Problem Set 3

Diyasha Basu

12-02-2026

Problem 2: Qualitative (Nominal) Predictors in Multiple Linear Regression

Loading Required Libraries and Data

```
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

# Load the Credit data
data("Credit")
```

(a) Regress “Balance” on “Gender” only

```
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) Regress “Balance” on “Gender” and “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
## -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) Regress “Balance” on “Gender”, “Ethnicity”, and “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 regressions in a single table

```
stargazer(model_a, model_b, model_c, type = "html", out = "f0.html")
```

	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
R^2	0.0005	0.001	0.216
Adjusted R^2	-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 is not statistically significant (p > 0.05). Gender alone does not significantly predict credit card balance.

Model (b): - Gender remains insignificant (p > 0.05) - Ethnicity Asian is significant at 5% level (p < 0.05), with negative coefficient - Ethnicity Caucasian is significant at 5% level (p < 0.05), with negative coefficient

Model (c): - Gender remains insignificant ($p > 0.05$) - Ethnicity coefficients become insignificant when Income is included - Income is highly significant ($p < 0.001$), showing strong positive relationship with Balance

(e) How Gender affects "Balance" in each model

```
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):** Males have about $19.73 lower balance than females (but not significant)
- **Model (b):** Males have about $19.90 lower balance than females (but not significant)
- **Model (c):** Males have about $7.38 lower balance than females (but not significant)

The effect of gender diminishes when income is included, suggesting that income differences between genders may explain some of the initial gender difference.

(f) Compare average credit card balance: Male African vs Male Caucasian (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
```

```r
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
```

(g) Compare average credit card balance: Male African vs Male Caucasian with $100,000 income (Model c)

```r
# 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
```

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

```r
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
```

Comment:

In model (b) without controlling for income, the difference between Male African American and Male Caucasian is substantial. However, in model (c) when we control for income at $100,000, the difference becomes much smaller. This suggests that much of the ethnic difference in credit card balance can be explained by differences in income levels between ethnic groups. When income is held constant, ethnicity has a much smaller effect on balance.

(i) Predict credit card balance for Female Asian with income $200,000

```
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 ]
```

(j) Check goodness of fit: Adjusted R^2

```
# Extract Adjusted R²
adj_r2_a <- summary(model_a)$adj.r.squared
adj_r2_b <- summary(model_b)$adj.r.squared
adj_r2_c <- summary(model_c)$adj.r.squared

# Create comparison table
comparison <- data.frame(
  Model = c("Model (a)", "Model (b)", "Model (c)"),
  Adjusted_R2 = c(adj_r2_a, adj_r2_b, adj_r2_c)
)

print(comparison)
```

```
##       Model  Adjusted_R2
## 1 Model (a) -0.002050271
## 2 Model (b) -0.006876514
## 3 Model (c)  0.207773976
```

Model Selection:

Based on the goodness of fit measures: **Adjusted R^2:** Higher is better. Model (c) has the highest Adjusted R^2.

Conclusion: Model (c) is the preferred model as it has the highest Adjusted R^2. This model includes Gender, Ethnicity, and Income, with Income being the most significant predictor.

Problem 4: Impact of Ignoring Interaction Term

```r
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, .., 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 $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 (x_{1i} \times x_{2i}) + \varepsilon_i$, 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: (β_0, β_1, β_2, β_3) = (−2.5, 1.2, 2.3, 0.001) , (-2.5, 1.2. 2.3, 3.1). Set seed as 123.

```r
set.seed(123)

n <- 100
R <- 1000

param_list <- list(
```

```r
  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.