

PredictiveAssignment-3

Shreyansh Kumar Das

2026-02-11

Problem Set 3

Question 2

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 4.3.3

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
## $ Balance : int 333 903 580 964 331 1151 203 872 279 1350 ...
```

Regress “balance” on

(a) “gender” only.

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

(b) “gender” and “ethnicity” .

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

(c) “gender”, “ethnicity”, “income”

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



```

(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

```

Interpretation:

- Model 1 (Gender) -
 - GenderFemale = 19.733
 - There are no significance stars hence the coefficient is statistically insignificant.
 - $R^2 = 0.0005$
 - It means the model based only on gender as a predictor explains almost no variability in the response variable.
 - Hence the model is a very poor fit.
- Model 2 (Gender+Ethnicity) -
 - GenderFemale = 20.038
 - EthnicityAsian = -19.371
 - EthnicityCaucasian = -12.653
 - Again none of the coefficients have any significance stars which implies none of the coefficients are statistically important.
 - $R^2 = 0.001$
 - It means that the second model also explains almost no variability in the response variable.
 - The second model is a very poor fit too.
- Model 3 (Gender+Ethnicity+Income) -
 - GenderFemale → Not significant
 - EthnicityAsian → Not significant
 - EthnicityCaucasian → Not significant
 - Income = 6.054* (p < 0.01) (highly significant)

- $R^2 = 0.216$ (substantial improvement)
- This model explains around 22% of variability in the response variable.
Hence the model is a better fit than the last 2 models.

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

- Model 1 (Gender) :
 - GenderFemale = 19.733
 - This means females, on average, have a balance 19.733 units higher than males.
 - However, this effect is not statistically significant.
 - Therefore, we cannot conclude that gender affects balance.
- Model 2 (Gender+Ethnicity) :
 - GenderFemale = 20.038
 - After controlling for ethnicity, females have about 20 units higher balance than males which is almost equal to the previous model.
 - Gender is still not statistically significant.
 - It means even after considering the ethnicity, the conclusion does not change that gender has no significant effect.
- Model 3 (Gender+Ethnicity+Income) :
 - GenderFemale = 24.340
 - After controlling for ethnicity and income, females appear to have 24.34 units higher balance than males.
 - Still not statistically significant.
 - Even after controlling for income (which is significant), gender remains statistically insignificant.

Overall, we can conclude that though gender has positive coefficient, it remains statistically insignificant and does not have any effect on balance .

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

```
data_to_be_predicted = data.frame(Gender = " Male" , Ethnicity = c("African
American" , "Caucasian"))
predicted_balance = predict(model2,newdata=data_to_be_predicted)
predicted_balance[1] - predicted_balance[2]

##          1
## 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).

```
data_to_be_predicted = data.frame(Gender = " Male" , Ethnicity = c("African
American" , "Caucasian"),
                                Income = 100)
predicted_balance = predict(model3,newdata=data_to_be_predicted)
predicted_balance[1] - predicted_balance[2]

##          1
## -6.446938
```

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

The comparison in (f) measures the difference in average balance between a male African and a male Caucasian without controlling for income. Here, we see that an African male has 12.653 units higher average balance than a Caucasian male.

In contrast, part (g) compares the two individuals while holding income fixed at \$100,000. Here, we see that an African male has 6.445 units lower average balance than a Caucasian male. Since income significantly affects credit card balance, the estimate in (g) provides a more accurate measure of the pure effect of ethnicity. Therefore, the comparison in (g) is more reliable and economically meaningful.

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

```
data_to_be_predicted = list(Gender = "Female" , Ethnicity = "Asian",
                             Income = 2000)
predicted_balance = predict(model3,newdata=data_to_be_predicted)
predicted_balance

##          1
## 12364.46
```

(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?

```
summary(model1)$adj.r.squared
## [1] -0.002050271

summary(model2)$adj.r.squared
## [1] -0.006876514

summary(model3)$adj.r.squared
## [1] 0.207774
```

Model (c) has the highest Adjusted R^2 . This indicates that including Income substantially improves the model's explanatory power. Models (a) and (b) perform poorly as they exclude Income, which is a highly significant predictor. Therefore, Model (c) is preferred as it provides the best balance between goodness of fit and model complexity.

Question 4

Consider a simulation setting where the data is generated as follows:

Step 1: Generate x_{1i} from Normal(0,1) distribution, $i = 1, 2, \dots, n$

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

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

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.

```
set.seed(123)
R = 1000
n = 100

sim_study <- function(beta0, beta1, beta2, beta3) {

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

  for (i in 1:R) {

    x1 = rnorm(n, 0, 1)
    x2 = rbinom(n, 1, 0.3)

    epsilon = rnorm(n, 0, 1)

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

    model_correct = lm(y ~ x1 * x2)
    model_naive = lm(y ~ x1 + x2)

    mse_correct[i] = mean((y - fitted(model_correct))^2)
    mse_naive[i] = mean((y - fitted(model_naive))^2)
  }

  avg_mse_correct = mean(mse_correct)
```

```

    avg_mse_naive = mean(mse_naive)

    return(c(avg_mse_correct, avg_mse_naive))
}

result1 = sim_study(-2.5, 1.2, 2.3, 0.001)
result2 = sim_study(-2.5, 1.2, 2.3, 3.1)

results = data.frame(
  Configuration = c("Small Interaction ( $\beta_3 = 0.001$ )",
                    "Large Interaction ( $\beta_3 = 3.1$ )"),
  MSE_Correct_Model = c(result1[1], result2[1]),
  MSE_Naive_Model   = c(result1[2], result2[2])
)

results

```

##	Configuration	MSE_Correct_Model	MSE_Naive_Model
## 1	Small Interaction ($\beta_3 = 0.001$)	0.9631944	0.9739083
## 2	Large Interaction ($\beta_3 = 3.1$)	0.9577982	2.8633349

The simulation study shows that when $\beta_3 = 0.001$, both models yield nearly equal MSE, indicating that ignoring a negligible interaction term does not significantly affect prediction accuracy. However, when $\beta_3 = 3.1$, the naive model (without interaction) produces a much larger MSE (2.863) compared to the correct model (0.958). This demonstrates that omitting an important interaction term leads to model misspecification and significantly poorer predictive performance. Hence, interaction terms should be included when they are significant.
