Predictive Analysis Assignment 3

Priyanshu Dey 703

2026-02-12

```
#install.packages("stargazer")
```

Problem 2: Role of Qualitative Predictors in Multiple Linear Regression

This problem demonstrates the role of qualitative (nominal) predictors in addition to quantitative predictors in multiple linear regression using the "Credits" (Credit) data from R.

Data Preparation

We attach the `Credit` data from the `ISLR` library.

```
data("Credit")
df = Credit
head(df)

  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
```

Part (a), (b), (c):

Regression Models

Regress Balance on different sets of predictors:

- **(a)** Gender only.
- **(b)** Gender and Ethnicity.
- **(c)** Gender, Ethnicity, and Income.

```r
# Model (a): Regress Balance on Gender
model_a = lm(Balance ~ Gender, data = df)

# Model (b): Regress Balance on Gender and Ethnicity
model_b = lm(Balance ~ Gender + Ethnicity, data = df)

# Model (c): Regress Balance on Gender, Ethnicity, and Income
model_c = lm(Balance ~ Gender + Ethnicity + Income, data = df)
```

Part (d): Stargazer Output and Coefficient Analysis

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

```r
stargazer(model_a, model_b, model_c,
          type = "text",
          title = "Regression Results for Credit Balance",
          column.labels = c("Model A", "Model B", "Model C"),
          keep.stat = c("n", "rsq", "adj.rsq", "ser", "f"))
```

```
Regression Results for Credit Balance
=============================================================================
======
                                       Dependent variable:
                    ---------------------------------------------------------
------
                                             Balance
                          Model A            Model B                 Model C
                            (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
```

Comments on Significant Coefficients:

1. **Model (a):** The intercept is statistically significant . However, the coefficient for `GenderFemale` is **not significant**. This suggests that when considering Gender alone, there is no statistical evidence of a difference in average credit balance between males and females.
2. **Model (b):** Similar to Model (a), the coefficients for `GenderFemale`, `EthnicityAsian`, and `EthnicityCaucasian` are **not significant**. Ethnicity and Gender alone do not appear to be strong predictors of Balance.
3. **Model (c):** When `Income` is added, it is highly significant () with a positive coefficient. Interestingly, the intercept becomes significant and negative. Gender and Ethnicity remaining insignificant implies that even after controlling for Income, these demographic factors do not significantly influence the credit balance.

Part (e): Gender Effect

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

- **Model (a):** The coefficient for `GenderFemale` represents the difference in the average balance between Females and Males. Since the coefficient is approximately 19.73 (and insignificant), it suggests females have a slightly higher sample mean balance, but this is not statistically distinct from zero.
- **Model (b):** The coefficient for `GenderFemale` represents the difference in balance between Females and Males, **holding Ethnicity constant**. It remains insignificant.
- **Model (c):** The coefficient for `GenderFemale` represents the difference in balance between Females and Males, **holding Ethnicity and Income constant**. It remains statistically insignificant.

Part (f): Comparison (Model b)

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

In Model (b), the dummy variables are `EthnicityAsian` and `EthnicityCaucasian`. The baseline category (intercept) represents `EthnicityAfrican` (and `GenderMale`).

```
model_b


Call:
lm(formula = Balance ~ Gender + Ethnicity, data = df)

Coefficients:
       (Intercept)        GenderFemale      EthnicityAsian
EthnicityCaucasian
            520.88               20.04              -19.37              -
12.65
```

- **Male African:** Prediction = 520.88(the intercept)
- **Male Caucasian:** Prediction = 520.88 (intercept) + (-12.65) (EthnicityCaucasian) = 508.23

From the output, the coefficient for `EthnicityCaucasian` is approximately -12.65. Thus, a Male Caucasian has an average balance that is **12.65 lower** than a Male African. However, this difference is not statistically significant.

Part (g): Comparison (Model c)

Question: 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


Call:
lm(formula = Balance ~ Gender + Ethnicity + Income, data = df)

Coefficients:
       (Intercept)        GenderFemale      EthnicityAsian
EthnicityCaucasian
           230.029              24.340               1.637
6.447
            Income
             6.054
```

In Model (c):

- **Male African ($100k):** Prediction = 230.029 (intercept) + 0 (EthnicityAsian) + 0 (EthnicityCaucasian) + 6.054 (Income) * 100 = 835.429
- **Male Caucasian ($100k):** Prediction = 230.029 (intercept) + 0 (EthnicityAsian) + (6.447) (EthnicityCaucasian) + 6.054 (Income) * 100 = 835.429 + 6.447 = 841.876

The `Income` terms cancel out. The difference is solely the coefficient for `EthnicityCaucasian` in Model (c). From output (d), . Therefore, holding income fixed at $100,000, a Male Caucasian has an average balance 6.447 higher than a Male African. This difference is also not statistically significant.

Note: In the `Credit` dataset, Income is in thousands. Therefore, $100,000 is entered as 100.

Part (h): Comment on (f) and (g)

Question: Compare and comment on the answers in (f) and (g).

Both models suggest a very small, statistically insignificant difference between African and Caucasian males. The inclusion of `Income` in Model (c) changes the estimated difference in favor of Caucasians slightly. This indicates that `Income` is weakly correlated with `Ethnicity` in this dataset; otherwise, including Income would have drastically changed the ethnicity coefficients. In both cases, ethnicity is not a useful predictor.

Part (i): Prediction

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

Assumption: The prompt specifies "2000,000 dollars". Since the dataset represents Income in thousands (e.g., an income of 40 represents $40,000), an income of $2,000,000 corresponds to an input value of `2000`.

```
# new data for prediction
new_data = data.frame(
  Gender = "Female",
  Ethnicity = "Asian",
  Income = 2000
)

# Predicting from  Model C
prediction_val = predict(model_c, newdata = new_data, interval =
"prediction")
prediction_val

       fit      lwr      upr
1 12364.46 9985.223 14743.69
```

Note: Since this income (2000) is far outside the range of the training data (extrapolation), the prediction interval will be very wide and the point estimate may be unreliable.

Part (j): Goodness of Fit

Question: Check the goodness of fit of the different models in (a)-(c) in terms of AIC, BIC and adjusted . Which model would you prefer?

```r
# Calculating metrics
metrics = data.frame(
  Model = c("Model A", "Model B", "Model C"),
  Adj_R2 = c(summary(model_a)$adj.r.squared,
             summary(model_b)$adj.r.squared,
             summary(model_c)$adj.r.squared),
  AIC = c(AIC(model_a), AIC(model_b), AIC(model_c)),
  BIC = c(BIC(model_a), BIC(model_b), BIC(model_c))
)

metrics
```

```
    Model       Adj_R2      AIC      BIC
1 Model A -0.002050271 6044.527 6056.501
2 Model B -0.006876514 6048.434 6068.391
3 Model C  0.207773976 5953.518 5977.466
```

Conclusion: Model C is the preferred model. It has the highest Adjusted (indicating it explains the most variance after penalizing for complexity) and the lowest AIC and BIC scores (indicating the best fit vs. complexity trade-off). Models A and B have extremely low values, implying Gender and Ethnicity alone explain almost none of the variability in Balance.