Predictive Analysis Problem Set 2

Deeptarka Saha 732

2026-02-12

PROBLEM SET 2

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

```
library(ISLR)
## Warning: package 'ISLR' was built under R version 4.3.3
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
data(Credit)
attach(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) Regress balance on gender only
fit1 <- lm(Balance ~ Gender, data = Credit)
summary(fit1)
##
## 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) Regress balance on gender and ethnicity
fit2 <- lm(Balance ~ Gender + Ethnicity, data = Credit)
summary(fit2)
##
## 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) Regress balance on gender, ethnicity and income
fit3 <- lm(Balance ~ Gender + Ethnicity + Income, data = Credit)
summary(fit3)
##
## 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 regressions in one table
stargazer(fit1, fit2, fit3,
          type = "text",
          title = "Regression Results",
          dep.var.labels = "Credit Card Balance",
          covariate.labels = c("Male",
                               "Asian",
                               "Caucasian",
                               "Income"),
```

```
          omit.stat = c("f", "ser"))
##
## Regression Results
## ============================================
##                     Dependent variable:
##              ------------------------------
##                     Credit Card Balance
##                  (1)         (2)         (3)
## --------------------------------------------
## Male            19.733      20.038      24.340
##                (46.051)    (46.178)    (40.963)
##
## Asian                      -19.371      1.637
##                            (65.107)    (57.787)
##
## Caucasian                  -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
## ============================================
## Note:             *p<0.1; **p<0.05; ***p<0.01
```

(e)How Gender affects Balance in model (a):

Gender coefficient shows the difference in average balance between males and females.

If positive and significant → males carry higher balance than females.

model (b):

Gender effect is now controlled for ethnicity.

model (c):

Gender effect is controlled for ethnicity and income.

Usually becomes statistically insignificant.

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

```
coef(fit2)
##        (Intercept)       GenderFemale     EthnicityAsian EthnicityCaucasian
##          520.87967           20.03825          -19.37088          -12.65305
#balance_estimate_african=beta_not + beta1
avg.balance.male.african=531.00+46.01;
avg.balance.male.african
## [1] 577.01
#balance_estimate_caucasian=beta_not + beta1 + beta3
avg.balance.male.caucasian=531.00+46.01+12;
avg.balance.male.caucasian
## [1] 589.01
# therefore diff = beta3 = 12
```

So the difference equals the coefficient of Caucasian.If β_3 is small and insignificant → 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).
balance.male.african_amr=230.029 + 6.054*100000
balance.male.african_amr
## [1] 605630
balance.male.caucasian= 230.029 + 6.054*100000 + 6.447
balance.male.caucasian
## [1] 605636.5
diff1=6.447
```

Ans: Its the coefficient of Caucasian

h. The difference remains β_3 in both cases.

Adding income does NOT change the difference when income is fixed equal.

This suggests part of ethnicity difference is explained by income.

i. Predict balance for Female Asian with income = 2,000,000

For model (c):

```
newdata = data.frame(
  Gender = "Female",
  Ethnicity = "Asian",
  Income = 2000000
```

```
)
predict(fit3, newdata)
##        1
## 12108706
```

j. Compare Adjusted R^2

```
summary(fit1)$adj.r.squared
## [1] -0.002050271
summary(fit2)$adj.r.squared
## [1] -0.006876514
summary(fit3)$adj.r.squared
## [1] 0.207774
```

We observe :

Model (a): Very low Adjusted R^2

Model (b): Slight improvement

Model (c): Much higher Adjusted R^2

Therefore Model (c) is the preferred here because:

Highest Adjusted R^2

Income is highly significant

Better explanatory power

Problem 4: Problem to demonstrate the impact of ignoring interaction term in multiple linear regression.

```
set.seed(123)

# Simulation settings
n = 100
R = 1000

simulate_mse = function(beta0, beta1, beta2, beta3) {

  mse_correct = numeric(R)
  mse_naive = numeric(R)

  for (r in 1:R) {
```

```
    # Step 1
    x1 = rnorm(n, 0, 1);
    x1

    # Step 2
    x2 = rbinom(n, 1, 0.3);
    x2

    # Step 3
    eps = rnorm(n, 0, 1)

    y = beta0 + beta1*x1 + beta2*x2 + beta3*(x1*x2) + eps;
    y

    # Step 4(i): Correct model
    fit1 = lm(y ~ x1 * x2)
    yhat1 = predict(fit1)
    mse_correct[r] = mean((y - yhat1)^2);
    mse_correct

    # Step 4(ii): Naive model (no interaction)
    fit2 = lm(y ~ x1 + x2)
    yhat2 = predict(fit2)
    mse_naive[r] = mean((y - yhat2)^2);
    mse_naive
  }

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

# Case 1: beta3 = 0.001 (almost no interaction)
res1 = simulate_mse(-2.5, 1.2, 2.3, 0.001)

# Case 2: beta3 = 3.1 (strong interaction)
res2 = simulate_mse(-2.5, 1.2, 2.3, 3.1)

res1
## [1] 0.9631944 0.9739083
```

```
res2
```

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

Final Interpretation :

1. When the interaction coefficient is nearly zero, omitting the interaction term does not materially affect MSE.
2. When the interaction effect is large, ignoring it substantially increases MSE.
3. Therefore, excluding important interaction terms leads to model misspecification and poorer predictive performance.