PROBLEM SHEET 3_720

Srijita Dey

2026-02-12

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 (a) “gender” only. (b) “gender” and “ethnicity” . (c) “gender”, “ethnicity”, “income”. (d) Output all the regressions in (a)-(c) in a single table using stargazer. Com- ment on the significant coefficients in each of the models. (e) Explain how gender affects “balance” in each of the models (a)- (c) . (f) Compare the average credit card balance of a male African with a male Caucasian on the basis of model (b). (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). (h) Compare and comment on the answers in (f) and (g) (i) Based on the model in (c), predict the credit card balance of a female Asian whose income is 2000,000 dollars. (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?

```
rm(list=ls())
library(ISLR)

## Warning: package 'ISLR' was built under R version 4.5.2

??Credit

## starting httpd help server ... done

attach(Credit)

#a
m1=lm(Balance~Gender)
summary(m1)

##
## 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
m2=lm(Balance~Gender+Ethnicity)
m2

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

#c
m3=lm(Balance~Gender+Ethnicity+Income)
m3

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

#d
library(stargazer)

## Warning: package 'stargazer' was built under R version 4.5.2

##
## 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(m1,m2,m3,type="html",out="f3.html")

##
## <table style="text-align:center"><tr><td colspan="4" style="border-bottom:
```

```
1px solid black"></td></tr><tr><td style="text-align:left"></td><td
colspan="3"><em>Dependent variable:</em></td></tr>
## <tr><td></td><td colspan="3" style="border-bottom: 1px solid
black"></td></tr>
## <tr><td style="text-align:left"></td><td colspan="3">Balance</td></tr>
## <tr><td style="text-
align:left"></td><td>(1)</td><td>(2)</td><td>(3)</td></tr>
## <tr><td colspan="4" style="border-bottom: 1px solid
black"></td></tr><tr><td style="text-
align:left">GenderFemale</td><td>19.733</td><td>20.038</td><td>24.340</td></t
r>
## <tr><td style="text-
align:left"></td><td>(46.051)</td><td>(46.178)</td><td>(40.963)</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td style="text-align:left">EthnicityAsian</td><td></td><td>-
19.371</td><td>1.637</td></tr>
## <tr><td style="text-
align:left"></td><td></td><td>(65.107)</td><td>(57.787)</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td style="text-align:left">EthnicityCaucasian</td><td></td><td>-
12.653</td><td>6.447</td></tr>
## <tr><td style="text-
align:left"></td><td></td><td>(56.740)</td><td>(50.363)</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td style="text-
align:left">Income</td><td></td><td></td><td>6.054<sup>***</sup></td></tr>
## <tr><td style="text-
align:left"></td><td></td><td></td><td>(0.582)</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td style="text-
align:left">Constant</td><td>509.803<sup>***</sup></td><td>520.880<sup>***</s
up></td><td>230.029<sup>***</sup></td></tr>
## <tr><td style="text-
align:left"></td><td>(33.128)</td><td>(51.901)</td><td>(53.857)</td></tr>
## <tr><td style="text-align:left"></td><td></td><td></td><td></td></tr>
## <tr><td colspan="4" style="border-bottom: 1px solid
black"></td></tr><tr><td style="text-
align:left">Observations</td><td>400</td><td>400</td><td>400</td></tr>
## <tr><td style="text-
align:left">R<sup>2</sup></td><td>0.0005</td><td>0.001</td><td>0.216</td></tr
>
## <tr><td style="text-align:left">Adjusted R<sup>2</sup></td><td>-
0.002</td><td>-0.007</td><td>0.208</td></tr>
## <tr><td style="text-align:left">Residual Std. Error</td><td>460.230 (df =
398)</td><td>461.337 (df = 396)</td><td>409.218 (df = 395)</td></tr>
## <tr><td style="text-align:left">F Statistic</td><td>0.184 (df = 1;
398)</td><td>0.092 (df = 3; 396)</td><td>27.161<sup>***</sup> (df = 4;
395)</td></tr>
## <tr><td colspan="4" style="border-bottom: 1px solid
black"></td></tr><tr><td style="text-align:left"><em>Note:</em></td><td
```

```
colspan="3" style="text-align:right"><sup>*</sup>p<0.1; <sup>**</sup>p<0.05;
<sup>***</sup>p<0.01</td></tr>
## </table>
```

```
================================================================================
                                      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
```

From the table above we see the presence of stars in the coefficient of Income only and hence we conclude that Income is a significant predictor used in predicting Credit.

(e) The regression equation of model (a) is

$$\widehat{\text{Balance}} = \beta_0 + 20.04\ (\text{GenderFemale})$$

Here the baseline is male for gender. Here, females have 20.04units of higher average balance than males . The regression equation of the model(b) is given by

$$\widehat{\text{Balance}} = 520.88 + 20.04\ (\text{GenderFemale}) - 19.37\ (\text{EthnicityAsian}) - 12.65\ (\text{EthnicityCaucasian})$$

Here the baseline Categories are Male in Gender and African in Ethnicity. When ethnicity is held constant, Females have a higher balance by $20.04 in comparison to males. The regression equation of the model(c) is given by

$$\widehat{\text{Balance}} = 230.029 + 24.340\ (\text{GenderFemale}) + 1.637\ (\text{EthnicityAsian}) + 6.447\ (\text{EthnicityCaucasian}) + 6.054\ (\text{Income})$$

Here the baseline Categories are Male in Gender and African in Ethnicity. When ethnicity and income are held constant, Females have a higher credit balance by $24.340 in comparison to Males.

(f) The regression equation of the model(b) is given by $$ = 520.88

- 20.04,()
- 19.37,()
- 12.65,() $$ The baseline Categories are Male in Gender and African in Ethnicity. To calculate Average Credit balance of a male African we put Gender_Female=0 and Ethnnicity_Asian and Ethnnicity_Caucasian both equal to 0. Hence from the fitted model, the intercept ($520.88) represents the average balance of a male African. To calculate Average Credit balance of a male Caucasian we put Gender_Female=0 and Ethnnicity_Asian=0 and Ethnicity_Caucasian=1. On putting the values in the fitted model, $508.23 the average balance of a male African. Hence, the average credit card balance of a male African exceeds that of a male Caucasian by $12.65 when all other factors are held constant.

(g)The regression equation of the model(c) is given by

$$\widehat{\text{Balance}} = 230.029 + 24.340\ (\text{GenderFemale}) + 1.637\ (\text{EthnicityAsian}) + 6.447\ (\text{EthnicityCaucasian}) + 6.054\ (\text{Income})$$

The baseline Categories are Male in Gender and African in Ethnicity. To calculate Average Credit balance of a male African when income is $100000, we put Gender_Female=0 and Ethnnicity_Asian and Ethnnicity_Caucasian both equal to 0. Hence from the fitted model, the average balance of a male African is calculated as $605630.029. To calculate Average Credit balance of a male Caucasian with income is 100000, we put Gender_Female=0,Ethnnicity_Asian=0 and Ethnicity_Caucasian=1. Hence from the fitted model we get the average credit balance of a male Caucasian as $605636.476. A male Caucasian has a higher average credit balance(by 6.447 dollars)in comparison to a male African when income of both are $100,000 and all other factors are held constant.

(h) From model (b), the average credit card balance of a male African exceeds that of a male Caucasian by $12.65 when all other factors are held constant.

From model (c),a male Caucasian has a higher average credit balance(by 6.447 dollars)in comparison to a male African when income of both are $100,000 and all other factors are held constant.

(i)In case of the model (c),the baseline Categories are Male in Gender and African in Ethnicity. To calculate Average Credit balance of a female Asian when income is $2000000, we put Gender_Female=1,Ethnnicity_Asian=1 and Ethnnicity_Caucasian= 0. Hence from the fitted model, the average balance of a female Asian is calculated as $12108256.006

(j)

	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

Here the adjusted R^2 is least for model (b). Since the adjusted R^2 accounts for model complexity and the adjusted R^2 is maximum here in case of model (c) which suggests that out of the 3 models given here, model (c) has the best fit.

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 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 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's

for 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.

```
rm(list=ls())
set.seed(123)
sim_fun=function(n, beta0, beta1, beta2, beta3, R=1000){
  mse_correct=numeric(R)
  mse_naive=numeric(R)
  for(r in 1:R){
    x1=rnorm(n, 0, 1)
    x2=rbinom(n, 1, 0.3)
    eps=rnorm(n, 0, 1)
    y=beta0 + beta1*x1 + beta2*x2 + beta3*(x1*x2) + eps

    fit_correct=lm(y ~ x1*x2)
    yhat_correct= predict(fit_correct)
    mse_correct[r]=mean((y - yhat_correct)^2)

    fit_naive=lm(y ~ x1 + x2)
    yhat_naive=predict(fit_naive)
    mse_naive[r]=mean((y - yhat_naive)^2)
  }

  return(c(mean(mse_correct), mean(mse_naive)))
}

sim_fun(100,-2.5, 1.2, 2.3, 0.001)

## [1] 0.9631944 0.9739083

sim_fun(100,-2.5, 1.2,2.3, 3.1)

## [1] 0.9577982 2.8633349
```

When the interaction effect is very small (β_3= 0.001), the average MSE of the correct and naive models are almost identical (0.9632 and 0.9739 respectively). Hence, ignoring the interaction term in this case does not affect model performance much.

However, when the interaction effect is strong (β_3= 3.1), the naive model produces a much larger average MSE=2.8633 compared to the correct model=0.9578. This shows that omitting a significant interaction term leads to substantial loss in predictive accuracy.

Therefore, ignoring important interaction terms results in model misspecification and significantly higher prediction error.