Problem set 3

Sayan Ganguly,Roll:714

2026-02-12

Problem 2

Attach "Credits" data from R. Regress "balance" on (a) "gender" only

```
library(ISLR)
head(Credit)
```

```
##   ID  Income Limit Rating Cards Age Education Gender Student Married
Ethnicity
## 1  1  14.891  3606    283     2  34        11   Male      No     Yes
Caucasian
## 2  2 106.025  6645    483     3  82        15 Female     Yes     Yes
Asian
## 3  3 104.593  7075    514     4  71        11   Male      No      No
Asian
## 4  4 148.924  9504    681     3  36        11 Female      No      No
Asian
## 5  5  55.882  4897    357     2  68        16   Male      No     Yes
Caucasian
## 6  6  80.180  8047    569     4  77        10   Male      No      No
Caucasian
##   Balance
## 1     333
## 2     903
## 3     580
## 4     964
## 5     331
## 6    1151
```

```
model1=lm(Balance~Gender,data=Credit)
model1
```

```
##
## Call:
## lm(formula = Balance ~ Gender, data = Credit)
##
## Coefficients:
##  (Intercept)  GenderFemale
##       509.80         19.73
```

(b) "gender" and "ethnicity"

```
model2=lm(Balance~Gender+Ethnicity,data=Credit)
model2
```

```
##
## Call:
## lm(formula = Balance ~ Gender + Ethnicity, data = Credit)
##
## Coefficients:
##        (Intercept)        GenderFemale      EthnicityAsian
EthnicityCaucasian
##             520.88               20.04              -19.37              -
12.65
```

(c) “gender”, “ethnicity”, “income”

```
model3=lm(Balance~Gender+Ethnicity+Income,data=Credit)
model3

##
## Call:
## lm(formula = Balance ~ Gender + Ethnicity + Income, data = Credit)
##
## Coefficients:
##        (Intercept)        GenderFemale      EthnicityAsian
EthnicityCaucasian
##            230.029              24.340               1.637
6.447
##              Income
##               6.054
```

(d) Output all the regressions in (a)-(c) in a single table using stargazer. Comment on the significant coefficients in each of the models.

```
library(stargazer)

##
## Please cite as:

##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary
Statistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer

stargazer(model1,model2,model3,type="text",title="output")

##
## output
##
================================================================================
======
##                                         Dependent variable:
##                     ---------------------------------------------------
---------
##                                                 Balance
##                             (1)                 (2)                   (3)
## --------------------------------------------------------------------------
```

```
---------
## GenderFemale              19.733              20.038                24.340
##                          (46.051)            (46.178)
(40.963)
##
## EthnicityAsian                                -19.371                1.637
##                                              (65.107)
(57.787)
##
## EthnicityCaucasian                            -12.653                6.447
##                                              (56.740)
(50.363)
##
## Income
6.054***
##
(0.582)
##
## Constant                509.803***          520.880***
230.029***
##                          (33.128)            (51.901)
(53.857)
##
## ------------------------------------------------------------------------
---------
## Observations                400                 400                   400
## R2                         0.0005              0.001                0.216
## Adjusted R2                -0.002              -0.007               0.208
## Residual Std. Error 460.230 (df = 398)  461.337 (df = 396)    409.218 (df
= 395)
## F Statistic         0.184 (df = 1; 398) 0.092 (df = 3; 396) 27.161*** (df
= 4; 395)
##
========================================================================
======
## Note:                                                *p<0.1; **p<0.05;
***p<0.01
```

(e) Explain how gender affects “balance” in each of the models (a)- (c)

Ans: In each of the three models if the gender is female then the average balance is increased and the increase are approximately 20,20 and 24 units respectively.

(f) Compare the average credit card balance of a male African with a male Caucasian on the basis of model (b).

Ans: Based on model (b) we can say that for a male Caucasian the average credit card balance is approximately 13 units less than an African male.

(g) Compare the average credit card balance of a male African with a male Caucasian when each earns 100,000 dollars. For comparison, use the model in (c).

Ans: When the income is fixed,on the basis of model (c) we can conclude that average credit card balance of a male Caucasian is approximately 7 units more than that of a male African.

(h) Compare and comment on the answers in (f) and (g)

Ans: The two answers are contradictory.So we see that if we include "income" as another predictor and form a new model the interpretation with respect to the other predictor "ethnicity" also changes.

(i) Based on the model in (c), predict the credit card balance of a female Asian whose income is 2000,000 dollars

Ans:

```
response=24.340+1.637+6.054*2000
response
```

```
## [1] 12133.98
```

(j) Check the goodness of fit of the different models in (a)-(c) in terms of adjusted R2. Which model would you prefer? **Ans:**Since the adjusted R squared value is greatest in case of the third model therefore the fit is most satisfactory in this case and we prefer the third model.

Problem 4

Step 1: Generate x1i from Normal(0,1) distribution, i = 1,2,..,n Step 2: Generate x2i from Bernoulli (0.3) distribution, i = 1,2,..,n Step 3: Generate ϵi from Normal(0,1) and hence generate the response yi =β0 +β1x1i +β2x2i +β3(x1i ×x2i)+ϵi, i = 1,2,,,n. Step 4: Run two separate multiple linear regressions (i) using the model in Step 3 and (ii) using the model in Step 3 without the interaction term.Repeat Steps 1-4 , R = 1000 times. At each simulation compute the MSE forthe correct model (i.e. model with the interaction term) and the naive model(i.e. the model without the interaction term). Finally find the average MSE'sfor each model. From the output, demonstrate the impact of ignoring the interaction term.Carry out the analysis for n = 100 and the following parametric configurations: (β0,β1,β2,β3) = (−2.5,1.2,2.3,0.001) , (-2.5, 1.2. 2.3, 3.1). Set seed as 123

For the first choice of the parameters

```
set.seed(123)
for(i in 1:1000){
  x1=rnorm(100,0,1)
  x2=rbinom(100,1,0.3)
  e=rnorm(100,0,1)
```

```
  y1= -2.5 + 1.2*x1 + 2.3*x2 + 0.001*(x1*x2) + e
  model1=lm(y1~x1+x2+x1*x2)
  model2=lm(y1~x1+x2)
  mse1=mean((y1-predict(model1))^2)
  avg_mse1=mse1/1000
  mse2=mean((y1-predict(model2))^2)
  avg_mse2=mse2/1000
}
avg_mse1

## [1] 0.001015287

avg_mse2

## [1] 0.001015437
```

Conclusion: So if we do not include the interaction term the average mse increases a little.

For the second choice of the parameters

```
set.seed(123)
for(i in 1:1000){
  x1=rnorm(100,0,1)
  x2=rbinom(100,1,0.3)
  e=rnorm(100,0,1)
  y1= -2.5 + 1.2*x1 + 2.3*x2 + 3.1*(x1*x2) + e
  model1=lm(y1~x1+x2+x1*x2)
  model2=lm(y1~x1+x2)
  mse1=mean((y1-predict(model1))^2)
  avg_mse1=mse1/1000
  mse2=mean((y1-predict(model2))^2)
  avg_mse2=mse2/1000
}
avg_mse1

## [1] 0.001015287

avg_mse2

## [1] 0.002816845
```

Conclusion: here we see that by not including the interaction term the average mse increases by quite some amount.