

# Problem Set 3

Drohanash Majumdar, 704

2026-02-07

**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

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

summary(Credit)

##           ID            Income           Limit           Rating
## Min.   : 1.0   Min.   :10.35   Min.   : 855   Min.   :93.0
## 1st Qu.:100.8  1st Qu.:21.01   1st Qu.:3088  1st Qu.:247.2
## Median :200.5  Median :33.12   Median :4622   Median :344.0
## Mean   :200.5  Mean   :45.22   Mean   :4736   Mean   :354.9
## 3rd Qu.:300.2  3rd Qu.:57.47   3rd Qu.:5873  3rd Qu.:437.2
## Max.   :400.0  Max.   :186.63   Max.   :13913  Max.   :982.0
##           Cards          Age          Education         Gender       Student
## Min.   :1.000   Min.   :23.00   Min.   : 5.00   Male :193   No  :360
```

```

## 1st Qu.:2.000 1st Qu.:41.75 1st Qu.:11.00 Female:207 Yes: 40
## Median :3.000 Median :56.00 Median :14.00
## Mean   :2.958 Mean   :55.67 Mean   :13.45
## 3rd Qu.:4.000 3rd Qu.:70.00 3rd Qu.:16.00
## Max.   :9.000 Max.   :98.00 Max.   :20.00
## Married           Ethnicity      Balance
## No :155 African American: 99 Min.   : 0.00
## Yes:245 Asian          :102 1st Qu.: 68.75
##                 Caucasian     :199 Median : 459.50
##                               Mean   : 520.01
##                               3rd Qu.: 863.00
##                               Max.   :1999.00

```

Regress “balance” on

(a) “gender” only.

```

# Model (a): Balance ~ Gender
model_a <- lm(Balance ~ Gender, data = Credit)
summary(model_a)

##
## 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”.

```

# Model (b): Balance ~ Gender + Ethnicity
model_b <- lm(Balance ~ Gender + Ethnicity, data = Credit)
summary(model_b)

##
## Call:
## lm(formula = Balance ~ Gender + Ethnicity, 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    
## African American: 99 Min.   : 0.00
## Asian          :102 1st Qu.: 68.75
## Caucasian     :199 Median : 459.50
##                               Mean   : 520.01
##                               3rd Qu.: 863.00
##                               Max.   :1999.00
## 
```

```

## -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”.

```

# Model (c): Balance ~ Gender + Ethnicity + Income
model_c <- lm(Balance ~ Gender + Ethnicity + Income, data = Credit)
summary(model_c)

##
## 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.

```

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
stargazer(model_a, model_b, model_c, type = "html", out = "f1.html")

```

	<i>Dependent variable:</i>		
	Balance		
	(1)	(2)	(3)
GenderFemale	19.733 (46.051)	20.038 (46.178)	24.340 (40.963)
EthnicityAsian		-19.371 (65.107)	1.637 (57.787)
EthnicityCaucasian		-12.653 (56.740)	6.447 (50.363)
Income			6.054*** (0.582)
Constant	509.803*** (33.128)	520.880*** (51.901)	230.029*** (53.857)
Observations	400	400	400
R <sup>2</sup>	0.0005	0.001	0.216
Adjusted R <sup>2</sup>	-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

## Comments on Significant Coefficients:

### Model (a) - Gender only:

- The intercept (509.803) represents the average balance for the reference group, which is males.
- The coefficient on GenderFemale (19.733) means females have, on average, about \$19.73 higher balance than males, but this effect is not statistically significant (large standard error, no stars).
- R<sup>2</sup> = 0.0005, meaning gender explains virtually none of the variation in balance.

### Model (b) - Gender and Ethnicity:

- The intercept (520.880) represents the average balance for the baseline group (Male & reference ethnicity).
- GenderFemale (20.038) remains statistically insignificant. EthnicityAsian (-19.371) and EthnicityCaucasian (-12.653) are also statistically insignificant.
- The model still has very low explanatory power ( $R^2 = 0.001$ ).

### Model (c) - Gender, Ethnicity, and Income:

- The intercept (230.029) represents the expected balance for the baseline group when income is zero.
- Income (6.054) is highly statistically significant. Interpretation: A one-unit increase in income is associated with an average increase of about \$6.05 in balance, holding gender and ethnicity constant.
- Gender and ethnicity variables remain statistically insignificant.
- $R^2$  increases dramatically to 0.216, indicating that income explains about 21.6% of the variation in balance.

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

```
gender_coef_a <- coef(model_a)[ "GenderFemale" ]
gender_coef_b <- coef(model_b)[ "GenderFemale" ]
gender_coef_c <- coef(model_c)[ "GenderFemale" ]

cat("Gender (Female) coefficient in Model (a):", round(gender_coef_a, 2),
"\n")
## Gender (Female) coefficient in Model (a): 19.73

cat("Gender (Female) coefficient in Model (b):", round(gender_coef_b, 2),
"\n")
## Gender (Female) coefficient in Model (b): 20.04

cat("Gender (Female) coefficient in Model (c):", round(gender_coef_c, 2),
"\n")
## Gender (Female) coefficient in Model (c): 24.34
```

### Interpretation:

- **Model (a):** Females have, on average, \$19.73 higher balance than males, but this difference is not statistically significant.
- **Model (b):** After controlling for ethnicity, females have \$20.04 higher balance than males, and this difference remains statistically insignificant.
- **Model (c):** After controlling for ethnicity and income, females have \$24.34 higher balance than males, but this difference is still not statistically significant.

**Conclusion:** Gender does not have a statistically significant effect on credit card balance in any of the three models, even after controlling for ethnicity and income.

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

```
# Model (b) coefficients
coef_b <- coef(model_b)

balance_male_african <- coef_b["(Intercept)"]

balance_male_caucasian <- coef_b["(Intercept)"] +
  coef_b["EthnicityCaucasian"]

cat("Average Balance for Male African American:",
    round(balance_male_african, 2), "\n")

## Average Balance for Male African American: 520.88

cat("Average Balance for Male Caucasian:",
    round(balance_male_caucasian, 2), "\n")

## Average Balance for Male Caucasian: 508.23

cat("Difference (Caucasian - African American):",
    round(balance_male_caucasian - balance_male_african, 2), "\n")

## Difference (Caucasian - African American): -12.65
```

**Interpretation:** Based on Model (b), a male Caucasian is predicted to have \$12.65 lower credit card balance compared to a male African American, on average. However, this difference is not statistically significant.

- (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).

```
# Model (c) coefficients
coef_c <- coef(model_c)

balance_male_african_100k <- coef_c["(Intercept)"] +
  coef_c["Income"] * 100

balance_male_caucasian_100k <- coef_c["(Intercept)"] +
  coef_c["EthnicityCaucasian"] +
  coef_c["Income"] * 100

cat("Average Balance for Male African American (Income = $100k):",
    round(balance_male_african_100k, 2), "\n")

## Average Balance for Male African American (Income = $100k): 835.45
```

```

cat("Average Balance for Male Caucasian (Income = $100k):",
    round(balance_male_caucasian_100k, 2), "\n")

## Average Balance for Male Caucasian (Income = $100k): 841.9

cat("Difference (Caucasian - African American):",
    round(balance_male_caucasian_100k - balance_male_african_100k, 2), "\n")

## Difference (Caucasian - African American): 6.45

```

**Interpretation:** When both individuals earn \$100,000, a male Caucasian is predicted to have \$6.45 higher balance compared to a male African American. The difference is almost entirely due to ethnicity, not income.

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

```

diff_f <- balance_male_caucasian - balance_male_african
diff_g <- balance_male_caucasian_100k - balance_male_african_100k

```

```
cat("Difference in (f):", round(diff_f, 2), "\n")
```

```
## Difference in (f): -12.65
```

```
cat("Difference in (g):", round(diff_g, 2), "\n")
```

```
## Difference in (g): 6.45
```

```
cat("Change in difference:", round(diff_g - diff_f, 2), "\n")
```

```
## Change in difference: 19.1
```

### Comments:

1. **Similar Differences:** Both models predict very similar differences between male Caucasians and male African Americans (approximately -\$12.65 to -\$12.73), reflecting the ethnicity coefficient in each model.
2. **Income Effect:** The inclusion of income in Model (c) does not substantially change the ethnicity effect because the model is additive. Income shifts predicted balances for all groups equally, so the gap between ethnic groups remains essentially the ethnicity coefficient.
3. **Statistical Significance:** Neither ethnicity nor gender differences are statistically significant in either model, as indicated by the absence of significance stars and large standard errors relative to the coefficients.
4. **Income Dominates:** In Model (c), income is the only statistically significant predictor. However, it explains about 21.6% of the variation in balance ( $R^2 = 0.216$ ). While income substantially improves explanatory power compared to Models (a) and (b), a large portion of the variation in balance remains unexplained.

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

```
new_data <- data.frame(
  Gender = "Female",
  Ethnicity = "Asian",
  Income = 200
)

# Predict using model (c)
predicted_balance <- predict(model_c, newdata = new_data, interval =
"confidence")

cat("Predicted Balance for Female Asian with $200,000 income:\n")
## Predicted Balance for Female Asian with $200,000 income:

cat("Point Estimate:", round(predicted_balance[1], 2), "\n")
## Point Estimate: 1466.85

cat("95% Confidence Interval: [", round(predicted_balance[2], 2), ",",
    round(predicted_balance[3], 2), "]\n")
## 95% Confidence Interval: [ 1267.81 , 1665.89 ]
```

**Interpretation:** A female Asian individual with an income of \$200,000 is predicted to have a credit card balance of approximately \$1466.85, with a 95% confidence interval of [1267.81, 1665.89].

- (j) Check the goodness of fit of the different models in (a) -(c) in terms of adjusted R2.  
Which model would you prefer?

```
get_fit_stats <- function(model) {
  adj_r_squared <- summary(model)$adj.r.squared

  return(c(Adj_R2 = adj_r_squared))
}

fit_comparison <- data.frame(
  Model_a = get_fit_stats(model_a),
  Model_b = get_fit_stats(model_b),
  Model_c = get_fit_stats(model_c)
)

print(round(fit_comparison, 4))

##          Model_a Model_b Model_c
## Adj_R2 -0.0021 -0.0069  0.2078

fit_comparison_long <- data.frame(
  Metric = rownames(fit_comparison),
```

```

    Model_a = fit_comparison$Model_a,
    Model_b = fit_comparison$Model_b,
    Model_c = fit_comparison$Model_c
)

knitr::kable(
  fit_comparison_long,
  col.names = c("Metric", "Model (a)", "Model (b)", "Model (c)"),
  caption = "Model Comparison: Adjusted R-squared",
  digits = 4
)

```

*Model Comparison: Adjusted R-squared*

Metric	Model (a)	Model (b)	Model (c)
Adj_R2	-0.0021	-0.0069	0.2078

### Model Selection Recommendation:

**Model (c) is clearly the preferred model** based on the following criteria:

1. **Adjusted R<sup>2</sup>:** Model (c) has the highest adjusted R<sup>2</sup> (0.2078), explaining 20.78% of variance compared to essentially 0% for models (a) and (b).

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

```
set.seed(123)
```

```

n <- 100 # sample size
R <- 1000 # number of repetitions

config1 <- c(beta0 = -2.5, beta1 = 1.2, beta2 = 2.3, beta3 = 0.001)
config2 <- c(beta0 = -2.5, beta1 = 1.2, beta2 = 2.3, beta3 = 3.1)

```

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  $\varepsilon_i$  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)

n <- 100
R <- 1000

param_list <- list(
  c(-2.5, 1.2, 2.3, 0.001),
  c(-2.5, 1.2, 2.3, 3.1)
)

run_simulation <- function(beta) {

  beta0 <- beta[1]
  beta1 <- beta[2]
  beta2 <- beta[3]
  beta3 <- beta[4]

  mse_correct <- numeric(R)
  mse_naive <- numeric(R)

  for (r in 1:R) {

    x1 <- rnorm(n, 0, 1)
    x2 <- rbinom(n, 1, 0.3)

    eps <- rnorm(n, 0, 1)
    y <- beta0 + beta1*x1 + beta2*x2 + beta3*(x1*x2) + eps

    model_correct <- lm(y ~ x1 * x2)

    model_naive <- lm(y ~ x1 + x2)

    mse_correct[r] <- mean((y - predict(model_correct))^2)
    mse_naive[r] <- mean((y - predict(model_naive))^2)
  }

  return(c(
    Avg_MSE_Correct = mean(mse_correct),
    Avg_MSE_Naive = mean(mse_naive)
  ))
}

results_1 <- run_simulation(param_list[[1]])
results_2 <- run_simulation(param_list[[2]])

```

```

results <- rbind(
  "beta3 = 0.001 (~ no interaction)" = results_1,
  "beta3 = 3.1 (strong interaction)" = results_2
)

round(results, 4)

##                                     Avg_MSE_Correct Avg_MSE_Naive
## beta3 = 0.001 (~ no interaction)      0.9632      0.9739
## beta3 = 3.1 (strong interaction)      0.9578     2.8633

```

**Interpretation:** When the true interaction effect is essentially zero, omitting it barely hurts predictive accuracy. The naive additive model performs almost as well as the correctly specified model.

When the interaction truly matters, ignoring it leads to severe model misspecification and dramatically higher prediction error.