

# Problem Set 3

SIMLIN SAHA

2026-02-12

## PREDICTIVE ANALYSIS

### PROBLEM SET 3

Roll no.: 731

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 4.5.2

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
```



```

##          -----
-----
##                               Balance
##                               (2)
##          (1)                               (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 all three models, gender does not have a statistically significant effect on credit card balance. Including ethnicity in model (b) does not alter this conclusion, as ethnicity also shows no meaningful statistical impact. In model (c), income emerges as a highly significant and positive predictor of balance, indicating a strong association between higher income and larger credit card balances. Even after accounting for income, both

gender and ethnicity continue to remain statistically insignificant. This suggests that income is the primary factor influencing credit card balance in this dataset.

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

In model (a), gender captures the basic difference in average credit card balance between males and females and is generally not statistically significant. In model (b), after adjusting for ethnicity, the effect of gender continues to be insignificant. In model (c), even after controlling for both ethnicity and income, gender still does not show a significant impact. Overall, gender does not appear to have a meaningful influence on credit card balance across 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 given by the coefficient on the African American indicator. Because both individuals are male, the gender term is the same for each and therefore drops out of the comparison. As a result, the difference depends solely on the ethnicity coefficient: a positive value implies that the African American male has a higher average balance by that amount, while a negative value implies a lower average 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 in predicted balance between a male African American and a male Caucasian, both earning \$100,000, is again determined by the coefficient on the African American indicator. Because both individuals share the same gender and income, the effects of these variables offset each other. Consequently, the comparison depends only on the ethnicity coefficient and is identical to that in model (b)

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

The conclusions in parts (f) and (g) are identical because, in both cases, the difference in predicted balance depends solely on the coefficient of the African American indicator. In part (g), although income is included in the model, both individuals have the same income level, so the income effect offsets itself. Since the model does not include any interaction terms between income and ethnicity, the effect of ethnicity remains unchanged. As a result, adding income does not alter the comparison.

**(j) Check the goodness of fit of the different models in (a) -(c) in terms of AIC,BIC and adjusted R<sup>2</sup>.Which model would you prefer?**

Based on the  $R^2$  values, model (c) performs the best, as it explains the largest proportion of variation in credit card balance. Although model (b) has a slightly higher adjusted  $R^2$  than model (a), the improvement is marginal. Model (a) shows the lowest adjusted  $R^2$ , indicating the weakest explanatory power among the three. Therefore, considering adjusted  $R^2$  as the evaluation criterion, model (c) is the preferred model.

#### 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,  $i = 1, 2, \dots, n$

**Step 2:** Generate  $x_{2i}$  from Bernoulli (0.3) distribution,  $i = 1, 2, \dots, n$

**Step 3:** Generate  $\epsilon_i$  from Normal(0,1) and hence generate the response  $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3(x_{1i} \times x_{2i}) + \epsilon_i$ ,  $i = 1, 2, \dots, n$ .

**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)

# 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)
    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 both the correct and naive models is nearly identical, indicating that omitting the interaction term has negligible impact on model performance. In contrast, when the interaction coefficient is large ( $\beta_3 = 3.1$ ), the naive model (which excludes the interaction) exhibits a substantially higher average MSE than the correct model. This demonstrates a significant loss of predictive accuracy due to model misspecification. Hence, ignoring a meaningful interaction term can severely degrade model performance.