Assignment_3_710

PROTYAY CHATTERJEE

2026-02-11

Problem Set 3: Multiple Linear Regression

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

Attach "Credits" data from R.

```
library(ISLR)
```

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

```
data(Credit)
attach(Credit)
```

Regress "balance" on

(a) "gender" only.

```
m_a=lm(Balance~Gender,data=Credit)
summary(m_a)
```

```
##
## Call:
## lm(formula = Balance ~ Gender, data = Credit)
##
## 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"

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

```
##
## Call:
## lm(formula = Balance ~ Gender + Ethnicity, data = Credit)
##
## 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 ***
## GenderFemale          20.04      46.18   0.434    0.665
## EthnicityAsian       -19.37      65.11  -0.298    0.766
## EthnicityCaucasian   -12.65      56.74  -0.223    0.824
## ---
## 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".

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

```
##
## Call:
## lm(formula = Balance ~ Gender + Ethnicity + Income, data = Credit)
##
## 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 ***
## GenderFemale       24.3396    40.9630   0.594    0.553
## EthnicityAsian      1.6372    57.7867   0.028    0.977
## EthnicityCaucasian  6.4469    50.3634   0.128    0.898
## 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
```

```
#install.packages("stargazer")
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
```

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

```
stargazer(m_a, m_b, m_c, type = "text", out="text.txt")

##
##
================================================================================
======
##                                          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
```

Comments:

1. Model a(m_a) and Model b(m_b): None of them yield statistically significant coefficients at the standard $\alpha = 0.05$ level.

2. Model c(m_c) : When Income is added, its coefficient is highly significant (shown by 6.054***)

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

Answer :

In models a,b and c the coefficients for GenderFemale are 19.733, 20.038, and 24.340, respectively.

This means the model mathematically estimates that females have a slightly higher balance than males (who act as the baseline) by those exact dollar amounts, holding other variables in the respective models constant.

However, because none of these coefficients are statistically significant, there is no reliable evidence that gender actually affects the credit card balance in the broader population.

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

```
diff_b = predict(m_b, data.frame(Gender = " Male", Ethnicity = "African
American")) -
          predict(m_b, data.frame(Gender = " Male", Ethnicity = "Caucasian"))
diff_b

##        1
## 12.65305
```

Comment:

"African American" is the baseline for ethnicity.

To compare a male Caucasian to a male African American, we look at the EthnicityCaucasian coefficient in column 2 i.e, he coefficient is -12.653.

On average ,a male Caucasian has a credit card balance that is 12.653 dollars less than a male African American.

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

```
diff_c = predict(m_c, data.frame(Gender = " Male", Ethnicity = "African
American", Income = 100)) -
          predict(m_c, data.frame(Gender = " Male", Ethnicity = "Caucasian",
Income = 100))

diff_c
```

```
##         1
## -6.446938
```

Observation:

Model (3) controls for income. When income is held constant (i.e., both individuals earn exactly $100,000), the difference between the two demographic groups is found by looking at the EthnicityCaucasian coefficient in column (3).

The coefficient is 6.447.

Therefore, when both earn the exact same amount, a male Caucasian is estimated to have a balance 6.447 dollars more than a male African American.

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

Observation:

Model (2) difference: Caucasian is 12.653 dollars lower.

Model (3) difference: Caucasian is 6.447 dollars higher.

When we ignore income in Model (2), Caucasians appear to have lower balances. But when you control for income in Model (3), the direction of the effect completely flips. This happens because income is correlated with both ethnicity and balance.

This is termed as Simpson's Paradox.

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

```
pred_i <- predict(m_c, data.frame(Gender = "Female", Ethnicity = "Asian",
Income = 2000))
pred_i
```

```
##        1
## 12364.46
```

Model:

$$\text{Balance} = \text{Constant} + \text{GenderFemale} + \text{EthnicityAsian} + (\text{Income} \times 2000)$$

$$\text{Balance} = 230.029 + 24.340 + 1.637 + (6.054 \times 2000)\text{Balance} = 256.006 + 12108\text{Balance} = 12364.006$$

(j) Check the goodness of fit of the different models in (a) -(c) in terms of adjusted R^2. Which model would be prefererable ?

####Observations:

From looking at the "Adjusted R^2" row at the bottom of table:

Model (1): -0.002

Model (2): -0.007

Model (3): 0.208

Model (3) is the clearly preferred model. Models (1) and (2) have negative adjusted R^2 values, meaning that gender and ethnicity explain essentially 0% of the variance in balance (and the model penalizes them for being useless predictors). By adding Income, Model (3) is able to explain 20.8% of the variance in credit card balances.

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 x_{1i} from Normal(0,1) distribution,i=1,2,…,n

Step 2: Generate x_{2i} from Bernoulli (0.3) distribution,i = 1, 2, .., n

Step 3: Generate ϵ_i from Normal(0,1) and hence generate the response

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

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'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:

$(\beta_0, \beta_1, \beta_2, \beta_3)$ = (−2.5, 1.2, 2.3, 0.001) , (-2.5, 1.2. 2.3, 3.1). Set seed as 123.

```
set.seed(123)
n = 100
R = 1000

run_simulation = function(b0, b1, b2, b3) {
  mse_correct = numeric(R)
  mse_naive = numeric(R)
  for (i in 1:R) {
    x1 = rnorm(n, mean = 0, sd = 1)
    x2 = rbinom(n, size = 1, prob = 0.3)
    eps = rnorm(n, mean = 0, sd = 1)

    y = b0 + b1 * x1 + b2 * x2 + b3 * (x1 * x2) + eps
```

```r
    mod_correct = lm(y ~ x1 * x2)
    mod_naive = lm(y ~ x1 + x2)

    mse_correct[i] = mean(mod_correct$residuals^2)
    mse_naive[i] = mean(mod_naive$residuals^2)
  }

  return(c(Average_MSE_Correct = mean(mse_correct),
           Average_MSE_Naive = mean(mse_naive)))

}

results_config1 = run_simulation(-2.5, 1.2, 2.3, 0.001)
results_config2 = run_simulation(-2.5, 1.2, 2.3, 3.1)

print("Configuration 1 (Beta3 = 0.001):")

## [1] "Configuration 1 (Beta3 = 0.001):"

print(results_config1)

## Average_MSE_Correct   Average_MSE_Naive
##           0.9631944           0.9739083

print("Configuration 2 (Beta3 = 3.1):")

## [1] "Configuration 2 (Beta3 = 3.1):"

print(results_config2)

## Average_MSE_Correct   Average_MSE_Naive
##           0.9577982           2.8633349
```

The Impact of Ignoring the Interaction Term1.

First Configuration: $(\beta_0, \beta_1, \beta_2, \beta_3) = (-2.5, 1.2, 2.3, 0.001)$.

In this scenario, the true coefficient for the interaction term (β_3) is 0.001, which is almost exactly zero.

Average MSE Correct: ≈ 0.96

Average MSE Naive: ≈ 0.97

Impact: Because the actual interaction effect in the data generating process is practically non-existent, omitting it from the model has no meaningful impact. Both models fit the data equally well, with an MSE hovering just below the variance of the random error term ($\epsilon \sim N(0,1)$).

2. Second Configuration: $(\beta_0, \beta_1, \beta_2, \beta_3) = (-2.5, 1.2, 2.3, 3.1)$

In this scenario, the true coefficient for the interaction term (β_3) is 3.1, representing a very strong interaction between x_1 and x_2.

Average MSE Correct: ≈ 0.96

Average MSE Naive: ≈ 2.86

Impact: The correct model captures the interaction and maintains an expected MSE of approximately 1. The naive model, by ignoring the strong interaction term, experiences a massive spike in its Mean Squared Error. The variance introduced by the true interaction effect ($3.1 \times x_1 \times x_2$) is entirely absorbed into the naive model's residuals, making its predictions highly inaccurate.