Linear and Logistic Regression with Data Gathering
Design of experiments

Anna Lindgren – Mathematical statistics

4 April, 2017
FMSN40 Some old projects

- What influences the price of a used car?
- Biomarkers in ovarian cancer
- Cancer rates in Sweden, have they changed over time?
- Carbon dioxide emissions in different car models
- Stock price and activity on social media
- Modelling of total seasonal score for an NBA player
- Regression av huspriser i Sverige
- Övervikt i världen
- Bil utan tak — andelen cabrioletter i olika kommuner
Choice of variables to measure

What variables do we want to measure?

▶ They should be those that best describe what we want to study.
▶ Often it is impossible to measure these directly. We then need to choose some other variable that we hope will reflect the behaviour of the one we wanted.
▶ Unfortunately, we cannot draw any conclusions from of model regarding these wished for variables, only the ones we managed to measure.
Questionnaires

When gathering data from/about persons a questionnaire can be useful.

▶ Minimize the number of questions asked! Too many questions tend do give large numbers of missing values.

▶ Formulate your questions so that they do not contain ambiguities. The respondents are not mind readers and cannot know how you intended the question to be interpreted. They will make their own interpretations. Test your questions on some friends first so that you can rephrase problematic questions.

▶ Phrase the answer alternatives so that they can be used in a regression model. (No free text answers here!)
A warning against using time series data in the project

- In regression analysis we model (the mean of) a response variable $Y$, as a function of a number of independent variables, $x_1, \ldots, x_p$. We assume that the errors are independent of ”everything else”, in particular the errors for one observation should be independent of the errors of other observations.

- It is possible to deal with a situation with a simple short term dependency but if there is a complicated time dependency between successive observations, we need a different approach: time series.

- This course does not deal with times series. Take FMSF10 Stationary stochastic processes next period instead, followed by FMSN45 (FMS051) Time series analysis, followed by FMSN60 (FMS161) Financial statistics.

- Exception: ”repeated measurements” where a large number of persons are measured at a limited number (3–4) of fixed times points, is OK.
Study types

Experimental study
Select the values of the independent variables and then measure the corresponding responses.

▶ avoids multicollinearity
▶ design points can be made to have equal influence on the estimation of the regression coefficients
▶ possible to detect the need for nonlinear terms
▶ unfortunately cannot be done in many practical situation (impossible and/or unethical)
An experimental design

- Choose a number of fixed values (levels) for each of your independent variables:
  e.g. \( x_1 = (10, 20, 30, 40), x_2 = (-1, 0, 1), x_3 = (3, 6, 9) \)
- Take all \((4 \cdot 3 \cdot 3 = 36)\) possible combinations of these values and make at least one, but preferably more, independent measurements of the response variable for each combination.
- The more levels you have the easier to find nonlinear terms.
- The more levels you have the more measurements you will need.

The course FMSF65 (FMS072) Design of Experiments deals much more with this.
Observational study

The independent variables are measured alongside their response.

- You have no control over the independent variables.
- Collinearity might be an issue.
- You may lack observations for important variable combinations.
- But it is often the only alternative outside a laboratory setting.

An observational design

- Choose independent variables that you think will not be overly correlated with each other.
- The independent variables should be expected to vary sufficiently in your study population.
Variable types

Variables can be measured on a number of different scales:

▶ Continuous, e.g. length, temperature, cost,
▶ Count data, e.g. number of deaths, number of sold items
▶ Ordinal = ordered categories, e.g. CEQ-questions, small/medium/large
▶ Nominal = un-ordered categories, e.g. mode of transport, yes/no

The type of regression you should chose depends on the type of your response variable.
Linear regression should only be performed using a continuous variable. Logistic regression can only be performed using a nominal variable with two categories (yes/no).
Other types of regressions can be used for other types of response variables (Poisson, multinomial or ordinal logistic, etc.)
Continuous variables

Measurements on a continuous scale, e.g. length (m), cost (SEK), temperature (°C), time (days).

- Can be normally distributed but other distributions are also possible, e.g. exponential, uniform, lognormal, etc.
- Model the mean (or median or other quantile) of $Y_i$ as a (linear?) function of the independent variables.
Count data

Measures the number of occurrences during some interval, e.g. number of sales during a month, number of cars passing during a time interval, number of children, etc.

- Often Poisson distributed with mean proportional to the length of the time intervals. Can also have a Negative binomial distribution if the variation is larger than a Poisson distribution.
- Model the logarithm of the mean of $Y_i$ as a (linear?) function of the independent variables.
Ordinal categories

The measurements are only recorded as belonging to a specific category where the possible categories have a natural order, e.g. the CEQ-questions. Degree of illness: mild, moderate, severe. I love meatballs: strongly disagree, somewhat disagree, neither disagree or agree, somewhat agree, strongly agree.

- Used when there is no easy way to measure (or express in words) some underlying continuous variable.
- The difference between categories has no sensible size and any numerical values we assign them are totally arbitrary (so long as the order between them is preserved).
- Model the probability to fall in the different categories using ordinal logistic regression.
Nominal categories

The measurements are recorded as belonging to a specific category where the possible categories have *no* natural order, e.g. gender: male, female.
Example: How did you get to school today: walked, bike, bus, train, other.

- Used when data have natural categories.
- Any numerical values we assign to the categories are totally arbitrary and no use should be made of the order, since there is none.
- Model the probability to fall in the different categories using logistic regression (multinominal logistic regression in the case of more than two categories).
Sample size

- Under ideal conditions we need a minimum of 7 observations for each parameter in order to make reasonably good estimates.

- Ideal conditions include: an experimental design where the model is correct, our independent variables are all orthogonal to each other and sufficiently spread out.

- Each linear variable thus needs 7 observations, each categorical variable needs 7 observations for each category, except the first. If the response variable is categorical it also needs 7 observations extra for each additional category. Any interaction terms will need an additional 7 observations each, etc.
Sample size calculation

- If we have a model with two continuous variables and one categorical variable with three categories we thus have 5 $\beta$-parameters and thus need an absolute minimum of $5 \cdot 7 = 35$ observations.

- If we suspect that we might have to add a quadratic term to one or both of the continuous variables we have 7 parameters and thus need 49 observations.

- The addition of an interaction term between one of the continuous variables and the categorical variable requires another 2 parameters (one for each extra category) requiring a total of 63 observations.

- Since we probably do not have ideal conditions we might want to double or triple that number and make 200 observations.
Project 3: . . . with Data Gathering

▶ Come up with a situation where the variablilty of one variable might be explained by some (3+) other variables.
▶ Collect the relevant data.
▶ Fit the model and validate it.
▶ Present the results.

**Deadline** for the project plan: 16.00 on Thursday, 27 April.
**Deadline** for the revised plan and collected data: 16.00 on Monday, 15 May.
**Oral presentations** 22–23 May