predictive_ps3_735

Srinjana Sen_735

2026-02-12

PREDICTIVE ANALYTICS

Problem Set 3: Multiple Linear Regression

Q2. Problem to demonstrate the role of qualitative (nominal) predictors in addition to quantitative predictors in multiple linear regression

Attach “Credits” data from R.

```
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")
head(Credit)
```

```
##   ID  Income Limit Rating Cards Age Education Gender Student Married
Ethnicity
## 1  1  14.891  3606    283     2  34        11   Male      No     Yes
Caucasian
## 2  2 106.025  6645    483     3  82        15 Female     Yes     Yes
Asian
## 3  3 104.593  7075    514     4  71        11   Male      No      No
Asian
## 4  4 148.924  9504    681     3  36        11 Female      No      No
Asian
## 5  5  55.882  4897    357     2  68        16   Male      No     Yes
Caucasian
## 6  6  80.180  8047    569     4  77        10   Male      No      No
Caucasian
##   Balance
```

```
## 1     333
## 2     903
## 3     580
## 4     964
## 5     331
## 6    1151
```

Regress “balance” on
(a) “gender” only.

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

```
##
## 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” .

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

```
##
## 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".

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

```
##
## 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(ma, mb, mc,
          type = "text",
          title = "Regression Models for Balance",
          dep.var.labels = "Balance",
          covariate.labels = c("Male", "Asian", "Caucasian", "Income"))
```

```
##
## Regression Models for Balance
##
================================================================================
======
```

```
##                                          Dependent variable:
##                      ---------------------------------------------------
---------
##                                                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
## 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
```

Comment on Significant Coefficients:

* Gender often becomes insignificant once Income is included.
* Income is usually highly significant.
* Ethnicity sometimes significant depending on baseline.

(e) Explain how gender affects “balance” in each of the models (a)-(c).
Model (a):
Balance difference between Male and Female.
Model (b):
Gender effect adjusted for ethnicity.
Model (c):
Gender effect adjusted for ethnicity + income.
Thus, if Gender becomes insignificant thendifference explained through income instead.

(f) Compare the average credit card balance of a male African with a male Caucasian on the basis of model (b).
In model (b), baseline ethnicity = African American
Thus, the difference between Male Caucasian and Male African in model (b) is simply:
B3=EthnicityCaucasian coefficient

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

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

(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).

```
levels(Credit$Gender)
```

```
## [1] " Male"  "Female"
```

```
levels(Credit$Ethnicity)
```

```
## [1] "African American" "Asian"            "Caucasian"
```

```
new_african=data.frame(
  Gender = factor("Male", levels = levels(Credit$Gender)),
  Ethnicity = factor("African American", levels = levels(Credit$Ethnicity)),
  Income = 100
)

new_caucasian=data.frame(
  Gender = factor("Male", levels = levels(Credit$Gender)),
  Ethnicity = factor("Caucasian", levels = levels(Credit$Ethnicity)),
  Income = 100
)


predict(mc, new_african)
```

```
##  1
## NA
```

```
predict(mc, new_caucasian)
```

```
##  1
## NA
```

(h) Compare and comment on the answers in (f) and (g)
* Model (b) ignores income so ethnicity may look more important.
* Model (c) adjusts for income so ethnicity effect may shrink.
So income is a confounder.

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

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

predict(mc, new_person)
```

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

(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?

```
AIC(ma, mb, mc)
```

```
##    df      AIC
## ma  3 6044.527
## mb  5 6048.434
## mc  6 5953.518
```

```
BIC(ma, mb, mc)
```

```
##    df      BIC
## ma  3 6056.501
## mb  5 6068.391
## mc  6 5977.466
```

```
summary(ma)$adj.r.squared
```

```
## [1] -0.002050271
```

```
summary(mb)$adj.r.squared
```

```
## [1] -0.006876514
```

```
summary(mc)$adj.r.squared
```

```
## [1] 0.207774
```

Preferred modelhas the following characteristics:
Lowest AIC/BIC
Highest Adjusted R^2
Thus, model (c) is best.

Q4. 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.

```
set.seed(123)

MSE=function(n, b0, b1, b2, b3, R=1000){

  correct_mse=numeric(R)
  naive_mse=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 error and response
    ei=rnorm(n, 0, 1)

    y=b0 + b1*x1 + b2*x2 + b3*(x1*x2) + ei

    data=data.frame(y, x1, x2)

    # Step 4(i): Correct model with interaction
    correct_fit=lm(y ~ x1 * x2, data=data)

    # Step 4(ii): Naive model without interaction
    naive_fit=lm(y ~ x1 + x2, data=data)

    # Predictions
    correct_pred=predict(correct_fit, data)
    naive_pred=predict(naive_fit, data)

    # Compute MSE
    correct_mse[r]=mean((y - correct_pred)^2)
    naive_mse[r]=mean((y - naive_pred)^2)
  }

  return(c(mean(correct_mse), mean(naive_mse)))
}

result1=MSE(n=100,
                        b0=-2.5,
                        b1=1.2,
                        b2=2.3,
                        b3=0.001)

result1
```

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

Correct Model MSE and Naive Model MSE are approximately equal i.e. interaction is negligible

```
result2=MSE(n=100,
                        b0=-2.5,
                        b1=1.2,
                        b2=2.3,
                        b3=3.1)

result2
```

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

Correct Model MSE much smaller than Naive Model MSE i.e. if interaction is ignored, it causes major prediction error