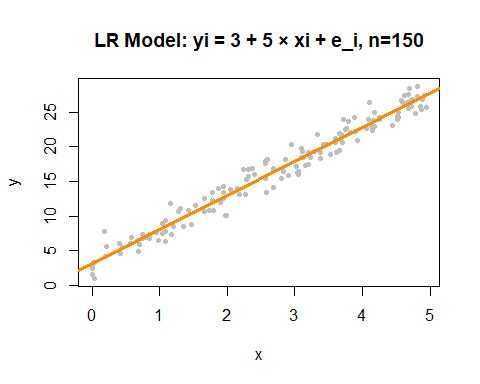
sim\_2 = function(sample\_size = 100) {  
 x = runif(n=sample\_size)\*5  
 y = 3 + 5 \* x + rnorm(n = sample\_size, mean = 0, sd = sqrt(2))  
 data.frame(x, y)  
}  
  
set.seed(42)  
sim\_data\_2 = sim\_2()

plot(y ~ x, data = sim\_data\_2, col = "grey", pch = 20,  
 main = "LR Model: yi = 3 + 5 × xi + e\_i, n=100")  
fit\_2 = lm(y ~ x, data = sim\_data\_2)  
abline(fit\_2, col = "darkorange", lwd = 3)



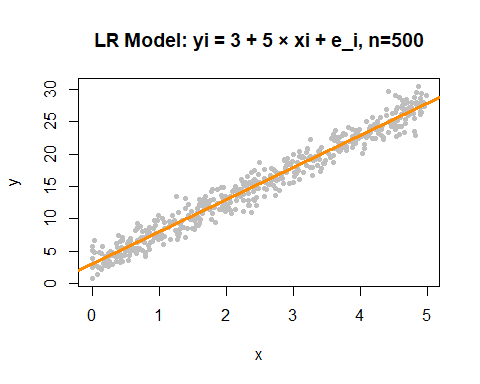
sim\_3 = function(sample\_size = 150) {  
 x = runif(n=sample\_size)\*5  
 y = 3 + 5 \* x + rnorm(n = sample\_size, mean = 0, sd = sqrt(2))  
 data.frame(x, y)  
}  
  
set.seed(42)  
sim\_data\_3 = sim\_3()

plot(y ~ x, data = sim\_data\_3, col = "grey", pch = 20,  
 main = "LR Model: yi = 3 + 5 × xi + e\_i, n=150")  
fit\_3 = lm(y ~ x, data = sim\_data\_3)  
abline(fit\_3, col = "darkorange", lwd = 3)



sim\_4 = function(sample\_size = 500) {  
 x = runif(n=sample\_size)\*5  
 y = 3 + 5 \* x + rnorm(n = sample\_size, mean = 0, sd = sqrt(2))  
 data.frame(x, y)  
}  
  
set.seed(42)  
sim\_data\_4 = sim\_4()

plot(y ~ x, data = sim\_data\_4, col = "grey", pch = 20,  
 main = "LR Model: yi = 3 + 5 × xi + e\_i, n=500")  
fit\_4 = lm(y ~ x, data = sim\_data\_4)  
abline(fit\_4, col = "darkorange", lwd = 3)



get coef:

coef(fit\_1)

## (Intercept) x   
## 2.938828 4.861178

coef(fit\_2)

## (Intercept) x   
## 2.641063 5.123244

coef(fit\_3)

## (Intercept) x   
## 3.109669 4.921563

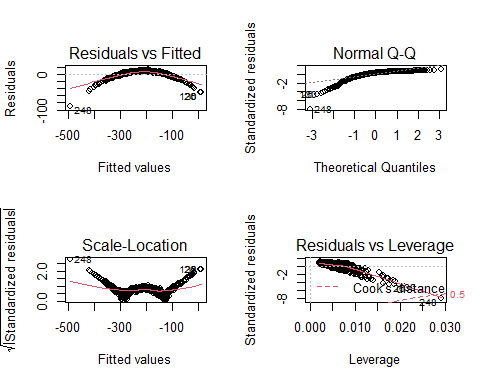
coef(fit\_4)

## (Intercept) x   
## 2.961743 4.986785

The true value is intercept = 3 and slope = 5. From the results, as sample size n increases, the estimated slope becomes more accurate, ie. it is closer and closer to the true slope 5.

## Q2

data<-read.csv("C:/Users/69544/Desktop/MGEC11/A2/data1.csv", header = T)  
model<-lm(y~cbind(x1,x2),data = data)  
par(mfrow = c(2, 2))  
plot(model)



summary(model)

##   
## Call:  
## lm(formula = y ~ cbind(x1, x2), data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -89.128 -3.086 3.990 7.095 12.131   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 202.1666 2.6892 75.176 < 2e-16 \*\*\*  
## cbind(x1, x2)x1 1.0059 0.3522 2.856 0.00447 \*\*   
## cbind(x1, x2)x2 -40.7289 0.2516 -161.866 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11.31 on 497 degrees of freedom  
## Multiple R-squared: 0.9814, Adjusted R-squared: 0.9813   
## F-statistic: 1.312e+04 on 2 and 497 DF, p-value: < 2.2e-16

From the residuals vs fitted figure, we don’t have a horizontal line, the linearity is violated. also, we can see there is a pattern in the residuals, they first increase and then decrease. Thus, the model is not fit because the residuals do not distribute evenly. Considering the normal Q-Q plot, we get heavy tails and residuals points don’t follow the line very well, so normality is violated. Since the residuals moves from far to close to the linear model and move from close to far, we guess the model should be y=√x or y=lnx whose derivative becomes smaller when x becomes larger. Therefore, the model seems not appropriate.

model2<- lm(y~log10(x1)+log10(x2),data=data)  
par(mfrow = c(2, 2))  
plot(model2)

Diagram, schematic

Description automatically generated

library(lmtest)

bptest(model)

##   
## studentized Breusch-Pagan test  
##   
## data: model  
## BP = 3.1522, df = 2, p-value = 0.2068

bptest(model2)

##   
## studentized Breusch-Pagan test  
##   
## data: model2  
## BP = 1.1333, df = 2, p-value = 0.5674

we see a large p-value, so we do not reject the null of homoscedasticity.

## Q3

文本, 信件

描述已自动生成

文本, 信件

描述已自动生成

文本, 信件

描述已自动生成