Bus 701: Advanced Statistics

Fall 2011

Example: Bank Customers

File `bank_customers_v2010-10-19.xls` contains data concerning 5868 customers of a bank. Each row corresponds to one customer. The variables and their values are:

symbol	description	values
deficit	Credit limit exceeded within last 6 months?	0: no 1: yes
age	age in completed years	positive integer number
m.status	marital status	0: single or not specified 1: married
edu	highest level of education achieved	1: no graduation or general-education graduation 2: vocational school, apprenticeship 3: university of applied sciences, university
econ.act	economic activity	0: unemployed or inactive 1: employee or public servant 2: self-employed (or family member)
urban	Is place of residence in urban area?	0: no 1: yes
stability	Is residential area stable? (Many moves during recent period indicate "no".)	0: no 1: yes
cellphone	number of cellular phone contracts	0: no contract 1: one contract 2: two or more contracts

Elementary exercises:

1. Determine the scaling of each variable.
2. Find the percentage of customers with `deficit = 1`.
3. Determine the distribution of the variable *deficit*. Make a suitable plot of it.
4. Compute the average age (here: the arithmetic mean) of customers with `deficit = 0` and `deficit = 1`, respectively.
5. Compute the percentage of customers with `deficit = 1` among those with no (one, two or more) cellular phone contract(s).
6. Establish a contingency table of the variables *deficit* and *econ.act*.
7. Draw a stemplot of the variable *age* for those with `deficit = 1`. Is this meaningful?
8. Draw a histogram of the variable *age* for those with `deficit = 0` and `deficit = 1`, respectively.
9. Also draw this histogram into a pdf file.

Not so elementary exercises:

1. Can we explain the variable *deficit* using the variable *age*? Fit a logistic regression model to the data, with *deficit* as dependent and *age* as independent variable.
2. Fit a model as before, but with age groups [18,35], (35,50], (50,65], (65,80], (80,95] substituted for age. Does this model beat model (1)?
3. Now, extend model (2) by including *urban* as additional independent variable. Does this model beat either model (1) or model (2), or both?
4. Compute probabilities (scores!) for each customer to predict their value of *deficit* by applying model (3). Compute and compare mean scores for customers with `deficit = 1` and `deficit = 0`, respectively.