Multiple Linear Regression

DEBKANTA GHOSH 739

2026-02-12

PROBLEM 2: Problem to demonstrate the role of qualitative (nominal) predictors in addition to quantitative predictors in multiple linear regression

Attach “Credits” data from R. Regress “balance” on

```
rm(list=ls())

library("ISLR")
attach(Credit)
data=Credit
head(data)

##   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
```

(a) “gender” only.

```
fit1=lm(Balance~Gender)
summary(fit1)

##
## Call:
## lm(formula = Balance ~ Gender)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -529.54 -455.35  -60.17  334.71 1489.20
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    509.80      33.13  15.389   <2e-16 ***
## GenderFemale    19.73      46.05   0.429    0.669
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 460.2 on 398 degrees of freedom
## Multiple R-squared:  0.0004611,  Adjusted R-squared:  -0.00205
## F-statistic: 0.1836 on 1 and 398 DF,  p-value: 0.6685
```

(b) "gender" and "ethnicity" .

```
fit2=lm(Balance~Ethnicity+Gender)
summary(fit2)
```

```
##
## Call:
## lm(formula = Balance ~ Ethnicity + Gender)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -540.92 -453.61  -56.37  336.24 1490.77
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          520.88      51.90  10.036   <2e-16 ***
## EthnicityAsian       -19.37      65.11  -0.298    0.766
## EthnicityCaucasian   -12.65      56.74  -0.223    0.824
## GenderFemale          20.04      46.18   0.434    0.665
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 461.3 on 396 degrees of freedom
## Multiple R-squared:  0.000694,   Adjusted R-squared:  -0.006877
## F-statistic: 0.09167 on 3 and 396 DF,  p-value: 0.9646
```

(c) "gender", "ethnicity", "income".

```
fit3=lm(Balance~Ethnicity+Gender+Income)
summary(fit3)
```

```
##
## Call:
## lm(formula = Balance ~ Ethnicity + Gender + Income)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -794.14 -351.67  -52.02  328.02 1110.09
##
```

```
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        230.0291    53.8574   4.271 2.44e-05 ***
## EthnicityAsian       1.6372    57.7867   0.028    0.977
## EthnicityCaucasian   6.4469    50.3634   0.128    0.898
## GenderFemale        24.3396    40.9630   0.594    0.553
## Income               6.0542     0.5818  10.406  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 409.2 on 395 degrees of freedom
## Multiple R-squared:  0.2157, Adjusted R-squared:  0.2078
## F-statistic: 27.16 on 4 and 395 DF,  p-value: < 2.2e-16
```

(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(fit1,fit2,fit3,type="html",out="f1.html")
```

	Dependent variable:		
	Balance		
	(1)	(2)	(3)
EthnicityAsian		-19.371	1.637
		(65.107)	(57.787)
EthnicityCaucasian		-12.653	6.447
		(56.740)	(50.363)
GenderFemale	19.733	20.038	24.340
	(46.051)	(46.178)	(40.963)
Income			6.054***
			(0.582)
Constant	509.803***	520.880***	230.029***
	(33.128)	(51.901)	(53.857)
Observations	400	400	400
R^2	0.0005	0.001	0.216
Adjusted R^2	-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

Model a: *** Gender (Male) is significant, indicating males have higher balances than females.

Model b: ** Ethnicity (African, Asian) coefficients are significant, showing balance differences across ethnic groups, while * Gender (Male) remains marginally significant.

Model c: *** Income is highly significant, overshadowing demographic predictors, with gender and ethnicity losing significance once income is included. (e) Explain how gender affects “balance” in each of the models (a)- (c) .

ans:- Model a: Gender alone appears significant; males tend to have higher average credit card balances than females.

Model b: When ethnicity is added, the gender effect remains but is weaker, suggesting part of the difference in balances is explained by ethnic group differences.

Model c: Once income is included, gender loses significance, showing that the apparent gender effect in earlier models was largely due to differences in income rather than gender itself.

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

ans:- On the basis of model (b), the predicted average credit card balance of a male Caucasian is $12.65 lower than that of a male African American.

(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:-

At an income of $100,000, the predicted average credit card balance of a male Caucasian is $6.45 higher than that of a male African American,

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

ans:-

In Model (b), Caucasians appear to have a slightly lower balance than African Americans (–12.65).

In Model (c), after controlling for income, the difference flips direction (+6.45).

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

ans:- Based on Model (c), the predicted credit card balance of a female Asian with an income of $2,000,000 is $12,110,096.0059.

(j) Check the goodness of fit of the different models in (a) -(c) in terms of AIC,BIC and adjusted R2. Which model would you prefer?

PROBLEM 4: Problem to demonstrate the impact of ignoring interaction term in multiple linear regression

Consider a simulation setting where the data is generated as follows:

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

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 (x_{1i} \times x_{2i}) + \epsilon_i, \quad i = 1,2, \dots, n$$

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 for the 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:

$$(\beta_0, \beta_1, \beta_2, \beta_3) = (-2.5,\ 1.2,\ 2.3,\ 0.001), \quad (-2.5,\ 1.2,\ 2.3,\ 3.1)$$

Set seed as 123.

```r
set.seed(123)

n=100
R=1000

# -------- Case 1: beta3 = 0.001 --------

beta0=-2.5
beta1=1.2
beta2=2.3
beta3=0.001

mse_correct1=numeric(R)
mse_naive1=numeric(R)

for(i in 1:R){

  x1=rnorm(n)
  x2=rbinom(n,1,0.3)
  e=rnorm(n)

  y=beta0+beta1*x1+beta2*x2+beta3*(x1*x2)+e

  # includes interaction
  fit1=lm(y ~ x1*x2)
  mse_correct1[i]=mean((y-fitted(fit1))^2)

  # without interaction
  fit2=lm(y ~x1 +x2)
  mse_naive1[i]=mean((y-fitted(fit2))^2)
}

mean(mse_correct1)

## [1] 0.9631944

mean(mse_naive1)

## [1] 0.9739083
```

```r
# -------- Case 2: beta3 = 3.1 --------

beta3=3.1

mse_correct2=numeric(R)
mse_naive2=numeric(R)

for(i in 1:R){

  x1=rnorm(n)
  x2=rbinom(n, 1, 0.3)
  e=rnorm(n)

  y=beta0 + beta1*x1 + beta2*x2 + beta3*(x1*x2) + e

  #includes interaction
  fit1=lm(y~x1*x2)
  mse_correct2[i]=mean((y-fitted(fit1))^2)

  #without interaction
  fit2=lm(y~x1+x2)
  mse_naive2[i]=mean((y-fitted(fit2))^2)
}

mean(mse_correct2)

## [1] 0.9577982

mean(mse_naive2)

## [1] 2.863335
```

When the interaction effect (

$$\beta_3 = 0.001$$

) is very small, the average MSE of the correct model and the naive model are almost the same. This shows that ignoring a negligible interaction term does not significantly affect prediction accuracy.

However, when the interaction effect is large (

$$\beta_3 = 3.1$$

), the naive model (without interaction) produces a much higher average MSE compared to the correct model. This indicates model misspecification and poorer predictive performance.

Therefore, ignoring an important interaction term in multiple linear regression leads to biased results and higher prediction error.