Predictive Analysis Problem Set 3

Multiple Linear Regression

Atrijo Roy - 736

2026-02-12

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

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

(a) "gender" only.

(b) "gender" and "ethnicity" .

(c) "gender", "ethnicity", "income".

(d) Output all the regressions in (a)-(c) in a single table using stargazer. Comment on the significant coefficients in each of the models.

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

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

(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).

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

(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 R^2. Which model would you prefer?

```
rm(list=ls())
library(ISLR)

## Warning: package 'ISLR' was built under R version 4.5.1

attach(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
```

Now we regress Balance on Gender

```
#(a)
fit1=lm(Balance~Gender)
summary(fit1)

##
## Call:
## lm(formula = Balance ~ Gender)
##
## 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
```

The fitted model is given by

Balance.hat = 509.8 + 19.73GenderFemale

If Gender is male, predicted balance is 509.8 units on an average.

If gender is female, predicted balance is 529.53 units on an average.

Now we regress Balance on Gender and Ethnicity

```
#(b)
fit2=lm(Balance~Gender+Ethnicity)
summary(fit2)

##
## Call:
## lm(formula = Balance ~ Gender + Ethnicity)
##
## 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
```

Interpretation

Fitted model:

Balance.hat = 520.88 + 20.44GenderFemale -19.37EthnicityAsian -12.65EthnicityCaucasian

If Gender is male and Ethnicity is African American predicted balance is 520.88 units on an average.

Now we regress Balance on Ethnicity, Gender and Income

```r
#(c)
fit3=lm(Balance~Gender+Ethnicity+Income)
summary(fit3)
```

```
##
## Call:
## lm(formula = Balance ~ Gender + Ethnicity + Income)
##
## 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
```

Fitted model:

Balance.hat = 230.0291 + 24.3396GenderFemale + 1.6372EthnicityAsian + 6.4469EthnicityCaucasian + 6.0542Income

If income is zero, if Gender is male and Ethnicity is African American predicted balance is 230.0291 units on an average.

Now we show all the three models together using stargazer.

```r
#(d)
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
```

```r
stargazer(fit1,fit2,fit3,type="text",out="f2.txt")
```

```
##
##
================================================================================
======
##                                          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
```

Model (a)

Coefficient of GenderFemale = 19.733.

Interpretation: Females have, on average, about 19.7 units higher balance than males. However, the coefficient is not statistically significant (large SE 46.051), so gender has no meaningful effect in this model.

Model (b)

Coefficient = 20.038.

Interpretation: After controlling for ethnicity, females have about 20 units higher balance than males, again not statistically significant.

Model (c)

Coefficient = 24.340.

Interpretation: After controlling for ethnicity and income, females have about 24.3 units higher balance than males. The effect is still not statistically significant, indicating gender does not significantly affect balance in any model.

(f) Male African vs Male Caucasian using model (b)

Model (b) coefficient for EthnicityCaucasian = −12.653.

Both are males, so gender cancels out; African is the baseline (0), Caucasian has −12.653.

Difference (Caucasian − African) = −12.653.

Thus, a male Caucasian has about 12.65 units lower average balance than a male African (not statistically significant).

(g) Male African vs Male Caucasian when income = 100,000 (model (c))

Model (c) coefficient for EthnicityCaucasian = 6.447.

Income is the same for both individuals, so the income term cancels out.

Difference (Caucasian − African) = 6.447.

Thus, a male Caucasian has about 6.45 units higher balance than a male African american.

(h) Compare (f) and (g)

In model (b), the difference was −12.653 (Caucasians lower).

In model (c), after controlling for income, the difference becomes +6.447 (Caucasians slightly higher).

This change shows that income is a significant variable; once income is included, the ethnicity comparison changes substantially, suggesting that some of the difference observed in model (b) was due to income differences rather than ethnicity itself.

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

Model:

$$\hat{y} = 236.029 + 24.340(\text{Female}) + 1.637(\text{Asian}) + 6.054(\text{Income})$$

```
Female=1
Asian=1
Income=200000
y.hat=236.029 +
      24.340*Female +
      1.637*Asian +
      6.054*Income

y.hat
```

```
## [1] 1211062
```

We get the predicted balance as 1211062 dollars

(j) Check the goodness of fit of the different models in (a)-(c)

Goodness-of-fit based on (R^2) and Adjusted (R^2)

Model (1):

(R^2 = 0.0005), Adjusted (R^2 = -0.002).

This indicates the model does not explain almost any variation in the response; gender alone explains none of the variation in credit card balance. The negative adjusted (R^2) suggests that the model performs no better than a model containing only the mean.

Model (2):

(R^2 = 0.001), Adjusted (R^2 = -0.007).

After including ethnicity, the model still explains almost none of the variability in balance, indicating very poor goodness of fit.

Model (3):

(R^2 = 0.216), Adjusted (R^2 = 0.208).

Including income substantially improves the goodness of fit, with approximately 21% of the variation in credit card balance explained by the predictors. Although the fit improves

considerably compared to Models (1) and (2), the explanatory power is still moderate rather than strong.

Overall conclusion:

Models (1) and (2) show very poor fit, whereas Model (3) provides a noticeably better fit because income is an significant predictor of credit card balance.

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

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, $i = 1, 2, .., n$

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

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

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.

We perform step 1,2 and 3 for one iteration

```
rm(list=ls())
set.seed(123)
beta1=c(-2.5,1.2,2.3,0.001)
beta2=c(-2.5,1.2,2.3,3.1)
#Step 1
n=100
x1=rnorm(n)
#Step 2
x2=rbinom(n,1,0.3)
#Step 3
e=rnorm(n)
y1=beta1[1]+beta1[2]*x1+beta1[3]*x2+beta1[4]*(x1*x2)+e
y2=beta2[1]+beta2[2]*x1+beta2[3]*x2+beta2[4]*(x1*x2)+e
#Outputs of response y for two sets of beta values
cbind(y1,y2)
```

```
##                   y1           y2
##   [1,] -2.384831928 -2.384831928
##   [2,]  0.292599076 -0.420720964
##   [3,] -0.297347444 -0.297347444
##   [4,] -3.423766539 -3.423766539
##   [5,] -2.464307324 -2.464307324
##   [6,]  1.579397714  6.894384108
##   [7,] -1.383911020 -1.383911020
##   [8,] -4.390512238 -4.390512238
##   [9,] -2.347250036 -2.347250036
##  [10,] -3.409375222 -3.409375222
##  [11,]  0.021609623  0.021609623
##  [12,] -3.117400414 -3.117400414
##  [13,] -3.279229504 -3.279229504
##  [14,]  0.873859194  0.873859194
##  [15,] -3.583866950 -3.583866950
##  [16,]  2.244310269  7.781954080
##  [17,] -1.266009752 -1.266009752
##  [18,] -5.343721214 -5.343721214
##  [19,]  1.159190482  3.332692421
##  [20,] -0.398857953 -1.864038526
##  [21,] -3.996768955 -3.996768955
##  [22,] -0.396494839 -1.071999100
##  [23,] -1.466298596 -4.645886381
##  [24,] -1.246217576 -1.246217576
##  [25,] -3.991383218 -3.991383218
##  [26,] -5.620028240 -5.620028240
##  [27,] -1.456867147 -1.456867147
##  [28,] -2.005471509 -2.005471509
##  [29,] -3.429240846 -3.429240846
##  [30,]  0.847466387  4.733038828
##  [31,] -3.051569068 -3.051569068
```

```
##  [32,] -1.590900604 -1.590900604
##  [33,] -1.775499595 -1.775499595
##  [34,] -2.311752678 -2.311752678
##  [35,] -1.750382271 -1.750382271
##  [36,] -1.870807589 -1.870807589
##  [37,] -0.725378526 -0.725378526
##  [38,] -0.189618672 -0.381483063
##  [39,] -2.113101411 -2.113101411
##  [40,] -1.156237689 -2.335317322
##  [41,] -3.119203065 -3.119203065
##  [42,] -3.074186645 -3.074186645
##  [43,] -3.923892094 -3.923892094
##  [44,] -0.792616200 -0.792616200
##  [45,] -2.361247135 -2.361247135
##  [46,]  0.448359976 -3.032153523
##  [47,] -2.382752979 -2.382752979
##  [48,] -2.011724441 -3.457889382
##  [49,]  0.125572190  2.542684092
##  [50,] -1.485606333 -1.743967070
##  [51,]  0.002792566  0.002792566
##  [52,] -1.221843130 -1.221843130
##  [53,] -2.816589605 -2.816589605
##  [54,] -0.314483200 -0.314483200
##  [55,] -3.185265131 -3.185265131
##  [56,]  1.145034301  5.844576704
##  [57,] -5.147106203 -5.147106203
##  [58,] -2.393080768 -2.393080768
##  [59,] -0.700467440 -0.700467440
##  [60,]  0.005317699  0.674520621
##  [61,]  0.375192255  1.551695012
##  [62,] -0.559603038 -2.116303419
##  [63,] -1.667372982 -1.667372982
##  [64,] -1.939372866 -5.095937978
##  [65,] -4.778656622 -4.778656622
##  [66,] -0.460068698 -0.460068698
##  [67,] -2.403311483 -2.403311483
##  [68,] -3.159460898 -3.159460898
##  [69,] -2.629552157 -2.629552157
##  [70,] -1.324614100 -1.324614100
##  [71,] -1.363701910 -2.885407493
##  [72,] -4.653016834 -4.653016834
##  [73,] -0.183265632 -0.183265632
##  [74,] -2.643452561 -2.643452561
##  [75,] -1.389955645 -3.522094348
##  [76,]  1.091461152  4.269706827
##  [77,] -1.246608845 -2.129120394
##  [78,] -4.682079416 -4.682079416
##  [79,] -1.397785325 -1.397785325
##  [80,] -3.682262214 -3.682262214
##  [81,]  1.762216753  1.780079965
```

```
##  [82,] -2.127983113 -2.127983113
##  [83,] -2.730253212 -2.730253212
##  [84,] -2.465275847 -2.465275847
##  [85,] -3.338972564 -3.338972564
##  [86,] -3.418877776 -3.418877776
##  [87,] -1.366718573 -1.366718573
##  [88,] -1.558799806 -1.558799806
##  [89,] -2.566813558 -2.566813558
##  [90,] -1.902967345 -1.902967345
##  [91,] -2.096417344 -2.096417344
##  [92,] -2.344122367 -2.344122367
##  [93,] -0.717461248 -0.717461248
##  [94,] -2.091418818 -4.037299748
##  [95,]  1.255091996  5.471753934
##  [96,]  0.981450058 -0.878754403
##  [97,]  2.326012039  9.104556985
##  [98,] -2.020707952 -2.020707952
##  [99,] -3.447609866 -3.447609866
## [100,] -0.947271522 -4.128149892
```

```
#Regression models (For the first set of beta values)
fit1=lm(y1~x1+x2+x1*x2) #with the interaction term
fit2=lm(y1~x1+x2) #without the interaction term
library(stargazer)
stargazer(fit1,fit2,type="text",out="f1.txt")
```

```
##
## ===============================================================
##                                 Dependent variable:
##                     ------------------------------------------
##                                         y1
##                                (1)                  (2)
## ---------------------------------------------------------------
## x1                          1.070***             1.071***
##                             (0.125)               (0.103)
##
## x2                          2.306***             2.306***
##                             (0.209)               (0.207)
##
## x1:x2                        0.003
##                             (0.225)
##
## Constant                   -2.475***            -2.475***
##                             (0.112)               (0.111)
##
## ---------------------------------------------------------------
## Observations                  100                  100
## R2                           0.711                0.711
## Adjusted R2                  0.702                0.705
## Residual Std. Error    0.942 (df = 96)      0.937 (df = 97)
```

```
## F Statistic         78.588*** (df = 3; 96) 119.110*** (df = 2; 97)
## ========================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

Interpretation:

R^2 and Adjusted R^2 are identical for both models (0.711).

Residual standard error is also nearly the same (0.942 vs 0.937).

This indicates that adding the interaction term provides almost no improvement in the multiple linear regression.

Now for 1000 iterations:

```
#Step 4: Repeating step 1,2 and 3 for R=1000 times
rm(list=ls())
set.seed(123)
f=function(beta,n,R)
{
  mse.correct=c()
  mse.naive=c()
  for(i in 1:R)
  {
    x1=rnorm(n)
    x2=rbinom(n,1,0.3)
    e=rnorm(n)
    y=beta[1]+beta[2]*x1+beta[3]*x2+beta[4]*(x1*x2)+e
    fit1=lm(y~x1+x2+x1*x2)
    mse.correct[i]=mean((y-predict(fit1))^2)
    fit2=lm(y~x1+x2)
    mse.naive[i]=mean((y-predict(fit2))^2)
  }
  c(avg.correct.mse=mean(mse.correct),avg.naive.mse=mean(mse.naive))
}
```

Now for the first set of beta values

```
beta1=c(-2.5,1.2,2.3,0.001)
f(beta1,100,1000)
```

```
## avg.correct.mse   avg.naive.mse
##       0.9631944       0.9739083
```

Interpretation:

The two MSE values are almost identical because the true interaction coefficient ($\beta_3 = 0.001$) is essentially zero. Hence, ignoring the interaction term produces almost similar predicted response. Both models perform similarly.

And for the second set of beta values

```r
beta2=c(-2.5,1.2,2.3,3.1)
f(beta2,100,1000)

## avg.correct.mse   avg.naive.mse
##       0.9577982       2.8633349
```

Interpretation:

When the interaction effect has a large coefficient ($\beta_3 = 3.1$), the naive model that omits the interaction term has a much larger MSE, showing substantial loss of model accuracy. The correct model, which includes the interaction, maintains low MSE. This clearly demonstrates the serious impact of ignoring an important interaction term in multiple linear regression.

Conclusion:

The simulation shows that when the interaction effect is negligible, excluding the interaction term has little impact on model performance. However, when the interaction effect is strong, ignoring it leads to a substantial increase in prediction error, confirming the importance of including relevant interaction terms in regression models.