Lab 5 - Computational Statistics (732A89)

Helena Llorens Lluís (hllor282), Yi Yang (yiyan338)

2025-02-20

QUESTION 1 Bootstrap for regression

First, we use `lm()` to fit a cubic regression model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$$

,where x represnts the *concentration* of the fertilizer(%) and y represents the *yield*(mg).

```
##
## Call:
## lm(formula = Yield ~ poly(Fertilizer, 3, raw = TRUE), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.893 -17.142  -3.893  17.716  58.107
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       199.284      5.442  36.616   <2e-16 ***
## poly(Fertilizer, 3, raw = TRUE)1   62.634     72.165   0.868   0.3881
## poly(Fertilizer, 3, raw = TRUE)2 -298.766    165.952  -1.800   0.0757 .
## poly(Fertilizer, 3, raw = TRUE)3  158.664     91.587   1.732   0.0872 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.47 on 77 degrees of freedom
## Multiple R-squared:  0.646,  Adjusted R-squared:  0.6322
## F-statistic: 46.84 on 3 and 77 DF,  p-value: < 2.2e-16
```

To reduce model complexity, we remove the cubic term and fit a *quardic model*

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

We then estimate the coefficients with their 95% confidence intervals. The regression curve plot and confidence intervals for the coefficients are shown below:

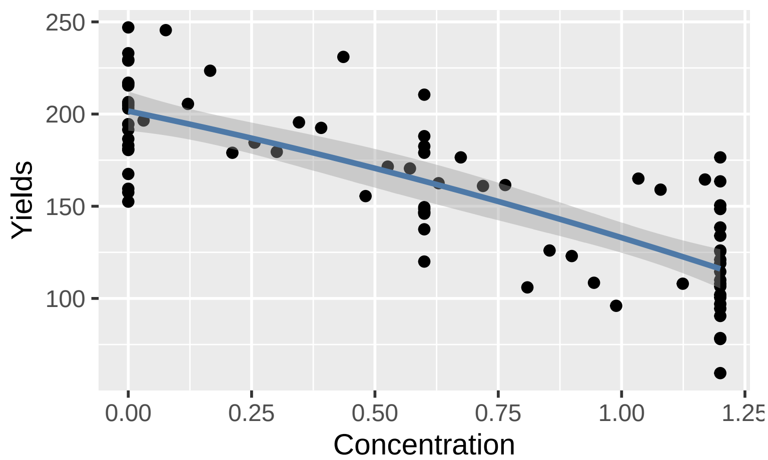

Figure 1: Yields vs Concentration

Table 1: 95% Confidence Interval of parameters

	2.5 %	97.5 %
(Intercept)	190.97432	212.242998
poly(Fertilizer, 2, raw = TRUE)1	-103.16614	-7.791748
poly(Fertilizer, 2, raw = TRUE)2	-51.57084	25.228592

We now use the **bootstrap method** with 10,000 replicates to derive a 95% confidence interval for β_1 using percentile method.The histogram and the 95% confidence interval are shown below:

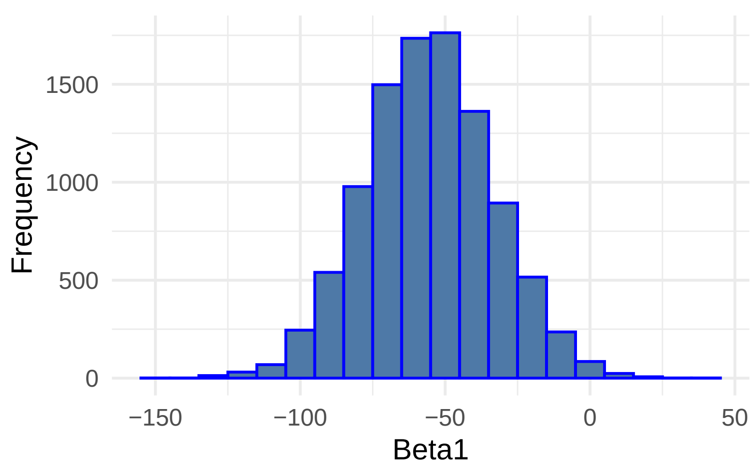

Figure 2: Bootstrap Coefficient Distribution

```
## 95% CI for beta 1 : -98.47781 -11.3802
```

Now we use the package `boot` to compute a 95% confidence interval for β_1 using both percentile method and BCa(bias-corrected and accelerated) method.The table below summarizes the 95% confidence intervals obtained using different methods:

Table 2: 95% Confidence Interval of different methods

	Lower Bound	Upper Bound	Interval Width
lm	-103.16614	-7.791748	95.3744
bootstrap	-98.47781	-11.380203	87.0976
boot_percentile	-99.46221	-12.067414	87.3948
boot_BCa	-98.38221	-10.665462	87.7167

Discussion of Confidence Interval Differences

From the results, we observe the following:

1.**The confidence interval(CI) using `lm()`is wider than the bootstrap CIs.** This is because `lm()` assumes that data follows a normal distribution.If the data sample size is small or the data is not normally distributed this assumption may not hold, leading to a less accurate CI.

2.**The bootstrap method produces narrower confidence interval without distributional assumption.** This allows it to better adapt to the actual data distribution.

3.**The CIs obtained using the manual bootstrap method and the `boot` package with the `percentile` method are similar.** However, the CI from the `boot` package with the `BCa` method is slightly different. This difference may be due to the BCa method correcting for bias, suggesting that the data may have some skewness.

Appendix

Question 1

```r
library(ggplot2)
library(boot)
data <- read.csv("kresseertrag.dat",header=FALSE,sep = "")
colnames(data) <- c("Number","Fertilizer","Yield")
data <- as.data.frame(data)

#a
modelA <- lm(Yield~poly(Fertilizer,3,raw = TRUE),data = data)
summary(modelA)

confint(modelA,level = 0.95)



#b
#remove a term
modelB <- lm(Yield~poly(Fertilizer,2,raw = TRUE),data = data)
summary(modelB)

#coefficients with 95% confidence interval
confint(modelB,level = 0.95)

#plot
#plot(data$Fertilizer,data$Yield,main="Yields vs Concentration",
 #    xlab = "Concentration",ylab = "Yields")
```

```r
#lines(data$Fertilizer,fitted(modelB),col="blue")

ggplot(data,aes(x=Fertilizer,y=Yield))+
  geom_point()+
  labs(title = "Yields vs Concentration",x="Concentration",y="Yields")+
  geom_line(aes(y=fitted(modelB)),color="#4E79A7")
  #geom_line(aes(y=fitted(modelA)),color="red")

#c bootstrap for beta
bo <- 10000                         #bootstrasp replicates
bs <- c()
set.seed(12345)
#save the results
for (i in 1:bo) {
  #sampling using indcies
  indices <- sample(1:nrow(data),size=nrow(data),replace = TRUE)
  bootstrapData <- data[indices,]
  model_bootstrap <- lm(Yield~poly(Fertilizer,2,raw = TRUE),data = bootstrapData)
  bs <- c(bs,coef(model_bootstrap)[2])
}
hist(bs)
bss <- sort(bs)
ci95 <- c(bss[round(bo*0.25)],bss[round(bo*0.975)])
ci95


#d bootstrap using boot package
beta1 <- function(data,i){
  model <- lm(Yield~poly(Fertilizer,2,raw = TRUE),subset=i,data = data)
  coef(model)[2]
}

cb <- boot(data,beta1,R=10000)
perc <- boot.ci(cb,type = "perc")
bca <- boot.ci(cb,type = "bca")
perc

 result <- c(CI[2,],95.3744,ci95,87.0976,perc$percent[4],perc$percent[5],87.3948,bca$bca[4],bca$bca[5],
 result_mt <- matrix(result,byrow = TRUE,ncol=3)
 rownames(result_mt) <- c("lm","bootstrap","boot_percentile","boot_BCa")
 colnames(result_mt) <- c("Lower Bound","Upper Bound","Interval Width")
 kable(result_mt,caption = "95% Confidence Interval of different methods")
```