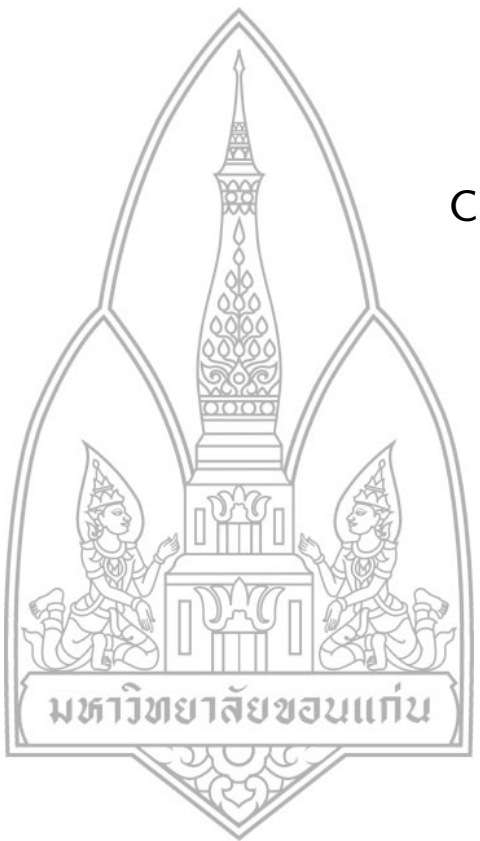
Biostatistics workshop series: Introduction to Modelling



Dr Cameron Hurst
cphurst@gmail.com

CEU, ACRO and DAMASAC, Khon Kaen University

26th June 2556

What we will cover....

1. Classical health study design
   - Experimental studies
   - Observational study designs
2. Study objectives
3. Model building approaches
   - Throw it all in
   - Computer algorithm approaches
   - The heirachical approach
   - Purposeful selection of Covariates (PSC)

Conventions

The conventions I will use:

Notes and Hints:.....

Things to note will occur in a green box

Pitfalls:.....

Common mistakes and things to watch out for will occur in a red box

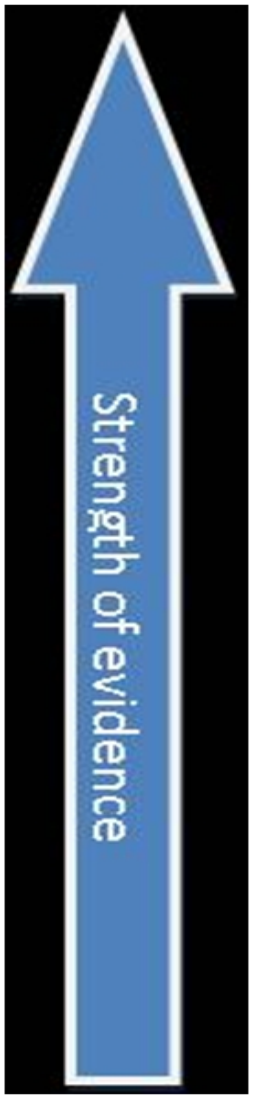
Introduction

- Often it is not clear what is the best approach to use both in designing our study, and performing the subsequent analysis
- Even after 20 (or so) years of consulting, I am still presented research questions that give me a headache
- Issues such as the nature of the research question, the likely nature of the data, and precedence set in other studies will often present a minefield
- Many of you will have to consult a biostatistican before moving on to the data collection phase
- Good news is that there are a few principles you can follow, that make life a little easier

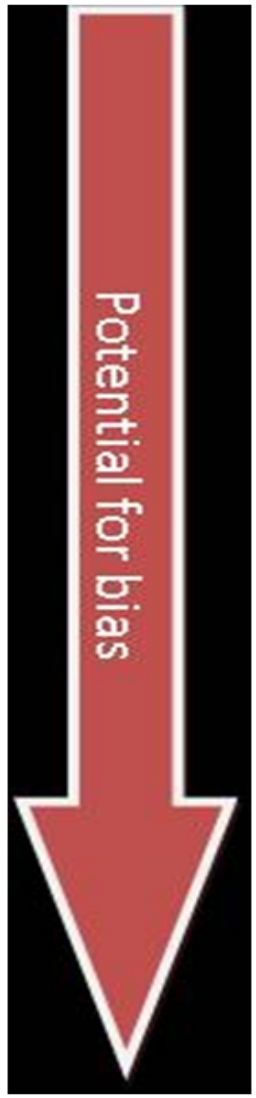
Strategy

- In this presentation, I will review the study designs commonly used in health studies
- I will also discuss how the study objective (i.e Why we are doing the research) also has an impact on how we should plan to conduct our study
- The fact is that there are no simple rules that we can follow that will always get us to the right study design and analytical approach
- Too many factors contribute to what represents the most optimal and/or practical approach
- Finally we will get to the main topic: How can we develop the best model to address our research question?

- We will start by examining some of the (classical) health and medical study designs
- The main factor that governs our choice of study design is the **research question**, **target population** (and ability to sample it) and (correspondingly) our **ability to control sources of bias**
- Some of you are lucky. You are interested in clinical intervention studies, which tend to occur at the stronger end of health study designs. For others.....well you don't want life to be boring do you????



Study design
Randomized Controlled Trial
Quasi-experiment
Cohort(Longitudinal)
Cross-sectional
Case-Control
Ecological
Case series
Case study



Randomized controlled trials

- Widely considered to be the gold standard of study designs
- Provide the strongest evidence
- Double blinding and randomization attempt to minimize selection and confounding bias
- As a true experiment, RCTs always involve an intervention (or treatment)
- That is, experimental manipulation by the researcher
- Often not possible in (population-based) epidemiological studies, especially those involving the study of risk factors (unethical to impose a protocol of risky behaviour)

Quasi-experiments

- Many study don't (or can't) include a randomization process, but still involve an intervention
- For example, it may be unethical or impractical to randomize patients, or it might be impossible to avoid contamination between different experimental arms
- Classical example of a quasi experiment is a pre-post study
  - All participants are observed at baseline
  - Administered treatment \ intervention
  - All participants are observed again post-treatment
- Problems (confounders) arise with the passage of time
  - For example, 'Pre' in winter, 'Post' in summer
  - Co-interventions: e.g. more severe patients will be treated differently
  - Pre and post observations for a given participant are correlated (often needing advanced statistical methods)

Cohort studies

- Generally the strongest of the 'observational' designs
- **Observational designs**: Researcher has no control over group membership (i.e. non-experimental)
- Longitudinal studies involve the observation of participants over time
- Advantage is that we have stronger evidence of causal relationships (as causes precede effect)
- Most common problem is longitudinal studies is loss-to-follow-up (LTFU)
- LTFU especially a problem when it confounds with effects/outcomes of interest
- That is LTFU cannot be assumed to occur at random

Cross-sectional studies

- Often called 'prevalence' studies as we can get an estimate of disease prevalence from our sample (or subgroups of the sample)
  - In contrast see case-control studies below
- One of the most common study designs in population-based epidemiology as they are generally cheaper and more practical than longitudinal studies
- Main problem is that without strong contextual evidence, associations can not be assumed to be causal (only associative)

Case-control studies

- Case-control studies are usually performed where there is a scarcity of participants with the 'condition' of interest (e.g. Low prevalence)
- Idea is to collect a group of cases with the outcome of interest and a corresponding group (controls) without
- In this respect the relative balance of cases to controls is **an artefact of study design** (so in no way reflects prevalence)
- Then (along with measuring observable traits) we ask participants about exposure history
- For this reason, case-control studies are prone to recall bias
- Important to note that (unlike any other health study design) it is the **outcome that defines group membership** (not intervention or risk factor group)
- In CC studies trying to find differences in the **prevalence of exposure** between cases and controls

Matched case-control studies

- There is also an important variant of case-control designs: matched case-control designs
- These designs attempt to control confounders by matching individuals in the case-control groups based on particular 'known' confounders (e.g. Age, Gender, SES)
- For example, we may age-gender match individuals between the case and control groups
- These designs should be analysed using appropriate methods (e.g. conditional logistic regression) that account for the matched pairs\sets (strata)

Ecological studies

- Finally we come to the last of the 'data-based' study designs, ecological studies
- Typically the outcomes from such studies represent aggregate counts (or rates) from different times (e.g. year) or spaces (e.g. countries or provinces)
- Often data are routinely collected (e.g. hospital or government departmental data)
- As the outcomes are often counts (or rates), Poisson regression (or related methods) are often used in these studies
- Probably the study design most prone to confounding, as there are always multiple sources of variation among observations ($\rightarrow$ ecological fallacy)
- We will consider an ecological study when we cover the Generalized Linear Model: *Poisson Regression*

Impact of study design

- The study design can have a major impact on analytical planning
- I would like to clarify what I mean by **analytical planning**. I mean:
  - What is the likely nature of our outcome(s) and predictor(s)?
  - What are the practical issues in sampling (e.g. Random sample, stratified samples, convenience samples)?
  - Based on our study design, what statistical method \ modelling approach should we use?
  - Prospective powering: What sample size is appropriate?
  - Given the above, what is my proposed modelling strategy:
    1. What method should I use (Linear, Logistic, Poisson, Cox regression etc)?
    2. How will I decide on **what modelling strategy (model building) to use?**

It is this last issue that is the main focus of this lecture.

Study objectives

There are a number of types of reasons (research output) that researchers may consider. In my experience these (mainly) fall into three different areas:

1. **Predictive modeling**: The main purpose of the study is to most accurately predict a particular patient characteristic
2. **Hypothesis testing**: The researcher has identified a main and particular **Study effect** on a particular outcome (and they want to test a hypothesis about this)
3. **Exploratory studies**: The study might be 'epidemioloical' where we are interested in which predictors represent risk/protective factors for a particular disease (and we don't have strong pre-concieved notions on what these might be)

Predictive modeling

In predictive modelling we are trying the model that BEST PREDICTS the outcome. One common application for this is in clinical epidemiology where we might be trying to develop a diagnostic, prognostic or screening tool.

For example, we may have a set of predictors and we are trying to find the best model to correctly diagnose a disease.

In this case, we would want the model with the best Area under the Reciever-Operator Characteristic curve (AUC-ROC) and the most desirable sensitivity and/or specificity. These crietria alone (rather than statitsical significance) would govern the choice of the 'best' model

(Research) hypothesis testing

When testing the significance of a particular study effect, we need to establish that the effect is accurate and approriate.

In this setting other explantory variables are only really imporant in terms of controlling for **confounding** and checking for **effect modification**. In all other respects we are not that interested in 'other' risk factors.

For our example today this is our approach.

Exploratory studies

In this setting we know little about the system under study, and we are trying to identify which risk factors play an important role in the 'epidemiology' of a disease.

Here all of the variables are considered (potentially) equally important in the modelling process, and we may want to 'throw everything in' and 'see what sticks'.

This is common in prevelance studies where the research objective is *'...aim is to estimate the prevalence of disease Z and investigate it's risk factors....'.*

The approach we use to build our model (select the 'best' model) needs to be considered in light of our research objectives (Predictive, hypothesis testing, exploratory).

I will discuss four different approaches to 'build our model':

1. **FORCED**: Throw everything in the model
2. **COMPUTER ALGORITHM**: Forward selection, Backwards elimination, Stepwise and best subsets
3. **HEIRACHICAL**: Collect variables into blocks (and then choose one of the above WITHIN each block)
4. Purposeful selection of covariates (PSC)

'Throw it all in' method

This is where we **FORCE** all of our covariates into the model.

This approach is OK if we just have a few (e.g. < 5 predictors), and our objective is to 1. Explore the relationships, or 2. Find a predictive model.

However, it doesn't really fit with a hypothesis testing framework. It doesn't allow us to obtain both crude and adjusted estimates of our effects (and therefore report on confounding effects.....which is a story that needs to be told in hypothesis testing studies)

Also, if there are a lot of predictors, there is a risk that the model will be overfit: That is, it might work well on the sample, but not be generalizable (externally valid) for the target population.

Computer algorithm approaches

A number of these available:

1. Forward (stepwise) selection
2. Backwards (stepwise) elimination
3. (full) Stepwise (Hybrid of 1 and 2)
4. Best subsets

These methods allow the computer(software) to make the decision on which variables should, and should not, be included in the model.

Computer algorithm approaches: Forward selection

The forward selection method:

1. Start with empty (null) model
2. Identify covariate most correlated (and $p<0.05$) with the outcome, add to the model
3. Of the covariates remaining, identify the next most correlated, add to the model and keep if significant ($p<0.05$)
4. Repeat step 3 until variable added to the model is not significant

Computer algorithm approaches: Backwards elimination

The Backwards elimination method:

1. Start with full model (all covariates included)
2. Identify covariate least correlated (and p$>$0.05) with the outcome, remove from model
3. Of the covariates remaining in the model, identify the next least correlated (and p$>$0.05), remove from model
4. Repeat step 3 until there are no more 'non-signifcant variables' remaining

Computer algorithm approaches: Forward selection

The forward selection method:

1. Start with empty (null) model
2. Identify covariate most correlated (and p$<$0.05) with the outcome, add to the model
3. Of the covariates remaining, identify the next most correlated, add to model
4. Look at model, see if any of the variables have become non-signifcant, if so remove
5. repeat steps 3 and 4 until we get a model that can no longer be improved

All three of the 'stepwise' models can use a number of criteria to decide on the inclusion (exclusion) of a covariate. These criteria can be at a variable (e.g. test of the individual β) or at the level of the model (No improvement in the F-statistic, or model R^2)

Computer algorithm approaches: Best subsets

The *Best Subsets* approach is a computational intensive algorithm that examines all possibles models (based on all permutations of covariates possible). It tends to use criteria like Mallow's C_p or Information Crietria (e.g. AIC, BIC etc.).

The advantage of the best subset approach (over other algorithmic approaches) is these criteria not only account for model fit (e.g. R^2) but also penalized for the model complexity (e.g. Number of variables in the model).

Problem: A little "Magic". Decisions are taken out of the hands of the researcher

Problem with computer algorithm approaches

There are a number of distinct problems with these 'automatic' ways of arriving at the 'best' model:

1. **Black box:** They take decisions away from the researcher. Contextual knowledge required in model building
2. **Multi-category predictors:** As categorical predictors (factors) are coded using at least 2 dummy (indicator) variables, we can have a dummy variable (from a set) removed from the model (Doesn't make sense and would totally destroy our sample size)
3. **Confounding:** None of the above approaches examines the impact of a added/removed variable on the effect of an existing model βs (confounding).
4. None of these approaches consider effect modification

We need an approach that allows us to make decisions (we understand the clinical implications), and uses the power of the computer to do all the 'heavy lifting'

Computer algorithm and study objectives

Using computer algorithm approaches is much worse for some research objectives than others. Think about this example:

We are interested in testing whether HT represents a risk factor for achieving diabetes clinical outcome (objectives). We use a stepwise approach, but find HT is excluded in the early stages (it doesn't appear in our 'best' model).

What do we conclude about the effect of the HT comorbidity on achieving diabeties clinical objectives???

Heirachical model building

The heirachical approach is very popular in psychology and the other social sciences. It involves themeing covariates into **BLOCKS**. For example:

Consider a study of schizophrenia where we are interested in certain SNPs (genetic markers) representing a genetic risk factor. We know that certain demographic and lifestyle exposures also represent risk factor for schizopherinia (most importantly certain illicit drug use). So in our study we theme our covariates into three blocks:

1. BLOCK 1:Demographics variables Sex, Age and Socio0economic status
2. BLOCK 2: Drug use history (especially cannabis and hallucinogens)
3. BLOCK 3: Genetic markers SNPa, SNPb, SNPc and SNPd

Heirachical model building

Then we:

1. Use **forward selection** for the demographics (we only want to include the demographics that impact on schizophrenia)
2. **Force** the drug use history variables (previous literature tells us they are important)
3. Individually test whether SNPa, SNPb, SNPc or SNPd is associated with schizophrenia

Advantage of the heirachical approach is we can 'residualize' the outcome variable (account for variation due to risk factors that don't interest us, but still relate to the outcome)

Disadvantage is that while we may (inadvertantly) adjust for some confounders, we can't comment of the effect of confounders (no comparison between crude and adjusted estimates).

Purposeful selection of covariate approach

A **WONDERFUL** approach to model building where we can be both empirically (from the results) and contextually guided.

I will cover two versions of the purposeful selection of covariate (PSC) approach:

1. Vanilla - PSC (*sensu* Lemeshow, Hosmer and May, 2008): This method treats all covariates as equally important (no study effect)
2. Modified - PSC (where we do have a study effect)

Vanilla-PSC

This approach to model building is a 7 step process:

1. **Crude effects**: Run a set of bivariate models for each covariate against the outcome of interest (Noting the effect: e.g. OR, and p-value)
2. **Initial multivariable model**:
   1. Include all covariates whose crude effects have a $p<0.25$ in a multivariable model and run it
   2. Remove all covariates not significant at 0.05 level (from model in step 2a)
3. ONE AT A TIME, reintroduce the variables removed in step 2b to see if:
   1. They are significant ($p<0.05$)
   2. They represent confounders: They substanitally change the effects of covariate already in the model (e.g. by 20%)
4. SAME AS STEP 3, but this time reintroduce the variables excluded in the bivariate analysis (Step 1).

Vanilla-PSC continued

Continuing:

6. Consider the **linearity** of the contiuous covariates against the outcome (difficulty of this depends on statistical method and software). After this step we have a **Main effects model**
7. (If approriate) **Effect modification**: Judiciously choose interaction effects that might go into the model
8. **Model adequacy**: Assess the significance, fit and validity (assumptions) of final model

This seems VERY involved, but once you have done it once or twice, it makes a lot of sense.

Question: Why did I use p$<$0.25 in step 2a??

Modified-PSC

In this version we have a study effect (a covariate that is more important to us than all the others). To account for this we slightly modify the Vanilla PSC (modifications in red)

1. **Crude effects**: Run a set of bivariate models for each covariate against the outcome of interest
2. **Initial multivariable model**:
   1. Include **the study effect (FORCED) AND** all covariates whose crude effects have a p<0.25 in a multivariable model and run it
   2. Remove covariates not significant at 0.05 level (from model in step 2a), **but keep the study effect in the model, regardless of its significance**
3. ONE AT A TIME, reintroduce the variables removed in step 2b to see if:
   1. They are significant (p<0.05)
   2. They represent confounders **of the study effect**: They substanitally change the **the study effect** (e.g. by 20%)

Modified-PSC

Continuing:

5. SAME AS STEP 3, but this time reintroduce the variables excluded in the bivariate analysis (Step 1).
6. Consider the **linearity** of the contiuous covariates against the outcome (difficulty of this depends on statistical method and software). AFter this step we have a **Main effects model**
7. (If approriate) **Effect modification**: Judiciously choose interaction effects that might go into the model. **Only include interactions involving the study effect**
8. **Model adequacy**: Assess the significance, fit and validity (assumptions) of final model

At first glance, the PSC approach seems very complex and involved but once you have used it (and see the logic). It makes a lot of sense.

We will run through two examples (one linear regression and one logistic regession) using the modified-PSC on the DMHT dataset.

REMEMBER: DOCUMENT YOUR DO FILE!!!!!!

Pros and cons of PSC

The disadavantage of the PSC approach to model building is that it is rather involved (and we, rather than the computer) have to do some of the work. BUT the advantages far outweigh the disadvantages. The advantages are:

- We get a VERY good feel for what is going on
- Unlike all other approach we can account for confounding (and examine its effect)
- Unlike all other approaches, we can consisder effect modification
- PSC is a very robust approach. If you present in well in your methods section, it is VERY difficult for a reviewer to criticize.
- It also considers the model adequacy (especially important for some types of models)

THANK-YOU!!

Questions??

YOUR TURN

Exercise: Using modified PSC on the DMHT dataset

We will run through two example of modified PSC:

1. A the continuous outcome hemoglobin A1C so we will use a General Linear Model (aka Linear Regression) approach
2. The Binary variable of hba1cyn (achieved clinical outcome: yes/no). In this case we will consider Binary Logistic Regression

In both cases we will consider the explanatory variables:

- **Hypertension comorbidity (ht=no, yes)**
- outpatient clinic type (opdg = general, specialized)
- sex (males, females)
- age
- bmi
- duration of diabetes (duradm)