Multiple Regression

Joshua Ibe

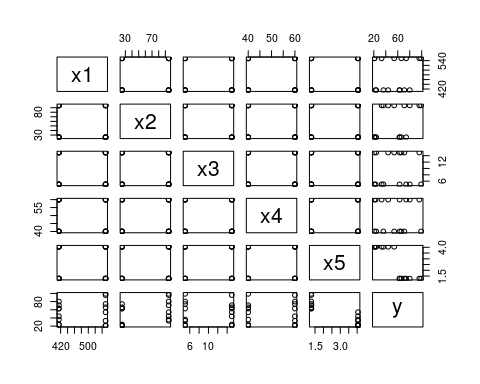
February 25, 2019

1. Fit a multiple linear regression model relating yield to these regressors

data=read.csv('/home/jayibex/Desktop/Data\_Sets/B7.csv', header=T)  
head(data)

## x1 x2 x3 x4 x5 y  
## 1 415 25 5 40 1.28 63  
## 2 550 25 5 40 4.05 21  
## 3 415 95 5 40 4.05 36  
## 4 550 95 5 40 1.28 99  
## 5 415 25 15 40 4.05 24  
## 6 550 25 15 40 1.28 66

plot(data)



y= data[,6]  
x1=data[,1]  
x2= data[,2]  
x3= data[,3]  
x4= data[,4]  
x5= data[,5]  
  
n=length(y)  
  
n

## [1] 16

fit=lm(y~ x1 +x2 +x3 +x4 +x5, data)  
fit

##   
## Call:  
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = data)  
##   
## Coefficients:  
## (Intercept) x1 x2 x3 x4   
## 5.208e+01 5.556e-02 2.821e-01 1.250e-01 1.776e-16   
## x5   
## -1.606e+01

From the plot is seems there is no instance of multicolinearity.

Model is y= 5.208e+01 +5.556e-02x1 +2.821e-01x2 + 1.250e-01x3 +1.776e-16x4 +-1.606e+01x5 Which is equivalent to (Total Yield of Oil per Batch of Peanuts) = 5.208e+01 +5.556e-02(pressure) +2.821e-01(temp) + 1.250e-01(moisture) +1.776e-16(flow rate) +-1.606e+01(peanut particle size)

1. Test for significance of regression. What canclusions can you draw?

We must draw the necessary components to conduct a p-value test.

summary(fit)

##   
## Call:  
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.250 -4.438 0.125 5.250 9.500   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.208e+01 1.889e+01 2.757 0.020218 \*   
## x1 5.556e-02 2.987e-02 1.860 0.092544 .   
## x2 2.821e-01 5.761e-02 4.897 0.000625 \*\*\*  
## x3 1.250e-01 4.033e-01 0.310 0.762949   
## x4 1.776e-16 2.016e-01 0.000 1.000000   
## x5 -1.606e+01 1.456e+00 -11.035 6.4e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.065 on 10 degrees of freedom  
## Multiple R-squared: 0.9372, Adjusted R-squared: 0.9058   
## F-statistic: 29.86 on 5 and 10 DF, p-value: 1.055e-05

anov=anova(fit)  
anov

## Analysis of Variance Table  
##   
## Response: y  
## Df Sum Sq Mean Sq F value Pr(>F)   
## x1 1 225.0 225.0 3.4589 0.0925445 .   
## x2 1 1560.2 1560.2 23.9854 0.0006254 \*\*\*  
## x3 1 6.2 6.2 0.0961 0.7629488   
## x4 1 0.0 0.0 0.0000 1.0000000   
## x5 1 7921.0 7921.0 121.7679 6.401e-07 \*\*\*  
## Residuals 10 650.5 65.1   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

k= 5 #number of x's  
p= 6 #number of coefficients(betas) in the model p = k +1  
num\_res= 10  
df\_f= (k+num\_res)-p   
SSr= 650.5 #Sum Square Residuals  
MSr= SSr/2   
MSres = 65.1 #Mean Square Residuals  
F0=MSr/MSres  
  
#p0 (critical value)  
abs(qt(0.025, df\_f))

## [1] 2.262157

#p-value  
pf(F0, df1=5, df2=n-p, lower.tail=FALSE)

## [1] 0.01490584

#Decision: ?  
# p0 is greater than p-val therefore we reject the null hypothesis

1. Use t test to access the contribution of each regressor to the model. Discuss your findings. The larger the F value the more likely fail to reject.

Each regressor coeffiecent is subjected to the hypothesis test $ H\_{0}: = 0 \ H\_{1}: = 0$

Test Statistic:

Reject if >   
or >

From the summary table, we observe on the column Pr(>|t|)

The intercept has a p-value of approximately 0.02 which is less than 0.05 (5% confidence interval), which implies that the intercept is not equal to zero

has a p-value of approximately 0.09 which is greater than 0.05 (5% confidence interval), which implies that the regressor is not contributing significantly to the model of y given that the rest of the regressors is in the model

has a p-value of approximately 0.0006 which is less than 0.05 (5% confidence interval), which implies that the regressor is contributing significantly to the model of y given that the rest of the regressors is in the model

has a p-value of approximately 0.7 which is greater than 0.05 (5% confidence interval), which implies that the regressor is not contributing significantly to the model of y given that the rest of the regressors is in the model

has a p-value of approximately 1.00 which is less than 0.05 (5% confidence interval), which implies that the regressor is not contributing significantly to the model of y given that the rest of the regressors is in the model

has a p-value of approximately 0.0006 which is less than 0.05 (5% confidence interval), which implies that the regressor is contributing significantly to the model of y given that the rest of the regressors is in the model

1. Ralculate R^2 and R\_adj^2 for this model. Compare these values to the R^2 and the R\_adj^2 for the multiple linear regression model relatinig yeild to tempuratre and the particle size. Discuss your results.

sum\_fit = summary(fit)  
  
sum\_fit$r.squared

## [1] 0.9372286

R-squared is 0.9372286 from the summary table, which implies that approimately 94% of the model is explained by the 5 regressor variables

Since there are regressors that are not contributing to the true model; , , . We need to get R-squared adjusted to show the true percentage that is explained by the relevant regressors and

sum\_fit$adj.r.squared

## [1] 0.9058429

0.9058429 or approximately 91% represents the true percentage of the y variable, the model explained by the two regressor variables and .

e.) Finally we would like to see the confidence intervals for the coeffiecent temperature which was in the model. Using the confidence interval function build into R

confit = confint(fit)  
confit

## 2.5 % 97.5 %  
## (Intercept) 9.99688896 94.1612109  
## x1 -0.01100273 0.1221138  
## x2 0.15378045 0.4105053  
## x3 -0.77353688 1.0235369  
## x4 -0.44926844 0.4492684  
## x5 -19.30879739 -12.8211665

confit[3, ]

## 2.5 % 97.5 %   
## 0.1537804 0.4105053

We are 95% confident that the temperature coeffieicnt (slope) is between 0.1537804 and 0.4105053 and zero is not in that interval.