Activity 8: Chapter7-Simple Linear Regression

Dr. Islam

March 31, 2022

# 7.1 Objectives

* Fit a linear regression model to predict mean response given a predictor
* Find variance and covariance of estimates of parameters
* Estimate and test hypothesis regarding model parameters
* Implement R for computation

**Remarks:**

* Please review all notes in Sections 7.1-7.5.
* It’s a brief summary of the text book materials

# 7.2 Simple Linear Regression Model

Two concepts are important

* A **response** () is an outcome, also called the dependent or explanined variable.
* A **predictor** or **explanatory variable** (), also called an independent variable, affects the response.

The true relationship between the response and the predictor is given by

where

is a function of the predictor , assumed to be linear, of the form:

is a random variable.

In other words, we can write the relationship between and as

Regression analysis models the mean response given a specific value of the predictor.

A simple linear regression model of on is defined as

The above equation at a specific value is given by

The above equation is a linear relation between mean response of at and a linear function of .

# 7.3 Estimation of a linear regression model

Given an ordered pair , an estimate of the linear regression model of on is the equation

where is the estimate of

Using the **method of the least squares**, one can verify that

and

The equation

is often called

* the least squares predicting equation
* the regression equation
* the equation of the best fit
* the predicting equation

The quantity is called the **prediction error** or **residual** due to the th observation.

**Theorem**

The variances of and and the covariance between and are given by

Estimates of SE of and and covariance between and are given by

where

# 7.4 Inferences

It appears that

Due to the above facts, we can test or using the following T-test statistics:

The CI estimates of and are given by

# 7.5 Prediction of Mean Response

In this section, we wish to predict the mean response, , at a specific value within the range of values of the predictor , usually called the interpolation.

The prediction of at is

Note that

and

After simple algebraic manipulation, it turns out that

Then,

Therefore, a CI estimate of is given by

where

and

**Activity 8**

In this activity, our main objective is to predict the circumference of trees using age for the default R data **Orange**, along with other relevant issues of the prediction. Note that the data **Orange** contains three columns Tree, age and circumference.

Let us introduce some notations:

Please review all notes in Sections 7.1-7.5 and answer the following questions:

1. Define response and explanatory variables in relation to the activity 10 objective.
2. Determine the number of rows in the dataset.
3. In order to estimate the least squares predicting equation, compute the estimates of and using the formula and notations of **ssqx**, **ssqy** and **ssqxy**.
4. Verify your estimates in (c) by fitting a linear model in R.
5. Find out an estimate of .
6. Find estimates of variance of , and estimate of covariance between and using the formula.
7. Verify the results in (f) using the results of the function **vcov()** applied to the result of the fitted model.
8. Find the predicted values of of on
9. Test the null hypotheses and against the two sided alternatives at 5% level of significance. Report the values of the test statistics and p-values.
10. Test the null hypotheses and against the two sided alternatives at 5% level of significance. Report the values of the test statistics and p-values.

#(a):  
#response=circumference, explanatory variable=age  
#(b)  
n=nrow(Orange)  
n

[1] 35

#(c)  
x=Orange$age  
y=Orange$circumference  
mx=mean(x)  
my=mean(y)  
ssqx=sum((x-mx)^2)  
ssqy=(n-1)\*var(y)  
ssqxy=sum((x-mx)\*(y-my))  
bet1=ssqxy/ssqx  
bet0=my-bet1\*mx  
#Estimates of beta\_0 and beta\_1  
c(bet0,bet1)

[1] 17.3996502 0.1067703

#(d)  
res=lm(circumference~age,data=Orange)  
attributes(res)

$names  
 [1] "coefficients" "residuals" "effects" "rank"   
 [5] "fitted.values" "assign" "qr" "df.residual"   
 [9] "xlevels" "call" "terms" "model"   
  
$class  
[1] "lm"

#estimates of beta\_0 and beta\_1 from fitted model  
betas=coef(res)  
# the results match with results in (c)  
betas

(Intercept) age   
 17.3996502 0.1067703

#(e) Estimates of sigmaHat  
sigmaHat=sqrt(sum(resid(res)^2)/(n-2))  
sigmaHat

[1] 23.73767

anova(res)

Analysis of Variance Table  
  
Response: circumference  
 Df Sum Sq Mean Sq F value Pr(>F)   
age 1 93772 93772 166.42 1.931e-14 \*\*\*  
Residuals 33 18595 563   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#(f)  
#Estimates of variance of beta\_0 and beta\_1 hats and their covariance   
varb1=sigmaHat^2/ssqx  
varb0=sigmaHat^2\*(1/n+mean(Orange$age)^2/ssqx)  
covb0b1=-mean(Orange$age)\*sigmaHat^2/ssqx  
c(varb0, varb1,covb0b1)

[1] 7.435026e+01 6.850249e-05 -6.316908e-02

#(g)variance and covariance obtained by cov() function  
vcov(res)

(Intercept) age  
(Intercept) 74.35026205 -6.316908e-02  
age -0.06316908 6.850249e-05

#variances are diagonal elements and covariance at off-diagonal of the cov() function above.  
#(h)  
prd1=sum(betas\*c(1, 284))  
prd2=sum(betas\*c(1, 250))  
c(prd1, prd2)

[1] 47.72242 44.09223

# We can get those results along with CIs, by the predict() function, defined on a dataframe datf.  
datf=data.frame(age=c(284,250))  
predict(res,datf,interval="confidence",level=0.90)

fit lwr upr  
1 47.72242 36.49717 58.94768  
2 44.09223 32.48418 55.70028

#i   
T0=bet0/sqrt(varb0)  
T1=bet1/sqrt(varb1)  
pvalue0=2\*pt(abs(T0), n-2,lower.tail = F)  
pvalue1=2\*pt(abs(T1), n-2,lower.tail = F)  
c(T0, pvalue0, T1, pvalue1)

[1] 2.017898e+00 5.179267e-02 1.290023e+01 1.930596e-14

#j   
T0=(bet0-15)/sqrt(varb0)  
T1=(bet1-0.15)/sqrt(varb1)  
pvalue0=2\*pt(abs(T0), n-2,lower.tail = F)  
pvalue1=2\*pt(abs(T1), n-2,lower.tail = F)  
c(T0, pvalue0, T1, pvalue1)

[1] 2.782958e-01 7.825217e-01 -5.223105e+00 9.552727e-06

#The summary(res) can be used to print results of above hypotheses tests:  
summary(res)

Call:  
lm(formula = circumference ~ age, data = Orange)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-46.310 -14.946 -0.076 19.697 45.111   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 17.399650 8.622660 2.018 0.0518 .   
age 0.106770 0.008277 12.900 1.93e-14 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 23.74 on 33 degrees of freedom  
Multiple R-squared: 0.8345, Adjusted R-squared: 0.8295   
F-statistic: 166.4 on 1 and 33 DF, p-value: 1.931e-14