Problem Set 3

SAURABH MISHRA

2026-02-12

PREDICTIVE ANALYSIS

PROBLEM SET 3

Roll no.: 709

---

2) Problem to demonstrate the role of qualita-tive (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 4.5.2
```

```
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
```

```
data(Credit)
# Check structure
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 ...

Credit$Gender <- as.factor(Credit$Gender)
Credit$Ethnicity <- as.factor(Credit$Ethnicity)
```

(a) "gender" only.

```
model_a <- lm(Balance ~ Gender, data = Credit)
summary(model_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" .

```
model_b <- lm(Balance ~ Gender + Ethnicity, data = Credit)
summary(model_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".

```
model_c <- lm(Balance ~ Gender + Ethnicity + Income, data = Credit)
summary(model_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
```

(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(model_a, model_b, model_c,
          type = "text",
          title = "Regression Results",
          dep.var.labels = "Balance",
          covariate.labels = c("Female",
                               "African American",
                               "Asian",
                               "Income"),
          digits = 3)
```

```
##
## Regression Results
##
================================================================================
======
```

```
##                                        Dependent variable:
##                     ---------------------------------------------------
---------
##                                             Balance
##                            (1)                (2)                    (3)
## ------------------------------------------------------------------------
---------
## Female                    19.733             20.038                24.340
##                          (46.051)           (46.178)
(40.963)
##
## African American                             -19.371                1.637
##                                            (65.107)
(57.787)
##
## Asian                                        -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
```

Across the three models, gender is not statistically significant when explaining credit card balance. Adding ethnicity in model (b) does not change this result, as ethnicity is also typically insignificant. In model (c), income enters as a highly significant and positive predictor of balance. This indicates that higher income is strongly associated with higher credit card balances. Once income is included, gender and ethnicity remain statistically

insignificant. Therefore, income appears to be the main driver of credit card balance in this dataset.

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

In model (a), gender measures the simple difference in average balance between males and females, and it is typically not statistically significant. In model (b), gender measures the difference after controlling for ethnicity, and it remains insignificant. In model (c), gender measures the difference after controlling for both ethnicity and income, and it still remains insignificant. Thus, gender does not have a meaningful effect on balance in any of the models.

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

Using model (b), the difference in average balance between a male African American and a male Caucasian is equal to the coefficient on the African American variable. Since both individuals are male, the gender effect is zero for both, so it cancels out. Therefore, the comparison depends only on the ethnicity coefficient. If the coefficient is positive, the African American male has a higher average balance by that amount; if negative, a lower balance.

(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).

Using model (c), the difference between a male African American and a male Caucasian who both earn $100,000 is still equal to the coefficient on the African American variable. Since both individuals are male and have the same income, the effects of gender and income cancel out. Therefore, the difference in predicted balance depends only on the ethnicity coefficient and is the same as in model (b).

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

The answers in (f) and (g) are the same because in both cases the difference depends only on the coefficient of the African American variable. In (g), even though income is included in the model, both individuals earn the same amount, so the income effect cancels out. Since there is no interaction between income and ethnicity, the ethnic difference remains constant. Therefore, adding income does not change the comparison. (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?

Comparing using R^2, model (c) has the highest value, meaning it explains the largest proportion of variation in credit card balance. Model (b) has a slightly higher adjusted R^2 than model (a), but the improvement is small. Model (a) has the lowest adjusted R^2,

indicating the weakest explanatory power. Therefore, based solely on adjusted R^2, model (c) is preferred.

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.

```r
set.seed(123)

# Parameters
n <- 100
R <- 1000

# Function to run simulation for given beta values
run_simulation <- function(beta0, beta1, beta2, beta3) {

  mse_correct <- numeric(R)
  mse_naive   <- numeric(R)

  for (r in 1:R) {

    # Step 1: Generate x1
    x1 <- rnorm(n, 0, 1)

    # Step 2: Generate x2
    x2 <- rbinom(n, 1, 0.3)

    # Step 3: Generate epsilon and y
    epsilon <- rnorm(n, 0, 1)
```

```r
    y <- beta0 + beta1*x1 + beta2*x2 + beta3*(x1*x2) + epsilon

    # Step 4(i): Correct model (with interaction)
    model_correct <- lm(y ~ x1 * x2)
    mse_correct[r] <- mean(residuals(model_correct)^2)

    # Step 4(ii): Naive model (without interaction)
    model_naive <- lm(y ~ x1 + x2)
    mse_naive[r] <- mean(residuals(model_naive)^2)
  }

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

# Case 1: Small interaction
result1 <- run_simulation(-2.5, 1.2, 2.3, 0.001)

# Case 2: Large interaction
result2 <- run_simulation(-2.5, 1.2, 2.3, 3.1)

# Output results
results <- data.frame(
  Case = c("Small Interaction (0.001)", "Large Interaction (3.1)"),
  MSE_Correct_Model = c(result1[1], result2[1]),
  MSE_Naive_Model = c(result1[2], result2[2])
)

print(results)
```

```
##                        Case MSE_Correct_Model MSE_Naive_Model
## 1 Small Interaction (0.001)         0.9631944       0.9739083
## 2   Large Interaction (3.1)         0.9577982       2.8633349
```

When the interaction coefficient is very small ($\beta_3 = 0.001$), the average MSE of the correct and naive models is almost the same, showing that ignoring the interaction has little impact. However, when the interaction is large ($\beta_3 = 3.1$), the naive model (without interaction) has a much higher average MSE than the correct model. This indicates substantial loss of accuracy due to model misspecification. Therefore, ignoring an important interaction term can significantly worsen model performance.