# 8)

## A)

data(Auto)

reg1 = lm(mpg ~ horsepower, data = Auto)

summary(reg1)

# I - Existe

# II - 60% da variabilidade de mpg é explicada por horsepower

# III - A relação é negativa

# IV -

predict(reg1, data.frame(horsepower = 98), interval = "confidence")

predict(reg1, data.frame(horsepower = 98), interval = "prediction")

## B

plot(Auto$horsepower, Auto$mpg, main = "mpg vs horsepower", xlab = "horsepower", ylab = "mpg", col = "green")

abline(reg1, col = "blue")

## C

par(mfrow = c(2, 2))

plot(reg1)

# Residuals vs Fitted: Os dados não são lineares.

# Residuals vs Leverage: Existem alguns outliers.

# 9)

## A)

pairs(Auto)

## B)

names(Auto)

cor(Auto[1:8])

## C)

reg1 = lm(mpg ~ . - name, data=Auto)

summary(reg1)

# I - Existe

# II - Todos as variáveis são significativas exceto cylinders, horsepower e acceleration

# III - Sugere que a cada ano a mais, espera-se um aumento médio de 0,75 em mpg

## D)

par(mfrow = c(2, 2))

plot(reg1)

# Residuals vs Fitted: Os dados não são lineares.

# Residuals vs Leverage: Existem alguns outliers.

# 10)

## A)

data(Carseats)

reg2 = lm(Sales ~ Price + Urban + US, data=Carseats)

summary(reg2)

## B)

# Price - A cada 1 dólar a mais, espera-se uma diminuição média de 54 unidades em vendas, controlando pelas outras variáveis.

# Urban - Em média, as vendas unitárias na área urbana são 21,91 unidades menores que na área rural.

# US - Em média, as vendas unitárias em uma loja nos EUA são 1200.57 unidades a mais que uma loja que não seja nos EUA.

## C)

# Vendas = 13,0434689 + (- 0,0544588) x Preço + (- 0,0219162) x Urbano + (1,2005727) × US + ε

# Urban = 1 se a loja estiver em um local urbano e 0 se não for, e US = 1 se a loja estiver nos EUA e 0 se estiver fora

## D)

# Price e US.

## E)

reg3 = lm(Sales ~ Price + US, data=Carseats)

summary(reg3)

## F)

# O R2 do modelo menor é melhor que o do modelo maior.

## G)

confint(reg3)

## H)

par(mfrow = c(2, 2))

plot(reg3)

# Existem alguns outliers e alguns pontos de alavancagem.

# 13)

## A)

set.seed(1)

x = rnorm(100)

## B)

eps = rnorm(100, sd = sqrt(0.25))

## C)

y = -1 + 0.5 \* x + eps

length(y)

# β0 = -1 E β1 = 0,5

## D)

plot(x, y)

# A relação entre x e y é linear.

## E)

reg4 = lm(y ~ x)

summary(reg4)

# Devido a estatística F e o valor de p próximo a zero, a hipótese nula pode ser rejeitada.

## F)

plot(x, y)

abline(reg4, col = "red")

abline(-1, 0.5, col = "blue")

legend("topleft", c("Least square", "Regression"), col = c("red", "blue"), lty = c(1, 1))

## G)

reg5 = lm(y ~ x + I(x^2))

summary(reg5)

# Não há evidências de que o termo quadrático melhore o ajuste do modelo.

## H)

set.seed(1)

eps = rnorm(100, sd = 0.125)

x = rnorm(100)

y = -1 + 0.5 \* x + eps

plot(x, y)

reg6 = lm(y ~ x)

summary(reg6)

abline(reg6, col = "red")

abline(-1, 0.5, col = "blue")

legend("topleft", c("Least square", "Regression"), col = c("red", "blue"), lty = c(1, 1))

# Diminuindo o ruído, a variância da distribuição normal foi reduzida.

# O R2 está mais alto e o RSE mais baixo.

## I)

confint(reg4)

confint(reg6)

# Quando o ruído aumenta, os intervalos de confiança aumentam. Quando o ruído diminui, há mais previsibilidade no conjunto de dados.

# 15)

## A)

library(MASS)

attach(Boston)

fit.zn <- lm(crim ~ zn)

summary(fit.zn)

fit.indus <- lm(crim ~ indus)

summary(fit.indus)

chas <- as.factor(chas)

fit.chas <- lm(crim ~ chas)

summary(fit.chas)

fit.nox <- lm(crim ~ nox)

summary(fit.nox)

fit.rm <- lm(crim ~ rm)

summary(fit.rm)

fit.age <- lm(crim ~ age)

summary(fit.age)

fit.dis <- lm(crim ~ dis)

summary(fit.dis)

fit.rad <- lm(crim ~ rad)

summary(fit.rad)

fit.tax <- lm(crim ~ tax)

summary(fit.tax)

fit.ptratio <- lm(crim ~ ptratio)

summary(fit.ptratio)

fit.black <- lm(crim ~ black)

summary(fit.black)

fit.lstat <- lm(crim ~ lstat)

summary(fit.lstat)

fit.medv <- lm(crim ~ medv)

summary(fit.medv)

# Todas as variáveis possuem um valor p menor que 0,05, exceto chas.

# Devido a isso, é possível concluir que cada variável é significativa com a resposta.

## B)

fit.all <- lm(crim ~ ., data = Boston)

summary(fit.all)

# Pode-se rejeitar a hipótese nula de zn, dis, rad, black e medv.

## C)

simple.reg <- vector("numeric",0)

simple.reg <- c(simple.reg, fit.zn$coefficient[2])

simple.reg <- c(simple.reg, fit.indus$coefficient[2])

simple.reg <- c(simple.reg, fit.chas$coefficient[2])

simple.reg <- c(simple.reg, fit.nox$coefficient[2])

simple.reg <- c(simple.reg, fit.rm$coefficient[2])

simple.reg <- c(simple.reg, fit.age$coefficient[2])

simple.reg <- c(simple.reg, fit.dis$coefficient[2])

simple.reg <- c(simple.reg, fit.rad$coefficient[2])

simple.reg <- c(simple.reg, fit.tax$coefficient[2])

simple.reg <- c(simple.reg, fit.ptratio$coefficient[2])

simple.reg <- c(simple.reg, fit.black$coefficient[2])

simple.reg <- c(simple.reg, fit.lstat$coefficient[2])

simple.reg <- c(simple.reg, fit.medv$coefficient[2])

mult.reg <- vector("numeric", 0)

mult.reg <- c(mult.reg, fit.all$coefficients)

mult.reg <- mult.reg[-1]

plot(simple.reg, mult.reg, col = "red")

cor(Boston[-c(1, 4)])

# D)

fit.zn2 <- lm(crim ~ poly(zn, 3))

summary(fit.zn2)

fit.indus2 <- lm(crim ~ poly(indus, 3))

summary(fit.indus2)

fit.nox2 <- lm(crim ~ poly(nox, 3))

summary(fit.nox2)

fit.rm2 <- lm(crim ~ poly(rm, 3))

summary(fit.rm2)

fit.age2 <- lm(crim ~ poly(age, 3))

summary(fit.age2)

fit.dis2 <- lm(crim ~ poly(dis, 3))

summary(fit.dis2)

fit.rad2 <- lm(crim ~ poly(rad, 3))

summary(fit.rad2)

fit.tax2 <- lm(crim ~ poly(tax, 3))

summary(fit.tax2)

fit.ptratio2 <- lm(crim ~ poly(ptratio, 3))

summary(fit.ptratio2)

fit.black2 <- lm(crim ~ poly(black, 3))

summary(fit.black2)

fit.lstat2 <- lm(crim ~ poly(lstat, 3))

summary(fit.lstat2)

fit.medv2 <- lm(crim ~ poly(medv, 3))

summary(fit.medv2)

# Para zn, rm, rad, tax e lstat como preditor,

# os valores p sugerem que o coeficiente cúbico não é estatisticamente significativo.

# Para indus, nox, idade, dis, ptratio e medv como preditor,

# os valores de p sugerem a adequação do ajuste cúbico.

# Para black como preditor, os valores de p sugerem que os coeficientes quadrático

# e cúbico não são estatisticamente significativos, por isso,

# neste último caso, nenhum efeito não linear é visível.