Regression using Dummy Variables

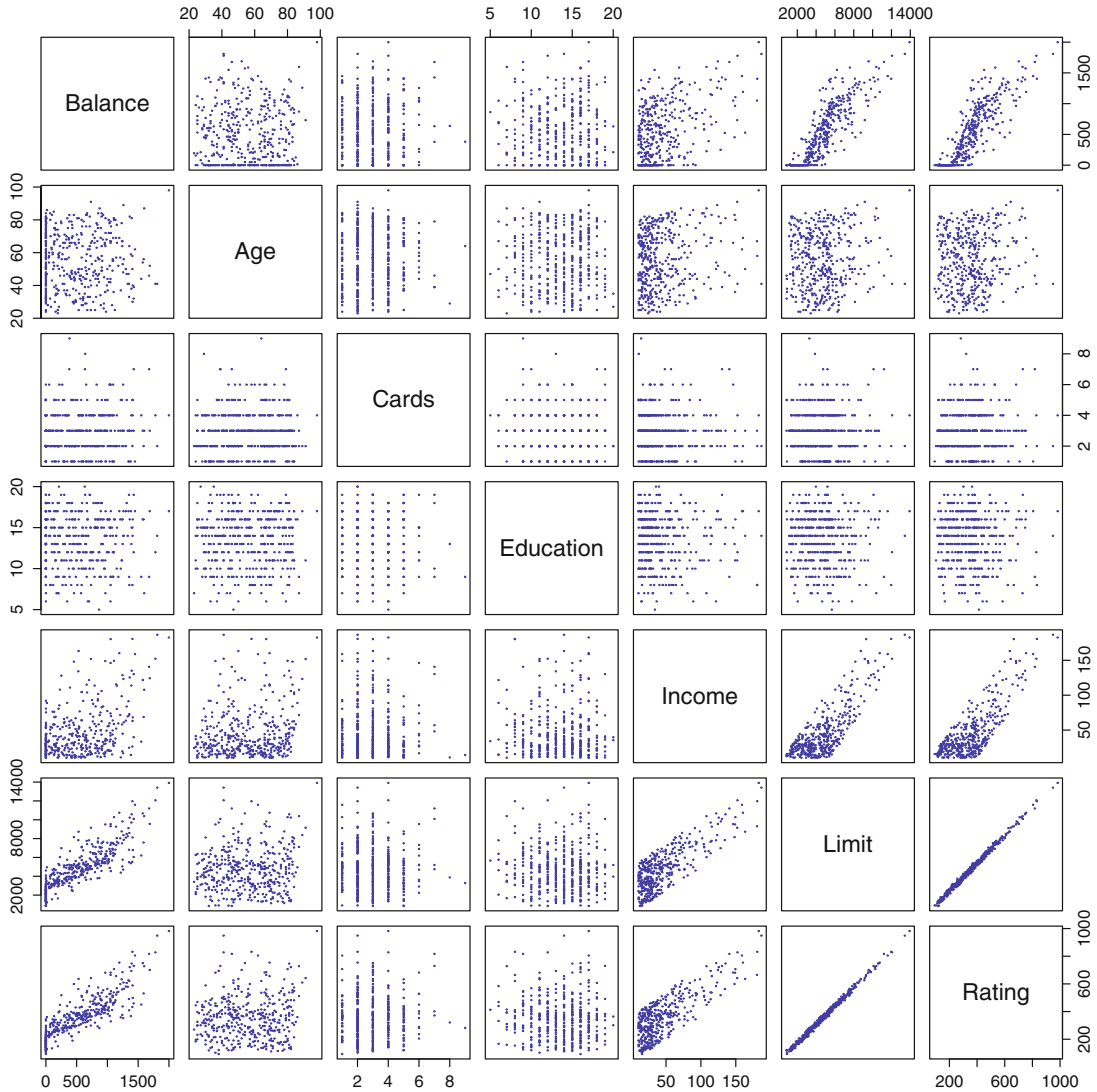
Rosalba Radice

Analytics Methods for Business

Qualitative Predictors

- Some predictors are not quantitative but are qualitative, taking a discrete set of values.
- These are also called categorical predictors or factor variables.
- See for example the scatterplot matrix of the credit card data in the next slide.
  In addition to the 7 quantitative variables shown, there are four qualitative variables: `gender`, `student` (student status), `status` (marital status), and `ethnicity` (Caucasian, African American (AA) or Asian).

Credit Card Data



The `Credit` data set contains information about `balance`, `age`, `cards`, `education`, `income`, `limit`, and `rating` for a number of potential customers.

Qualitative Predictors - continued

Example: investigate differences in credit card balance between males and females, ignoring the other variables. We create a new variable

$$x_i = \begin{cases} 1, & \text{unit } i \text{ is female} \\ 0, & \text{unit } i \text{ is male} \end{cases}$$

Resulting model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \text{unit } i \text{ is female} \\ \beta_0 + \epsilon_i, & \text{unit } i \text{ is male.} \end{cases}$$

Interpretation?

Credit Card Data - continued

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`gender[Female]`	19.73	46.05	0.429	0.6690

Least squares coefficient estimates associated with the regression of `balance` onto `gender` in the `Credit` data set.

Qualitative Predictors with More than Two Levels

With more than two levels, we create additional dummy variables. For example, for the `ethnicity` variable we create two dummy variables. The first could be

$$x_{i1} = \begin{cases} 1, & \text{unit } i \text{ is Asian} \\ 0, & \text{unit } i \text{ is not Asian,} \end{cases}$$

and the second could be

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

Qualitative Predictors with More than Two Levels - continued

- Then both of these variables can be used in the regression equation, in order to obtain the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & i \text{ is Asian} \\ \beta_0 + \beta_2 + \epsilon_i, & i \text{ is Caucasian} \\ \beta_0 + \epsilon_i, & i \text{ is African American} \end{cases}$$

- There will always be one fewer dummy variable than the number of levels. The level with no dummy variable - African American in this example - is known as the baseline.

Results for ethnicity

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	531.00	46.32	11.464	< 0.0001
`ethnicity[Asian]`	-18.69	65.02	-0.287	0.7740
`ethnicity[Caucasian]`	-12.50	56.68	-0.221	0.8260

Least squares coefficient estimates associated with the regression of balance onto ethnicity in the `Credit` data set. `ethnicity` is encoded via two dummy variables.

Quantitative Predictor and Qualitative Predictor with Two Levels

Consider the following model:

$$y_i = \beta_0 + \beta_1 x_i + \gamma z_i + \epsilon_i,$$

where x_i is a quantitative predictor and z_i is a binary variable, which takes the following values

$$z_i = \begin{cases} 0, & \text{unit } i \text{ belongs to group A} \\ 1, & \text{unit } i \text{ belongs to group B} \end{cases}$$

This results in the model

$$y_i = \beta_0 + \beta_1 x_i + \gamma z_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 x_i + \epsilon_i, & i \text{ belongs to group A} \\ (\beta_0 + \gamma) + \beta_1 x_i + \epsilon_i, & i \text{ belongs to group B.} \end{cases}$$

Quantitative Predictor and Qualitative Predictor with Two Levels - continued

Now consider following model with an interaction term:

$$y_i = \beta_0 + \beta_1 x_i + \gamma z_i + \delta x_i z_i + \epsilon_i.$$

This results in the model

$$y_i = \beta_0 + \beta_1 x_i + \gamma z_i + \delta x_i z_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 x_i + \epsilon_i, & i \text{ belongs to A} \\ (\beta_0 + \gamma) + (\beta_1 + \delta) x_i + \epsilon_i, & i \text{ belongs to B.} \end{cases}$$

Quantitative Predictor and Qualitative Predictor with more than Two Levels

Consider the model

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \gamma_3 z_{3i} + \epsilon_i \quad i = 1, \ldots, n, \\ &= (\beta_0 + \gamma_j) + \beta_1 x_i + \epsilon_i \quad , \; j = 1, 2, 3. \end{aligned}$$

where z_{ji}, $j = 1, 2, 3$ is the dummy regressor indicating whether an observation belongs to the jth category or factor level.

$$z_{ji} = \begin{cases} 1 & \text{if unit } i \text{ belongs to category } j \\ 0 & \text{otherwise} \end{cases}$$

The extension with more dummy regressors and interaction is trivial.

Inference for Main Effects and Interactions

The following four models can be compared, where $j = 1, \ldots, m-1$ represents the factor category:

Model		Description
	$y_i = \beta_0 + \epsilon_i$	constant
(A)	$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$	single line
(B)	$y_i = (\beta_0 + \gamma_j) + \beta_1 x_i + \epsilon_i$	parallel lines
(C)	$y_i = (\beta_0 + \gamma_j) + (\beta_1 + \delta_j) x_i + \epsilon_i$	separate lines

A sequential ANOVA table can be constructed to test between such models hierarchically.