A1

Rafsaan Sanvir

2024-06-06

The data file houses.txt contains measurements of house prices (in $1000s) and living area (in square feet) for 25 houses. Suppose we consider the simple normal linear regression to describe the relationship between the response Y (price) and the predictor X (living area). Answer the following data related questions:

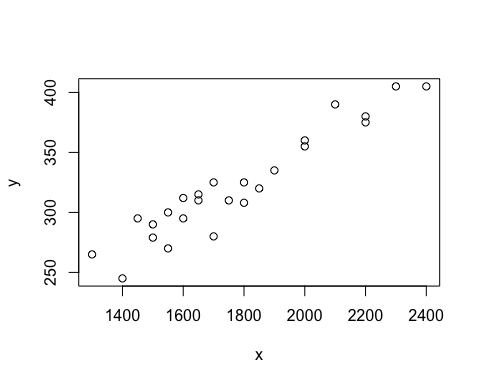
Import dataset and view

library(readr)  
houses <- read.table("houses.txt", header = TRUE)  
head(houses)

## price living\_area  
## 1 245 1400  
## 2 312 1600  
## 3 279 1500  
## 4 308 1800  
## 5 265 1300  
## 6 355 2000

3a) Plot the data in a scatter plot and comment on your plot. Solution:

x = houses$living\_area  
y = houses$price  
plot(x, y)

 The plot above seems to indicate a strong positive linear relationship.

3b) From now on we will assume that the Normal error regression model we discussed in class is appropriate for this data. Calculate the least squares regression line. Solution:

#manual calculation  
b1 = sum((x - mean(x))\*(y - mean(y))) / sum((x-mean(x))^2) #by least sq formula  
cat("least square est of b1 is", b1, "\n")

## least square est of b1 is 0.1419286

b0 = mean(y) - b1 \* mean(x)  
cat("least square est of b0 is", b0)

## least square est of b0 is 69.61087

fit = lm(y~x)  
summary(fit)

##   
## Call:  
## lm(formula = y ~ x)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -30.8896 -6.8539 -0.0824 10.3997 22.3390   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 69.610866 17.223584 4.042 0.000507 \*\*\*  
## x 0.141929 0.009562 14.843 2.85e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13.81 on 23 degrees of freedom  
## Multiple R-squared: 0.9055, Adjusted R-squared: 0.9014   
## F-statistic: 220.3 on 1 and 23 DF, p-value: 2.85e-13

Note how the least squares estimate matches the b0 and b1 value provided by R

3c) Conduct a t-test to determine whether or not there is a linear association between X and Y . Use α = 0.05 State the null and alternative hypotheses.

Solution: H0: B1 = 0 vs HA: B1 != 0 Note: From the lm summary above, stand error of b1 can be extracted, & n = 25

sxx = sum((x - mean(x))^2)  
alpha = 0.05  
sb1 = summary(fit)$coefficients[2,2]  
tstar = (b1 - 0)/sb1  
cat("test stat is:", tstar, "\n")

## test stat is: 14.84341

n = 25  
tcritical = qt(1 - alpha/2, n - 2)  
cat("critical value is:", tcritical, "\n")

## critical value is: 2.068658

Since Test statistic value 14.84 much greater than critical value 2.06, test stat falls in rejection region and we reject the null hypothesis. Meaning that a significant linear association between x and y exists.

1. Calculate the power of your test in part (c) above if actually β1 = 0.1. Assume σ = 44. Solution:

variance = 44^2  
sxx = sum((x - mean(x))^2)  
variance\_b1 = variance / sxx # by formula   
b1\_not = 0.1  
delta = abs(b1\_not - 0) / sqrt(variance\_b1) # by delta formula  
power = pt(qt(1 - (alpha / 2), n-2), n-2, ncp = delta, low = F) +   
 pt(-qt(1 - (alpha / 2), n-2 ), n-2, ncp = delta, low = T)  
power

## [1] 0.881453

This means that if b1 = 0.1, the probability would by 0.88 that we would be led to conclude with the alternate hypothesis that b1 != 0.

1. Calculate a 95% confidence interval for β1. Solution:

confint(fit, level = 0.95)

## 2.5 % 97.5 %  
## (Intercept) 33.9811683 105.2405637  
## x 0.1221487 0.1617086

95% CI for b1 is (0.1222, 0.1617). This means that if we conducted this analysis multiple times with different samples from the same population, 95% of the calculated confidence intervals would contain the true value of b1. The confidence interval also does not contain 0, meaning a statistically significant relationship exists between living area and house price.s

1. Calculate a 95% cofidence interval for the mean price of houses with a living area of 2000 square feet. Solution:

newdata = data.frame(x = 2000)  
predict(fit, newdata = newdata, interval = "confidence", level = 0.95)

## fit lwr upr  
## 1 353.4682 346.2627 360.6736

Confidence interval is (346.2627, 360.6736)

1. Calculate a 95% prediction interval for the price of a new house with a living area of 2000 square feet. Solution:

predict(fit, newdata = newdata, interval = "pred", level = 0.95)

## fit lwr upr  
## 1 353.4682 324.0093 382.927

Prediction interval is (324.0093, 382.927) We are 95% confident that the average price of houses with a living area of 2000 square feet is between 346.2627 and 360.6736.

1. Calculate the boundary values of the 95% confidence band for the re- gression line when the living area is Xh = 2000 square feet. Solution: Let’s use the boundary value formula

xh = 2000  
xbar = mean(x)  
yhat = fit$fitted.values  
SSE = sum((y - yhat)^2)  
MSE = SSE / (n - 2)  
s = sqrt(MSE)  
xbar = mean(x)  
yhath = fit$coefficients[1] + fit$coefficients[2] \*xh  
syhath = s \* sqrt(1/n+((xh-xbar)^2) / sxx)  
W = sqrt(2 \* qf(1 - alpha, df1 = 2, df2 = n-2))  
wh.lower = yhath - W \* syhath  
wh.upper = yhath + W\*syhath  
cat("lower value is ", wh.lower, "\n")

## lower value is 344.3557

cat("upper value is ", wh.upper, "\n")

## upper value is 362.5806

This means we are 95% confident that the true regression line (the mean response) for houses with a living area of 2000 square feet falls within this range.