ex4R

2023-04-24

data("marketing", package = "datarium")  
head(marketing)

## youtube facebook newspaper sales  
## 1 276.12 45.36 83.04 26.52  
## 2 53.40 47.16 54.12 12.48  
## 3 20.64 55.08 83.16 11.16  
## 4 181.80 49.56 70.20 22.20  
## 5 216.96 12.96 70.08 15.48  
## 6 10.44 58.68 90.00 8.64

Lets first see what a default simple linear model looks like

model = lm(sales ~ youtube + facebook + newspaper, data=marketing)  
summary(model)

##   
## Call:  
## lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.5932 -1.0690 0.2902 1.4272 3.3951   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.526667 0.374290 9.422 <2e-16 \*\*\*  
## youtube 0.045765 0.001395 32.809 <2e-16 \*\*\*  
## facebook 0.188530 0.008611 21.893 <2e-16 \*\*\*  
## newspaper -0.001037 0.005871 -0.177 0.86   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.023 on 196 degrees of freedom  
## Multiple R-squared: 0.8972, Adjusted R-squared: 0.8956   
## F-statistic: 570.3 on 3 and 196 DF, p-value: < 2.2e-16

We can see that the p-value of newspaper is 0.86 which makes it statistically insignificant. We can do variable selection to remove it for better results. Lets build 3 different models with glmnet and examine the results (MSE and SE).

library(glmnet)

## Warning: package 'glmnet' was built under R version 4.2.3

## Loading required package: Matrix

## Loaded glmnet 4.1-7

# predictor variable x excluding newspaper since the p value was high  
x <- as.matrix(marketing[, 1:2])  
y <- marketing$sales  
  
# alpha 0 = ridge regression  
mod\_cv <- cv.glmnet(x, y, alpha = 0, nfolds = 10, type.measure = "mse")  
cat("10 fold cv model: min MSE =",min(mod\_cv$cvm), "SE = ",min(mod\_cv$cvsd), "\n")

## 10 fold cv model: min MSE = 4.349579 SE = 0.7589857

# folds = nrow(x) means that looc is used  
mod\_loocv <- cv.glmnet(x, y, alpha = 0, type.measure = "mse", nfolds = nrow(x), grouped=FALSE)  
cat("loocv model: min MSE =",min(mod\_loocv$cvm), "SE = ",min(mod\_loocv$cvsd), "\n")

## loocv model: min MSE = 4.356527 SE = 0.729291

# alpha 1 = lasso regression  
lasso\_cv <- cv.glmnet(x, y, alpha = 1, nfolds = 10, type.measure = "mse")  
cat("lasso cv model: min MSE =",min(lasso\_cv$cvm), "SE = ",min(lasso\_cv$cvsd), "\n")

## lasso cv model: min MSE = 4.194974 SE = 0.6247761

Observing MSE and Standard Error, we can see that lasso model with 10 fold cv achieves the lowest MSE and Standard Error, while the two ridge models have very similar estimates. Using these estimates as a score for best fit, we conclude that the cross validated lasso model is the “best”. However, caliming that a model is best in this case (or any case to be fair) is very questionable as there are many other things to consider, so these claims should be considered with caution.

flu <- matrix(c(3, 10, 15, 13), 2)  
colnames(flu) <- c("Shot = True", "Shot = False")  
rownames(flu) <- c("Flu = True", "Flu = False")  
flu

## Shot = True Shot = False  
## Flu = True 3 15  
## Flu = False 10 13

#I Null hypothesis: There is no significant difference in the propotion of people who got flu between people who got flu shot and people who didn’t get flu shot.

Alternative hypotherisis: The propotion of people who got flu is significantly different between people who got flu shot and people who didn’t get flu shot.

We can use the fisher test for this case

fisher.test(flu)

##   
## Fisher's Exact Test for Count Data  
##   
## data: flu  
## p-value = 0.09548  
## alternative hypothesis: true odds ratio is not equal to 1  
## 95 percent confidence interval:  
## 0.03904735 1.35074988  
## sample estimates:  
## odds ratio   
## 0.2686768

#II The p-value is significant when alpha > 0.1, but not anything lower than this. So we can conclude that there is difference between the groups on 90% confidence level, but anything higher than this is not statistically significant.

#III

We can estimate p-value by calculating z-value first and then using normal distribution.

# flu shot  
n1 <- 13  
# no flu shot  
n2 <- 28  
# no flu  
x1 <- 3  
# got flu  
x2 <- 15   
  
# propotions  
p1 <- x1/n1  
p2 <- x2/n2  
  
p\_hat = (x1 + x2)/(n1 + n2)  
   
# standard error  
se = sqrt(p\_hat\*(1-p\_hat)\*(1/n1 + 1/n2))  
  
# z-value  
z <- (p1 - p2)/se  
  
# as z-value is normally distributed, we can calculate p-value using it  
p\_value <- 2 \* pnorm(-abs(z))  
  
p\_value

## [1] 0.06711584

The p-value is low, but still wont quite achieve significance level of 0.05. However the results from both of the tests indicate difference between the groups with 90 confidence level, we can reject the null hypothesis on both of the tests and conclude that there is statistical significance between the propotions of people with or without the flu shot.