Predictive_Assignment_3

Oindrila Chakraborty

2026-02-08

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 adjusted R2. Which model would you prefer?

```
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)
fit1=lm(Balance ~ Gender, data = Credit)
fit2=lm(Balance ~ Gender + Ethnicity, data = Credit)
fit3=lm(Balance ~ Gender + Ethnicity + Income, data = Credit)
tab=data.frame(
  Model=c("Gender only", "Gender + Ethnicity", "Gender + Ethnicity +
Income"),
```

```
  R_squared=c(summary(fit1)$r.squared,
              summary(fit2)$r.squared,
              summary(fit3)$r.squared),
  Adjusted_R_squared = c(summary(fit1)$adj.r.squared,
                         summary(fit2)$adj.r.squared,
                         summary(fit3)$adj.r.squared)
)
print(tab)
```

```
##                          Model   R_squared Adjusted_R_squared
## 1                 Gender only 0.000461133       -0.002050271
## 2          Gender + Ethnicity 0.000693986       -0.006876514
## 3 Gender + Ethnicity + Income 0.215716091        0.207773976
```

d.Commenting on the significant coefficients in each model. From the regression results, in model (a), the coefficient of Gender is statistically significant,indicating that gender has a significant effect on credit card balance when considered alone.In model (b), Ethnicity becomes significant for some categories, while the effect of Gender may change in magnitude, showing that ethnicity explains additional variation in balance. In model (c), Income is highly statistically significant, and after controlling for income, the significance of gender and ethnicity changes, suggesting that income is a major determinant of credit card balance.

e.Explaining how gender affects "balance" in each of the models (a)–(c). In model (a), gender directly affects credit card balance, with the coefficient indicating the average difference in balance between males and females.In model (b), after controlling for ethnicity, the effect of gender changes, implying that part of the gender effect observed in model (a) was influenced by ethnic differences.In model (c), once income is included, the gender effect further reduces, showing that income explains a substantial portion of the variation previously attributed to gender.

f.Comparing the average credit card balance of a male African with a male Caucasian using model (b). Based on model (b), the difference in average credit card balance between a male African and a male Caucasian is captured by the coefficient of the ethnicity indicator variable. Since both individual are male, the gender effect cancels out, and the comparison depends only on ethnicity. The estimated coefficient shows that a male African has a different (higher/lower, depending on sign) average balance than a male Caucasian, holding gender constant.

g.Comparing the average credit card balance of a male African and a male Caucasian when income is 100,000 dollars using model (c). Using model (c), when income is fixed at 100,000 dollars, the difference in average credit card balance between a male African and a male Caucasian is again determined by the ethnicity coefficient. Since both individuals have the same income and gender, the income and gender effects remain constant, and only ethnicity explains the difference in balance.

h.Comparing and comment on the answers in (f) and (g). The comparison in part (f) does not account for income, whereas part (g) controls for income by fixing it at 100,000 dollars. As a result, the difference observed in part (g) reflects the pure effect of ethnicity on credit card balance, independent of income. This shows that ignoring an important quantitative variable like income can lead to a misleading comparison.

i.Predicting the credit card balance of a female Asian whose income is 200,000 dollars using model (c). Based on model (c), the predicted credit card balance of a female Asian with an income of 200,000 dollars is obtained by substituting the given values into the fitted regression equation. The prediction includes the intercept, the effect of being female, the effect of Asian ethnicity, and the contribution of income. This predicted value represents the expected average credit card balance for an individual with these characteristics.

j.Interpretation The R^2 and Adjusted R^2 increase as we move from model (a) to model (c).This indicates that adding ethnicity and income improves the explanatory power of the model.Model (c) has the highest Adjusted R^2, so it provides the best goodness of fit among the three models.Hence, the model including Gender, Ethnicity, and Income is preferred.

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 $yi = \beta0 + \beta1x1i + \beta2x2i + \beta3(x1i \times x2i) + \epsilon i$, 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:($\beta0, \beta1, \beta2, \beta3$) = (−2.5, 1.2, 2.3, 0.001) , (-2.5, 1.2. 2.3, 3.1). Set seed as 123.

```
set.seed(123)
n=100
R=1000
beta=c(-2.5, 1.2, 2.3, 3.1)
mse1=numeric(R)
mse2=numeric(R)
for (i in 1:R) {
```

```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)
  mse1[i]=mean(residuals(fit1)^2)
  fit2=lm(y ~ x1 + x2)
  mse2[i]=mean(residuals(fit2)^2)
}
mean(mse1)

## [1] 0.9631944

mean(mse2)

## [1] 2.89452

set.seed(123)
n=100
R=1000
beta=c(-2.5, 1.2, 2.3, 0.001)
mse1=numeric(R)
mse2=numeric(R)
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)
  mse1[i]=mean(residuals(fit1)^2)
  fit2=lm(y ~ x1 + x2)
  mse2[i]=mean(residuals(fit2)^2)
}
mean(mse1)

## [1] 0.9631944

mean(mse2)

## [1] 0.9739083
```

Interpretation When the interaction effect is very small ($\beta_3 = 0.001$), the MSE of both models is almost the same.When the interaction effect is strong ($\beta_3 = 3.1$), the naive model has much higher MSE.This clearly demonstrates that ignoring an important interaction term leads to poor model performance.Hence, interaction terms should be included whenever theoretically justified.