Predictive Analytics Problem Set 3

746_Subhodeep Bhattacharjee

2026-02-10

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. Regress “balance” on

```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 4.2.3
```

```
data(Credit)

str(Credit)
```

```
## 'data.frame':    400 obs. of  12 variables:
##  $ ID       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Income   : num  14.9 106 104.6 148.9 55.9 ...
##  $ Limit    : int  3606 6645 7075 9504 4897 8047 3388 7114 3300 6819 ...
##  $ Rating   : int  283 483 514 681 357 569 259 512 266 491 ...
##  $ Cards    : int  2 3 4 3 2 4 2 2 5 3 ...
##  $ Age      : int  34 82 71 36 68 77 37 87 66 41 ...
##  $ Education: int  11 15 11 11 16 10 12 9 13 19 ...
##  $ Gender   : Factor w/ 2 levels " Male","Female": 1 2 1 2 1 1 2 1 2 2 ...
##  $ Student  : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 1 1 2 ...
##  $ Married  : Factor w/ 2 levels "No","Yes": 2 2 1 1 2 1 1 1 1 2 ...
##  $ Ethnicity: Factor w/ 3 levels "African American",..: 3 2 2 2 3 3 1 2 3
1 ...
##  $ Balance  : int  333 903 580 964 331 1151 203 872 279 1350 ...
```

(a) “gender” only.

```
m1 = lm(Balance ~ Gender, data = Credit)
summary(m1)
```

```
##
## 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
```

Interpretation

This model compares the mean balance of females vs males. By default, R uses Male as reference category or we can say baseline. The other values are compared relative to the baseline.

Intercept = mean balance for males

GenderFemale = difference (Female – Male)

(b) “gender” and “ethnicity” .

```
m2 = lm(Balance ~ Gender + Ethnicity, data = Credit)
summary(m2)
```

```
##
## 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
```

Baseline- Male and African American

(c) “gender”, “ethnicity”, “income”.

```
m3 = lm(Balance ~ Gender + Ethnicity + Income, data = Credit)
summary(m3)
```

```
##
## 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
```

Income here shows that it is major confounder in credit card balance.

(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(m1, m2, m3,
          type = "text",
          title = "Regression Results for Credit Balance",
          dep.var.labels = "Balance",
          column.labels = c("Gender only",
```

```
                           "Gender + Ethnicity",
                           "Gender + Ethnicity + Income"))

##
## Regression Results for Credit Balance
##
================================================================================
==========
##                                              Dependent variable:
##                      ---------------------------------------------------------
-------------
##                                                    Balance
##                           Gender only     Gender + Ethnicity  Gender +
Ethnicity + Income
##                              (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
```

Interpretation

Model 1 (GENDER) -

- GenderFemale is 19.733
- There are no significance stars hence the coefficient is statistically insignificant.
- R^2 = 0.0005
- It means the model based only on gender as a predictor explains almost no variability in the response variable.
- Hence the model is a very poor fit.

Model 2 (GENDER+ETHNICITY) -

- GenderFemale is 20.038
- EthnicityAsian is -19.371
- EthnicityCaucasian is -12.653
- Again none of the coefficients have any significance stars which implies none of the coefficients are statistically important.
- R^2 = 0.001
- It means that the second model also explains almost no variability in the response variable.
- The second model is a very poor fit too.

Model 3 (GENDER+ETHNICITY+INCOME) -

- GenderFemale is Not significant

- EthnicityAsian is Not significant
- EthnicityCaucasian is Not significant
- Income = 6.054* ($p < 0.01$) (highly significant)
- R^2 = 0.216 (substantial improvement)
- This model explains around 22% of variability in the response variable. Hence the model is a better fit than the last 2 models.

Observing the R-squared value we see that the model with Gender+Ethnicity+ Income is more better than others as it has highest value and captures the variation better.

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

Model (a): Gender effect usually weak
Model (b): Gender may still be insignificant Model (c): Gender often becomes insignificant after adding Income

Here we conclude that,Gender differences are largely explained by income differences.

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

```
coef(m2)["EthnicityCaucasian"]
```

```
## EthnicityCaucasian
##          -12.65305
```

Interpretation

Difference = coefficient of EthnicityCaucasian
A male Caucasian has this much lower average credit card balance than an African American 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).

```
coef(m3)["EthnicityCaucasian"]
```

```
## EthnicityCaucasian
##           6.446938
```

Income is held fixed i.e 100,000 dollars, so ethnicity difference remains same.

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

The comparison in (f) measures the difference in average balance between a male African and a male Caucasian without controlling for income. Here, we see that an African male has 12.653 units higher average balance than a Caucasian male.

In contrast, part (g) compares the two individuals while holding income fixed at $100,000. Here, we see that an African male has 6.445 units lower average balance than a Caucasian male.Since income significantly affects credit card balance, the estimate in (g) provides a more accurate measure of the pure effect of ethnicity. Therefore, the comparison in (g) is more reliable and economically meaningful.

Therefore, after controlling for income, the difference in balances across ethnic groups changes, indicating that income is an important confounding variable.

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

```
new_customer = data.frame(
  Gender = "Female",
  Ethnicity = "Asian",
  Income = 2000
)

predict(m3, new_customer)
```

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

Hence the Credit Card Balance of a Female Asian is 12364460 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?

On the basis of the adjusted R-squared value, we shall prefer model 3 as it is the highest among the three. So, the model with Gender+Ethnicity+Income is 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 $y_i = \beta_0 + +\beta_1 x_{1\,i} + \beta_2 x_{2\,i} + \beta_3 (x_{1\,i} \times x_{2\,i}) + \varepsilon_i$, 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.

```r
set.seed(123)
n=100
R=1000
simulate_mse = function(beta0, beta1, beta2, beta3) {

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)
    #true model
    y = beta0 + beta1*x1 + beta2*x2 + beta3*(x1*x2) + eps
    # Correct model (with interaction)
    model_correct = lm(y ~ x1 * x2)
    y_hat_correct = predict(model_correct)
    # Naive model (no interaction)
    model_naive = lm(y ~ x1 + x2)
```

```
    y_hat_naive = predict(model_naive)
    mse_correct[r] = mean((y - y_hat_correct)^2)
    mse_naive[r] = mean((y - y_hat_naive)^2)
  }

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

Case - 1 (β0, β1, β2, β3) = (−2.5, 1.2, 2.3, 0.001)

```
result1 = simulate_mse(-2.5, 1.2, 2.3, 0.001)
result1
```

```
## [1] 0.9631944 0.9739083
```

MSE_correct ≈ MSE_naive
Almost identical, because interaction is negligible.

Case - 2 (β0, β1, β2, β3) = (-2.5, 1.2. 2.3, 3.1)

```
result2 = simulate_mse(-2.5, 1.2, 2.3, 3.1)
result2
```

```
## [1] 0.9577982 2.8633349
```

MSE_correct << MSE_naive
Big difference, because ignoring interaction causes model misspecification.

```
comparison = data.frame(
  Case = c("Interaction ~ 0", "Strong Interaction"),
  MSE_Correct = c(result1[1], result2[1]),
  MSE_Naive = c(result1[2], result2[2])
)

comparison
```

```
##                 Case MSE_Correct MSE_Naive
## 1    Interaction ~ 0   0.9631944 0.9739083
## 2 Strong Interaction   0.9577982 2.8633349
```

When the interaction parameter is negligible, both models yield similar MSE values. However, when the interaction term is substantial, ignoring it significantly increases the MSE. This demonstrates that failing to include important interaction terms results in model misspecification and poorer predictive performance.