Experimental Designs leading to multiple regression analysis

1. (Randomized) designed experiments.
2. Randomized block experiments.
3. Observational studies: probability based sample surveys
4. Observational studies: sample of convenience.

Randomized designed experiments

- want to study the effect of variables $x_1, x_2, \ldots, x_p$ on a response variable Y.
- Experimenter **chooses** n sets of values of $x_1, x_2, \ldots, x_p$ and measures the response Y on n **experimental units**.
- Experimental Units are assigned **at random** to levels (that is to the particular combinations of x values).
- This is a much better method than other methods for deciding which experimental units get which x values.

Designed Experiments

Example:

- Experimental Unit is a batch of plaster
- $n = 18$ batches made.
- x_1 is the sand content and x_2 is the fibre content. We tried 3 settings of x_1, 3 of x_2 and tried each of the $3 \times 3 = 9$ combinations twice.

Randomized Block Designs

- Want to study the effect of variables $x_1, x_2, \ldots, x_p$ on a response variable Y of an experimental unit.
- BUT Y is probably influenced by variable B which the experimenter cannot control.

Example of Randomized Block Designs

- x_1 is log(Dose) of some drug.
- B = sex of patient (patient is the experimental unit).
- experimenter can assign patient to level of x_1 but NOT to the level of B.
- B is called a blocking factor.
- Blocking can serve useful purpose: increase precision of estimates of effects.

Another Example

- Y is lung capacity
- B_1 is cigarettes smoked per day
- B_2 is age
- B_3 is sex
- x_1 is daily vitamin C intake
- x_2 is daily Echinacea dose
- Key point is that x_1 and x_2 are under control of the experimenter but the other factors are not.

Observational Studies

- values of Y and variables $x_1, x_2, \ldots, x_p$ are determined by sampling from a population.
- covariates $x_1, x_2, \ldots, x_p$ are not controlled by the experimenter.

Example

- As in the previous example but suppose vitamin C and echinacea intakes are not controlled, just measured.

Vital Distinction

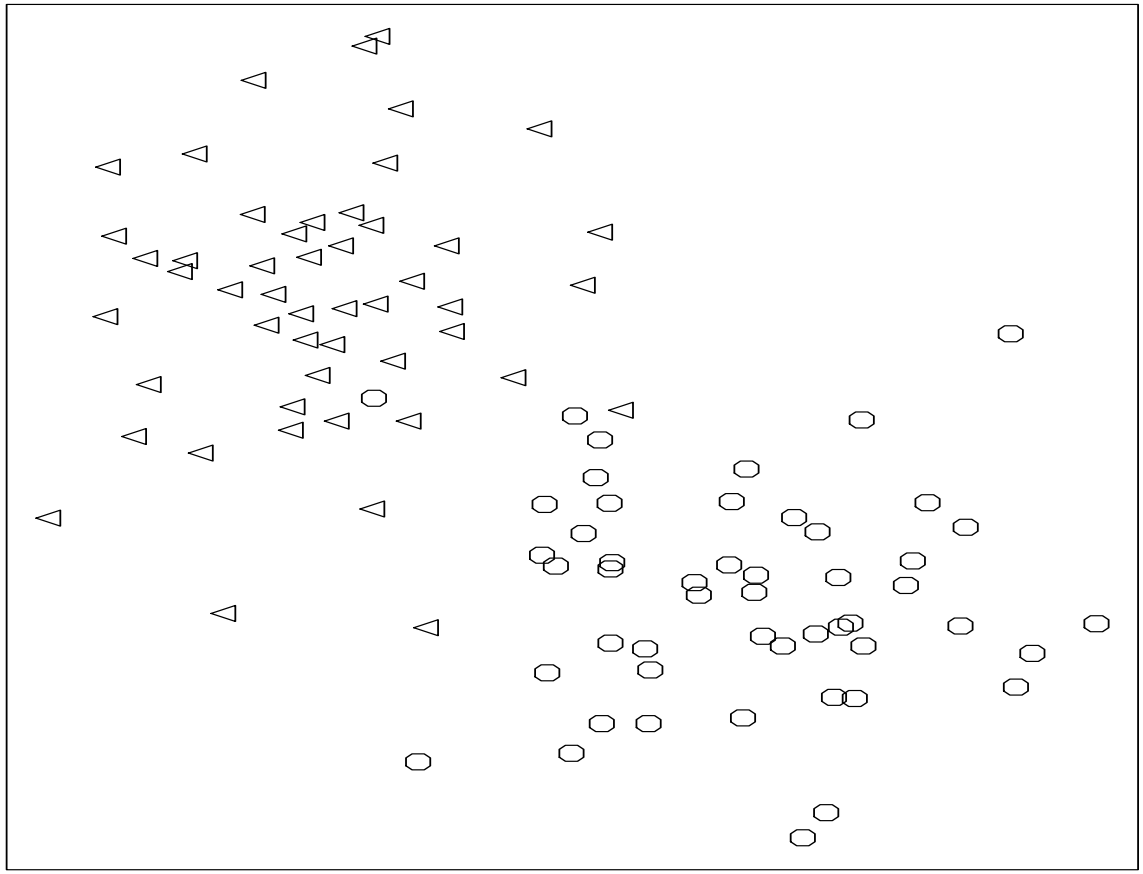
- Cause and effect relations are convincingly deduced only for controlled variables.
- Interpretation of regression coefficients is difficult in observational studies.

Cause and Effect

- Inference in an observational study is largely descriptive.
- BUT researchers in social science often want to know if changes in variable X **cause** changes in Y.
- The interpretation is that if X could be manipulated then Y would be changed.
- To demonstrate that changing X causes changes in Y we hold all other important variables constant and try experimental units at various settings of X.
- Variables we don't know about or can't control are equalized between the different levels of X by randomly assigning units to the different values of X.
- An observational study is one where X cannot be controlled and other variables cannot be held constant.

Hypothetical Example

- Think about a case where men have generally higher values of both X and Y and women have generally lower values but that among men there is no relation between X and Y
- Here is a possible plot, the triangles being men.



Discussion

- If you didn't know about the influence of sex you would see a positive correlation between X and Y.
- But if you compute separate correlations for the two groups you see the variables are unrelated.
- Remember, if you manipulate X in the picture you are either doing so for a women (and X and Y are unrelated for women) or for a man (and again X and Y are unrelated).
- In either case Y will be unaffected because you would not be affecting the sex of a person.

Discussion Continued

- Doing multiple regression is very much like this.
- Imagine you have a response variable Y, a variable X whose influence on Y is of primary interest and some other variables which probably influence Y and may influence X as well.
- You would like to look at the relation between X and Y in groups of cases where all the other covariate values are the same; this is not generally possible.
- Instead, we estimate the average value of Y for each possible combination of the variable X and the other variables.
- We ask if this mean depends on X. We say we are **adjusting** for the other covariates.
- The method works pretty well if we have identified all the possible **confounding** variables so that we can adjust for them all.

SENIC example

- Example to come: regress risk of infection on many variables.
- One is Nurses per patient. Estimated coefficient is positive.
- More nurses means more infection? Fire nurses?
- Trouble: no such deduction is rigorously possible.
- Need to be sure there is no 3rd variable correlated with both X and Y which causes variation in both and for which you haven't adjusted.
- Designed experiments deal with problem by **randomization**.
- The slope in a regression model corresponding to X measures the change expected in Y when X is changed by 1 unit and all the other variables in the regression are held constant.
- Regression method is used to *adjust* for the other covariates.
- Researchers say things like "Adjusted for Length of service and publication rate sex has no impact on salary of professors."
- But not many say that particular thing.

Observational Studies: samples of convenience

- Multiple regression logic depends on model assumptions.
- Sometimes justified by fact data is sample from population with suitable structure.
- Sometimes used when data is just gathered in some convenient way.
- Inference extends from sample to population from which it is sample.
- Sampling biases produce invalid regression results.

Compare and contrast

- Randomized controlled experiments provide most compelling proof of cause and effect.
- But experiments not entirely like real world – better medical care, for instance, in a clinical trial than is normal.
- So effect sizes in experiment may not match effects in real world.
- Observational studies nearly always leave room for doubt about cause and effect.
- Econometricians use technique called *instrumental variables*.
- Technique requires assumptions.
- Demonstration of cause and effect in such studies always uses assumptions.
- Many different observational studies "showing" same point in different contexts are more compelling.