Qualitative Predictors

Credit Data Set

- Balance: Average Credit Card Debt
- Age
- Cards: Number of credit cards
- Education: Years of education
- Income (in thousands of dollars)
- Limit: Credit Limit
- Rating: Credit Rating
- Gender
- Student
- Married
- Ethnicity: Asian, African American, Caucasian

Qualitative Predictors with Two Levels

- Consider only “Balance” (response variable) and Gender (qualitative predictor).
- For a qualitative predictor, we simply create an indicator/ dummy variable that takes on two possible values.
- Based on the “Gender” variable, we can have the following dummy variable:

$$x_i = \begin{cases} 1, & \text{if } i\text{th person is female} \\ 0, & \text{if } i\text{th person is male} \end{cases}.$$

Qualitative Predictors with Two Levels

- Fit the following model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i, & \text{if } i\text{th person is female} \\ \beta_0 + \varepsilon_i, & \text{if } i\text{th person is male} \end{cases}.$$

- So β_0 is the average credit card balance among males.
- And $\beta_0 + \beta_1$ as the average credit card balance among females.
- Therefore β_1 is the average difference in credit card balance between females and males.

Regression Table

	Coefficients	Std. Error	t-statistic	P-value
Intercept	509.80	33.13	15.39	<0.0001
Gender[Female]	19.73	46.05	0.43	0.67

Interpretation

- From the previous Table, we observe that the average credit card debt for males is $509.80.
- Females are expected to carry $19.73 in additional debt.
- However, we note that the p-value for the dummy variable is very high. This indicates that there is no statistical evidence of a difference in average credit card balance between the genders.

Qualitative Predictors with More than Two Levels

- Consider the variable “ethnicity” as a predictor.
- Ethnicity has three levels. So we need to create two dummy variables.
- The first dummy variable may be

$$x_{i1} = \begin{cases} 1, & \text{if } i\text{th person is Asian} \\ 0, & \text{if } i\text{th person is not Asian} \end{cases}.$$

- Similarly, the second dummy variable could be

$$x_{i2} = \begin{cases} 1, & \text{if } i\text{th person is Caucasian} \\ 0, & \text{if } i\text{th person is } not \text{ Caucasian} \end{cases}.$$

Regression Table

	Coefficients	Std. Error	t-statistic	P-value
Intercept	531.00	46.32	11.46	<0.0001
Ethnicity[Asian]	-18.69	65.02	-0.29	0.7740
Ethnicity[Caucasian]	-12.50	56.68	-0.22	0.8260

Interpretation

- From the previous Table, we see that the estimated average balance for the African American is $531.00.
- The Asian category are expected to have $18.69 less debt than the African American category.
- The Caucasian category are expected to have $12.50 less debt than the African American category.
- However, we observe that the p-values associated with the coefficient estimates for the two dummy variables are very large.
- This suggests that there is no statistical evidence of a real difference in credit card balance between the ethnicities.

Extension to Quantitative and Qualitative Variable

- Consider the predictors "income" (quantitative predictor) and "student" (qualitative predictor) along with balance as response variable.
- Suppose the model takes the following form:

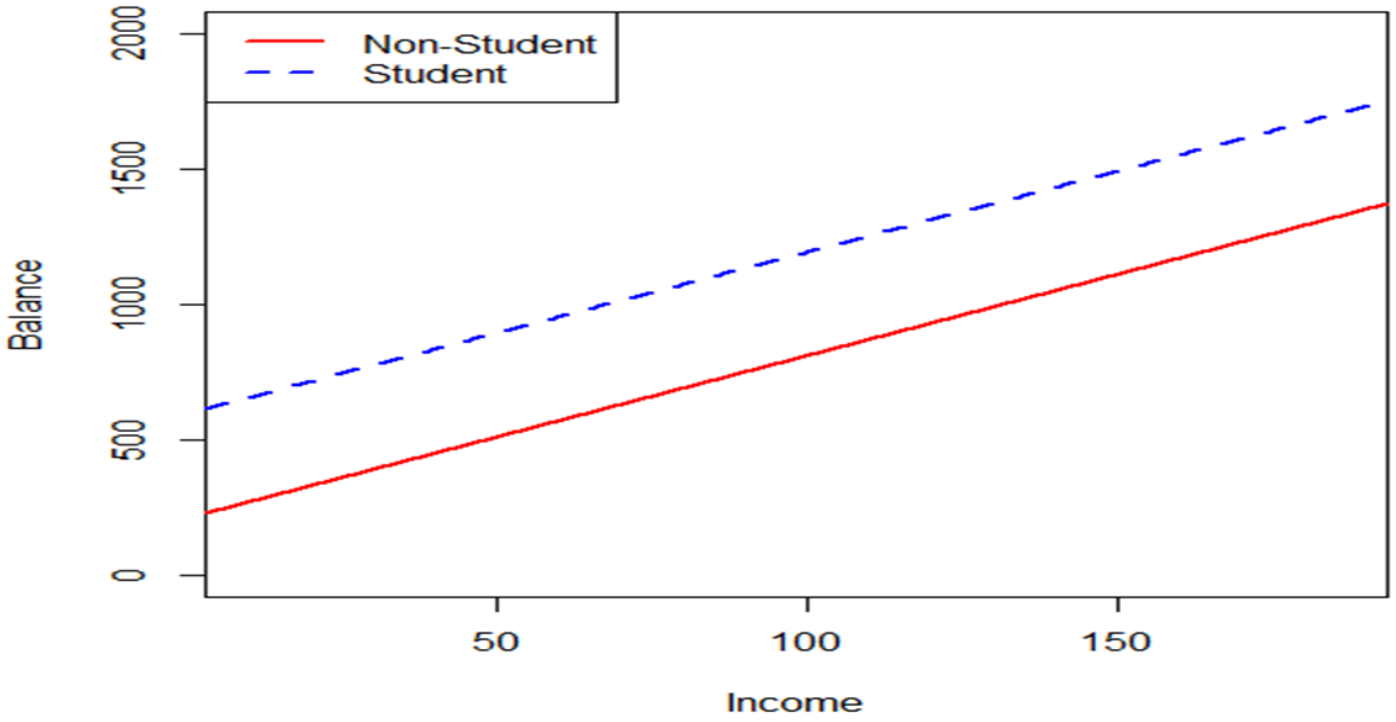
$$\begin{aligned}\text{balance}_i &\approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases}\beta_2, & \text{if } i\text{th person is student} \\ 0, & \text{if } i\text{th person is not a student}\end{cases}. \\ &= \beta_1 \times \text{income}_i + \begin{cases}\beta_0 + \beta_2, & \text{if } i\text{th person is a student} \\ \beta_0, & \text{if } i\text{th person is not a student}\end{cases}.\end{aligned}$$

Extension to Quantitative and Qualitative Variable

	Coefficients	Std. Error	t-statistic	P-value
Intercept	211.14	32.46	6.51	<0.0001
Income	5.98	0.56	10.75	<0.0001
Student[Yes]	382.67	65.31	5.86	<0.0001

Extension to Quantitative and Qualitative Variable

- This suggests that the average effect on balance of a one-unit increase in income does not depend on whether or not the individual is a student.
- This represents a potentially serious limitation of the model, since in fact a change in income may have a very different effect on the credit card balance of a student versus a non-student.



Inclusion of Interaction Term

- This limitation can be resolved if we add an interaction variable, created by multiplying income with the dummy variable for student.
- So the new model is

$$\begin{aligned}&\text{balance}_i \\ &\approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2 + \beta_3 \times \text{income}_i, & \text{if } i\text{th person is student} \\ 0, & \text{if } i\text{th person is not a student} \end{cases}. \\ &= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \text{income}_i, & \text{if } i\text{th person is student} \\ \beta_0 + \beta_1 \times \text{income}_i, & \text{if } i\text{th person is not a student} \end{cases}.\end{aligned}$$

Inclusion of Interaction Term

	Coefficients	Std. Error	t-statistic	P-value
Intercept	200.62	33.70	5.95	<0.0001
Income	6.22	0.59	10.50	<0.0001
Student[Yes]	476.68	104.35	4.57	<0.0001
Income × Student[Yes]	-2.00	1.73	-1.16	0.25

Inclusion of Interaction Term

- We observe that the slope for students is lower than the slope for non-students.
- This indicates that increases in income are associated with smaller increases in credit card balance among students as compared to non-students.