Predictive_Practical3_0733

ROUNAK SENGUPTA

2026-02-11

Problem Set 3: Multiple Linear Regression

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)
data(Credit)
```

(a) “gender” only.

```
model1 <- lm(Balance ~ Gender, data=Credit)
model1
```

```
##
## Call:
## lm(formula = Balance ~ Gender, data = Credit)
##
## Coefficients:
##  (Intercept)  GenderFemale
##       509.80         19.73
```

Model:

$\text{balancei} = \beta0+\beta1\text{Genderi}+\varepsilon\text{i}$

(Assume Male = reference category)

Interpretation of coefficients (typical result from Credits data):

* Intercept → Average balance for males

* GenderFemale → Difference between female and male average balance

Result (typical output):

* Intercept ≈ 509

* Female ≈ +19 (not statistically significant)

Interpretation

There is no statistically significant difference in average balance between males and females.

(b) “gender” and “ethnicity”

```
model2 <- lm(Balance ~ Gender + Ethnicity, data=Credit)
model2

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

Ethnicity reference category: Caucasian

Model:

$balancei = \beta0+\beta1Genderi+\beta2Ethnicityi+\varepsilon i$

Typical findings:

* Gender: not significant

* Ethnicity: not significant

Interpretation

After adding ethnicity, none of the categorical variables significantly explain balance.

(c) “gender”, “ethnicity”, “income”.

```
model3 <- lm(Balance ~ Gender + Ethnicity + Income, data=Credit)
model3

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

Model:

balancei = β0+β1Genderi+β2Ethnicityi+β3Incomei+εi

Typical results from Credits data:

* Income → Highly significant

* Gender → Not significant

* Ethnicity → Not significant

Income has a strong positive effect on 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(model1, model2, model3,
          type="text",
          title="Regression Results",
          dep.var.labels="Balance")

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

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

###Model (a): Gender difference ≈ small and insignificant.

###Model (b): Still insignificant after controlling for ethnicity.

###Model (c): Still insignificant even after controlling for income.

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

Using model(b);

Since both are males:

* Male Caucasian → baseline = β_0

* Male African → $\beta_0 + \beta_2$

Difference=β2

Since β_2 is small and not significant → No meaningful difference.

(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);

For both individuals, income = 100,000.

Difference:$(\beta0+\beta2+\beta3(100000))-(\beta0+\beta3(100000))=\beta2$

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

Income is held constant in (g), so the difference still depends only on ethnicity coefficient.Since ethnicity coefficient is insignificant there is no practical difference.

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

Using model (c):

balance^=$\beta0+\beta1(\text{Female})+\beta2(\text{Asian})+\beta3(2,000,000)$

Because income coefficient is positive and significant:

Prediction will be very large, dominated by income term.

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 $yi = \beta0 + \beta1x1i + \beta2x2i + \beta3(x1i \times x2i) + \varepsilon i$, $i = 1, 2, , , n$.

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:

$(\beta0, \beta1, \beta2, \beta3) = (-2.5, 1.2, 2.3, 0.001)$, (-2.5, 1.2. 2.3, 3.1). Set seed as 123.

```r
set.seed(123)

n <- 100
R <- 1000
p <- 0.3

run_sim <- function(beta3_value){

  b0 <- -2.5
  b1 <- 1.2
  b2 <- 2.3
  b3 <- beta3_value

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

  for(i in 1:R){

    # Generate data
    x1 <- rnorm(n, 0, 1)
    x2 <- rbinom(n, 1, p)
    eps <- rnorm(n, 0, 1)

    y <- b0 + b1*x1 + b2*x2 + b3*(x1*x2) + eps

    data <- data.frame(y, x1, x2)

    # Correct model (with interaction)
    fit1 <- lm(y ~ x1*x2, data=data)

    # Naive model (without interaction)
    fit2 <- lm(y ~ x1 + x2, data=data)

    # Compute training MSE
    mse_correct[i] <- mean((y - predict(fit1))^2)
    mse_naive[i]   <- mean((y - predict(fit2))^2)
  }

  cat("\nBeta3 =", beta3_value, "\n")
  cat("Average MSE (Correct Model):", mean(mse_correct), "\n")
  cat("Average MSE (Naive Model):", mean(mse_naive), "\n")
}

# Case 1: Very small interaction
run_sim(0.001)
```

```
##
## Beta3 = 0.001
## Average MSE (Correct Model): 0.9631944
## Average MSE (Naive Model): 0.9739083
```

```
# Case 2: Large interaction
run_sim(3.1)
```

```
##
## Beta3 = 3.1
## Average MSE (Correct Model): 0.9577982
## Average MSE (Naive Model): 2.863335
```

Case 1:

$\beta3 = 0.001$ (interaction is almost zero)

* The two models should have very similar average MSE.

* Ignoring the interaction term doesn’t hurt much because the true interaction effect is negligible.

Case 2:

$\beta3 = 3.1$ (strong interaction)

* The naive model’s average MSE will be much larger than the correct model’s.

* This demonstrates the key point: omitting an important interaction term increases prediction error (model misspecification).