7. Case studies from 'Group B'

• The case studies of 'Group B' are more complicated case studies than the ones from 'Group A'.

⇝ The statistical (analytics) problem is generally still well defined, but broader in scope than the 'Group A' case studies.

⇝ Several solutions may need to be evaluated.

◇ In what follows, we present six case studies.

B.1 Expenses in municipalities

• The objective is to provide to a governmental client an estimate of the cost impact on municipal expenditures resulting from the proposed construction of new housing projects in three towns of the client's state.

• Since many of the services provided by a municipality are funded largely through property taxes, it is clearly of interest to try to determine whether these projects will produce an increase in expenditures.

$\rightsquigarrow$ Hence, the client would like you to develop a suitable model for predicting per person expenditures ($\rightsquigarrow$ allows to make direct comparisons between different towns), based on a common set of demographic and income-related predictors.

$\rightsquigarrow$ Following your analysis, the client would like to use the model as the basis for predicting future expenditures in order to support the decision whether to approve permits for a new housing project in a certain town.

Data

- The governmental client collected for a chosen year data from all municipalities within his state.

- Source file on the course's Moodle page: `Bexpenses.xls`

- Size: 914 rows (*i.e.* towns), 7 columns (*i.e.* variables).

- Variables:
  - EXPEN: expenditure per person (in dollars);
  - INCOME: mean income per person (in dollars);
  - WEALTH: wealth per person (in dollars), *i.e.* wealth related to real estate property values (measures different things than INCOME);
  - POP: population;
  - PINTG: percent intergovernmental, *i.e.* percentage of revenue that comes from state and federal grants or subsidies (measures different things than WEALTH and INCOME, *e.g.* public services);
  - DENS: density (= population per square kilometre);
  - GROWR: growth rate of the population.

B.2 Car's mileage in driving

• Your client is a marketing company that tries to determine what factors, which are involved in the construction of a car, affect the car's mileage in driving.

$\rightsquigarrow$ To investigate this the client has data on 154 cars, which he obtained from a famous car buyer's guide.

$\rightsquigarrow$ Your task is to help the client in determining the factors affecting the car's mileage in driving.

Data

- Source file on the course's Moodle page: `Bcars.xls`
- Size: 154 rows (*i.e.* cars), 10 columns (*i.e.* variables).

- Variables:
  - `PRICE`: price (in dollars);
  - `WEIGHT`: weight (in pounds);
  - `MPG`: car's mileage in driving (in miles per gallon);
  - `DISP`: displacement (in cubic centimetres);
  - `COMP`: compression ratio;
  - `HP`: horsepower;
  - `TORQUE`: torque ($5'200$ rpm);
  - `AUTO`: type of transmission ($1 =$ 'automatic', $0 =$ 'manual');
  - `CYLIN`: number of cylinders;
  - `COUNTRY`: country of origin ($1 =$ 'Japan', $2 =$ 'USA', $3 =$ 'South Korea', $4 =$ 'Italy', $5 =$ 'United Kingdom', $6 =$ 'Germany', $7 =$ 'USA/Canada', $8 =$ 'USA/Mexico', $9 =$ 'Japan/USA', $10 =$ 'Japan/Canada', $11 =$ 'Australia', $12 =$ 'Sweden', $13 =$ 'Germany/Mexico', $14 =$ 'Sweden/Belgium').

B.3 Analysis of birth weights

• Your governmental client provides you with data from a random sample of 1'440 birth records of his state.

$\rightsquigarrow$ Of particular interest to your client are incidents of 'low infant birth weight' (*i.e.* defined as birth weight less than 2'500 grams).

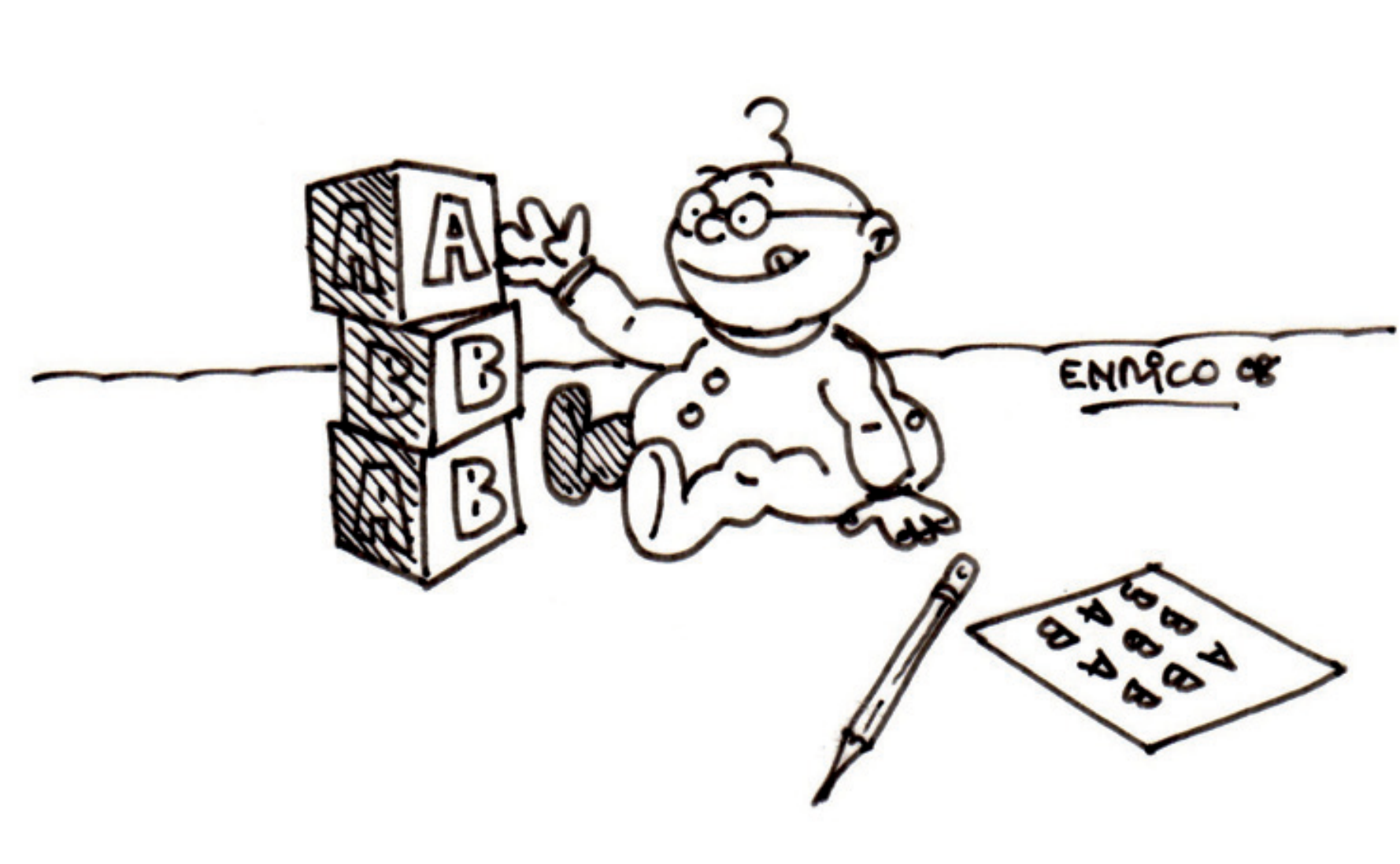
$\rightsquigarrow$ Incidence of low birth weight are associated with weaker development of many characteristics, *e.g.* intelligence, coordination or strength.

$\rightsquigarrow$ Your task is to help the client in determining the factors affecting low infant birth weight.

Data

- Source file on the course's Moodle page: `Bbirth.xls`
- Size: $1'440$ rows (*i.e.* births), 11 columns (*i.e.* variables).
- Variables:
  - `sex`: sex of child ($1 =$ male, $2 =$ female);
  - `race`: race of child ($0 =$ 'Other non-white', $1 =$ 'White', $2 =$ 'Black', $3 =$ 'American Indian', $4 =$ 'Chinese', $5 =$ 'Japanese', $6 =$ 'Hawaiian', $7 =$ 'Filipino', $8 =$ 'Other Asian or Pacific Islander');

- `age`: age of the mother (in years);
- `educ`: education level of the mother (in years);
- `gest`: completed weeks of gestation;
- `bwtgroup`: birth weight (grams) group (0 = '500 or less', 1 = '500–1'000', 2 = '1'001–1'500', 3 = '1'501–2'000', 4 = '2'001–2'500', 5 = '2'501–3'000', 6 = '3'001–3'500', 7 = '3'501–4'000', 8 = '4'001–4'500', 9 = '4'501 and over');
- `marital`: marital status (1 = 'married', 2 = 'not married');
- `cigs`: average number of cigarettes daily (98 = 'smokes an unknown amount');
- `drinks`: average number of alcoholic drinks weekly (98 = 'drinks an unknown amount');
- `plural`: number of children born of the pregnancy;
- `totounc`: weight of child in total ounces (1 ounce ≈ 28.35 grams).



B.4 Teaching improvement

• The aim of the client's study is to investigate whether teaching based on a person's learning style preferences would improve retention of the material taught.

$\rightsquigarrow$ As such, the client's sample consists of high school and elementary school students who were randomly allocated into two groups: 'Control' and 'Experiment'.

$\rightsquigarrow$ The traditional (textbook) teaching format was used for the 'Control' group.

$\rightsquigarrow$ For the 'Experiment' group, traditional teaching was augmented by activities specifically suited to the preferred learning styles of the students.

• Both groups were split into four sessions and each session received the same teaching formats.

• The so-called ‘Learning Style Inventory’ test instrument was first used to classify the preferences of the students for both groups.

• For assessment purposes, three quantitative measures were employed in this study: a pre-test, an attitude scale score and a post-test given one month after the teaching.

• The client wants you to investigate his research hypothesis that students in training sessions that utilise a processing activity that matches the students’ perceptual learning style preferences will demonstrate greater long-term retention of content than students in a traditional setting that has not utilised that processing activity.

Data

- Source file on the course's Moodle page: `Bteaching.xls`
- Size: 87 rows (*i.e.* students), 8 columns (*i.e.* variables).

- Variables:
  - `GROUP`: C = 'Control' (*i.e.* traditional teaching methods), E = 'Experiment' (*i.e.* incorporated learning styles);
  - `SESSION`: students were randomly assigned to 1 of 4 sessions (denoted by $S1$ to $S4$) within each `GROUP`;
  - `PREF`: students' learning style preference (T = 'Tactile', K = 'Kinesthetic', A = 'Auditory', V = 'Visual', N = 'No preference');
  - `GENDER`: M = male, F = female;
  - `SLEVEL`: school level (E = 'Elementary, H = 'High School');
  - `PRE`: pre-test score (out of 100);
  - `POST`: post-test score (out of 100);
  - `ATT`: attitude scale score (out of 60).

B.5 Segmentation of telecommunication customers

• Your client is a telecommunications company who provides you with data for 916 customers.

$\rightsquigarrow$ The data were preprocessed and result from an aggregation of different databases: basic customer information, call data aggregated by month and details of different tariff schemes used.

$\diamond$ The client would like to know whether there are any groupings of these customers.

Data

- Source file on the course's Moodle page: `Btelco.xls`
- Size: 916 rows (*i.e.* customers), 24 columns (*i.e.* variables).
- Variables:
  - `Age`: age in years;
  - `L_O_S`: length of service in months (since connect date);
  - `Dropped_Calls`: number of dropped calls during 6-month period;
  - `Peak_calls_Sum`: total number of peak-time calls in 6-month period;
  - `Peak_mins_Sum`: total number of peak-time call minutes in 6-month period;
  - `OffPeak_calls_Sum`: total number of off-peak calls in 6-month period;

- OffPeak_mins_Sum: total number of off-peak call minutes in 6-month period;
- Weekend_calls_Sum: total number of weekend calls in 6-month period;
- Weekend_mins_Sum: total number of weekend call minutes in 6-month period;
- International_mins_Sum: total number of international-call minutes in 6-month period;
- Nat_call_cost_Sum: total cost of national calls (peak + off-peak + weekend);
- AvePeak: average duration of peak-time calls during 6-month period;
- AveOffPeak: average duration of off-peak calls during 6-month period;
- AveWeekend: average duration of weekend calls during 6-month period;
- National_calls: total number of national calls in 6-month period;
- National_mins: total number of national call minutes in 6-month period;
- AveNational: average duration of national calls during 6-month period;

- `All_calls_mins`: total number of call minutes in 6-month period (national + international);
- `Mins_charge`: number of chargeable national call minutes in 6-month period (national minutes − free minutes);
- `call_cost_per_min`: cost of national calls per minute ignoring free minutes;
- `actual_call_cost`: cost of national calls after free minutes removes — indicates call mix;
- `Total_call_cost`: actual call cost + cost of international calls;
- `Total_Cost`: total call cost + fixed cost of tariff;
- `average_cost_min`: total cost / all call minutes (average call cost per minute including tariff cost and international calls).

B.6 Pain relief medication

• In the drug development process one of the important steps is to determine the minimum dosage amount of the drug that achieves efficacy. Sometimes this may involve estimating combinations of drugs that achieve maximum efficacy with a minimum drug amount.

• The experiment in this case study consists of administering a combination of two drugs to 10 mice at different doses. The drugs are morphine and marijuana and are used as pain relief medication ('tranquilisers').

$\rightsquigarrow$ In order to measure their effect 'flick tail' tests were performed for two drugs and their combination.

$\diamond$ The client's objective is to detect whether a synergy exists between these two drugs. In addition, the client is also interested to know the effective dosage at which 50% of the subjects would be expected to respond.

Data

• The experiment was repeated for 20 different drug combinations. For six of the experiments the mice were injected with a combination of both drugs, whereas for 13 experiments the mice were administered only one drug. One of the experiments was a control for which no drug was given to the 10 mice.

• Source file on the course's Moodle page: `Bpain.xls`

• Size: 20 rows (*i.e.* drug combinations), 7 columns (*i.e.* variables).

- Variables:
  - `Obs`: index number;
  - `Set`: one of four possible combinations (`O` = 'control', `A` = 'morphine', `B` = 'marijuana', `AB` = 'combination of both drugs');
  - `Morphine`: dose of morphine 'sulfate' (mg/kg) injected into study mice — the range is 0 to 8;
  - `Marijuana`: dose of marijuana 'Delta9-THC' (mg/kg) injected into study mice — the range is from 0 to 16;
  - `Flick`: the number of mice that did not flick the tail after being applied a heat stimulus from beneath ($\rightsquigarrow$ 'flick tail' test);
  - `Reps`: number of repetitions (with different mice);
  - `Prob`: proportion of mice (out of 10) that did not flick their tails after being administered a heat stimulus for a given drug combination (= `Flick`/`Reps`).