# Model-building in regression models

**Morten Frydenberg ©**
**Department of Biostatisics, Aarhus Univ, Denmark**

**Which model should I use?**

- Models are approximations
- What is the focus
  - One primary exposure - adjustment
  - Aetiology
  - Prediction

**Some comments on confounding and adjustment.**

**Automatic model selection**

- Why you never should use it

**Over-fitting**

**A general strategy – ten steps to heaven**

# Which model should I use?

This a hard question!

The first thing to remember is that all models are **approximations**!

The "true" , the "best" or the "correct" model **does not exist!**

The **quality** of a model depends on what you want to use it for.

So the first thing to clarify is:
What is the **purpose** of your analysis – what is the **aim** of your data collection?

Different purposes – different models!!!!!

When you have found out what you want, you still have an **infinity** of models to choose between.

# Which model should I use?

The choice is always a choice between **complicated** and **less complicated** models.

**Complicated** models are often better models, in the sense that they are **better approximations** to the truth.

But complicated models can be:

Very hard to **estimate** – many parameters.

Very hard to **understand**.

Very hard to **communicate**.

So in these senses they are **not so good** models.

# Which model should I use?

**Less complicated** models are often not as good models, in the sense that they are **not so good approximations** to the truth.

But less complicated models can be:

Easy to **estimate** – few parameters.

Easy to **understand**

Easy to **communicate**

So in these senses they are **better** models.

The first thing to remember is that all models are **approximations**!

**Statistical significance** has nothing to do with the quality of the model!

# Which model should I use?

You can often divide the explanatory variables **into groups**:

1: Variables of **primary interest- main exposure.**

2: Variables of **less interest** – variables you want to "**adjust**" for.

A good model will try to introduce the **first** group in an **interpretable** way into the model.

- You want to **know** "how they work".

E.g. if you specifically are interested in the "effect" of age you should model age in an **understandable** way.

Still you have to look out for collinearity.

# Which model should I use?

The **second** type of variables can be introduced any way you like.

It can be very complicated – you do not care- as long as they do the job - that is, **adjust sufficiently**.

If you are not interested in age itself – you just want to adjust – then age can be introduce in a complicated/weird way, e.g. a fourth order polynomial.

In **general**:

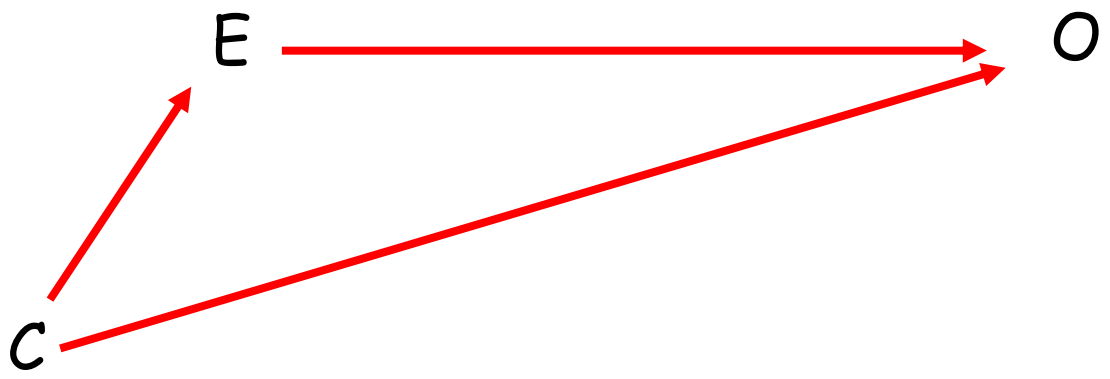
Models with many parameters need more data to obtain precise estimate.

Again few data – lower your ambition !

# Confounding and adjustment?

The traditional, but not very precise definition of a confounder, *C*, is:

1. *C* is a risk factor for O (the outcome)
2. *C* is associated with E (the exposure)
3. *C* is not in the pathway between E and O



# Confounding and adjustment?

When estimating the effect of an exposure, **E**, on an outcome, **O**, many authors routinely state that they want to adjust for confounders.

Often this is done without confounding is a relevant concept in the specific analysis.

Consider **E=smoking** and **O="problems with the heart"** and the covariates:
sex, age, alcohol consumption, BMI and daily exercise.

As we want to estimate the 'effect' of smoking on a personal level, i.e. how would my risk change, if I change my smoking habits, a relevant model should **include known risk factors** like sex and age. **Even if smoking habits are independent of sex and age.**

# Confounding and adjustment?

Whether or not one should adjust for BMI, alcohol consumption and daily exercise is a more complicated question.

If we adjust for alcohol consumption, then we look at the effect of changing smoking habits without changing drinking habits.

If we do not adjust for alcohol consumption we might have a confounding problem as some of the smoking **observed** effect might be due to 'extensive drinking' among smokers.

A similar problem exists for daily exercise as the association between smoking habits and daily exercise is bidirectional.

In short, which variables/information you want/need to include in your model, can seldom be determined by the simple confounding concept, but depend on the purpose of your analysis.

# Automatic model selection

Some programs (even Stata) have programmed algorithms for **automatic model selection**!

That is, procedures that will find the "best" model to answer your question without knowing what **you want**, **know** or anything else about the **problem**!

It is very rarely of any interest, especially if you have **little data**.

There are in general three types of such algorithms:

**Backward selection** : You specify a **start model** and the procedure will reduce the model by **removing** variables from the model until nothing can be removed.

The **criteria** for removing variables are typically **high p-values**.

# Automatic model selection

**Forward selection**: You specify a **start model** together with a **list** of variables that might be included in a model. The procedure will build the model by **adding variables** from the list to the model until nothing can be added.

The **criteria** for adding variables is typically **low p-values**.

**Best subset selection**: You specify a **list** of variables that might be included in a model and a **number** of variables you want in the model. The procedure will then search among the possible models and find the "best".

The **criteria** is typically the **highest likelihood** or related statistics.

# Automatic model selection

Some comments:

- These procedures **do not know anything about the subject**.
- They will not consider **transformation** of the variables.
- or **interaction**.
- They will choose **arbitrarily** between explanatory variables that are highly correlated.

# Model selection - some implications

Even when you do not use an automatic model selection procedure: The **final** model is selected!

That is, you have spent some time **working** with the model you present!

You might choose only to include **statistically significant** variables in the model.

You might group **two levels** of an explanatory variable **into one level** if there is no statistically significant difference between the two levels.

The implications of this selection:

- The **estimates** tend to be too far **away from null**.
- The **standard errors** are too **small**.
- The **CI's** are too **narrow** and the **p-values** too **small**.

# Over-fitting

Overfitting is what you get if you fit a very complicated, (i.e. a model with many parameters) to a small data set.

The fitted model will fit the data set too well and hence tell more about the actual data set and less about the general structure.

In order not to overfit one have to balance the complexity of the model and amount of information in the data.

# A general strategy

Start by clarifying **the purpose** of your study!!

Prediction / aetiology / exposure-study...

**Step one**: **Outcome**

Decide on your outcome variable.

**Step two: List of explanatory variables/covariates**

Make a **prioritised** list of the variables/information you would like to include in the model.
This list should be made solely based on **subject-matter** knowledge and the **purpose** of your study.

Often you would benefit by dividing it into **blocks** of information or variables, e.g. measures of blood pressure, measures of body composition, information on disease history , comorbidity, demography, previous births etc.

# A general strategy

**Step three**: Design and data

Decide on an **appropriate design** and **collect your data** according to the list you made in step one and two.

**Step four**: The maximum complexity of the model

Decide on the **maximum number of parameters**, P, allowed in your model, using the following rules:

**Continuous** outcome: P=n/15

**Binary** outcome: P=number of events/15

( i.e. if you have 100,000 persons but only 200 events, then P=200/15=13.3, so you model can only contain 13 parameters)

# A general strategy

**Step five**: **Explore the explanatory variables**

Explore the distribution of and association between variables on the list of explanatory variables – start **from the top** of your list.

This might guide you to choice of cutpoints, scales and potential problems with collinearity.

**Step six**: **List of interactions/effect modifications**

Make a **prioritised** list of effect modifications.

# A general strategy

## Step seven: Allocating the P parameters

This is the tricky part.....

Carefully, go through your list of explanatory variables/blocks of variables and interactions and decide, how **many parameters you will spend on each of them.**

Remember:

| | |
|---|---|
| one continuous variable | = 1 parameter |
| one binary variable | = 1 parameter |
| a categorical variable with k levels | =(k-1) parameters |
| one continuous variable with squared term | = 2 parameters |
| cubic splines with k knots | =(k-1) parameters |
| interaction with a binary and a cont. var | =1 parameters |
| interaction between two binary | =1 parameters |
| interaction between two categorical | =(R-1)*( C-1) par. |

# A general strategy

## Step eight: Representing blocks

At this stage you might have decided that you are willing to use **two** parameters on "body composition", but you have several variables in this block: BMI, waist-hip-ratio etc.

Now you have to decide on how to use the **two** parameters:

- Include BMI (continuous) and waist-hip-ratio (continuous) ?
- Only use BMI but divide into three levels?
- Only use waist-hip-ratio but divide into three levels?
- Only use BMI use with cubic splines with three knots?
- Only use waist-hip-ratio with cubic splines with three knots?
- Combine the variables into two 'dimensions' of body composition?

# A general strategy

## Step nine: Choosing scales, cut-points and knots

Imagine that you have allocated **one** parameter to a variable, and whether it is as a **continuous** variable or as a **binary** variable.

In the first case you have to decide on whether or not the variable should be **transformed**, e.g. using log to model that relative difference have constant effect.

In the second case you have to decide on a **cut-point**.

Imagine that you have allocated **k** parameters to a variable, and whether it is as a **continuous** variable or as a **binary** variable (or a mixture).

In the first case you have to decide on whether or not the variable should be **transformed**, e.g. using log to model that relative difference have constant effect and where the **k**+1 knots should be.

In the second case you must decide on the **k** the **cut-points**.

# A general strategy

**Step ten: Choosing scales, cut-points and knots of pairs of interacting variables**

If you have allocated **k** parameters for two variables and their interaction, then you have to decide how the **k** parameters are going to be used on 'main' effects and interactions.

Again you have the choice between continuous and categorical variables – scales, knots and cut-points.

# A general strategy

## Step seven to ten

- You often have to go through these steps several times in order to get 'the best' balance between the number of variables and the complexity of each variable/interaction.
- In none of this are you allowed to fit/estimate a model or to look at the outcome variable.
- In other words, the outcome is only used to determine **P** the maximum complexity of the model

When you have finished step one to ten then you have decided on your model and you can fit it!!!!

# A general strategy

## Further comments

- If you are working with a continuous **outcome** , you should of course also decide on the **appropriate scale** for this.
- I have here assumed that all observations are **independent** and all **effects are systematic**. Otherwise you also have to take this into consideration in your model building.
  I.e. how to model the **correlation structure** and the **random effects**.
- I have also ignored the problem with **missing** values, i.e. that you have missing values for some of your covariates. Only analysing persons with complete information will lead to loss of efficiency (power) and maybe bias.
  In some situations the problem can be solved by '**imputation**' – a statistical technique were one 'generates' the missing values.

# A general strategy

## Checking the model

- Checking the model is not the same as building the model! You check the model to see if there are **serious problems** with it.
- You do not check by looking at p-values!
  All models are wrong and with enough data any goodness-of-fit test will show that.
- A model is not wrong because some of the coefficients are statistically insignificant!!!!!!
- You check the model by looking at features of the model that worry you. Typically by inspection of the residuals.
- Only **serious** problems will suggest that you have to alter the model.

# A general strategy

## Checking the data

After you have fitted your model you should check the data for points with high residuals or leverage.

These data should be checked for errors and corrected.

You might refit the model without these data-points in order to see if your conclusion depend heavily on these 'strange' data.