VHM 812/802: Model Building Exercise

Data

This model building exercise will be based on a dataset from a cross-sectional study that investigated the effect of pig diseases on growth performance. The dataset is called: "pig_adg.dta" and includes the variables listed below (plus more, but these are the ones we will use for this exercise).

Variable	Description	Range
farm	Farm identification code	1-15
pig	Pig identification number	1-341
adg	Average daily gain in *g*/day	continuous
sex	Sex of the pig	0=female; 1=castrated male
worms	Count of nematodes in small intestine at time of slaughter	continuous
ar	Atrophic rhinitis score	0-5
pn	Enzootic pneumonia	0=absent; 1=present

For the following questions, we assume:

- -adg- varies between farms (due to different rations etc.),
- male and female pigs grow at different rates,
- all 3 "diseases" (worm infestation, -ar-, -pn-) might affect -adg-,
- one way that atrophic rhinitis severity (-ar-) might impact on -adg- is by increasing the susceptibility of the pig to -pn- (by destruction of the air filtering capacity of the turbinate bones),
- worm burdens are not related to either -pn- or -ar- (there is no biological reason to think they might be),
- it could be meaningful to consider quantifying the impact of -ar- only by the presence of high -ar- scores (corresponding to severe atrophic rhinitis), so define a dichotomous variable to represent -ar- scores > 4 (i.e., -ar_sev- = 1 when -ar- > 4),
- our primary interest for modelling is how any of the 3 "diseases" affect -adg-.

Use the data and the information above to go through the following analytical steps.

1. Identify the primary outcome of interest and main predictors of interest.
2. Draw the causal diagram for your full causal model.
3. Identify which variables are potential confounders for the disease –> outcome relationships of interest.
4. Identify any variables that might be considered intervening (intermediate) variables.
5. Identify any exposure-independent variables.

6. You have a very limited number of predictors in this dataset so we will not actually eliminate any on them on the basis of descriptive statistics or lack of unconditional associations. However, go through the exercise of computing descriptive statistics for each variable, evaluate unconditional associations and carry out a pairwise correlation/association analysis among all predictors. If you were looking to eliminate potential predictors at this stage, are there any likely candidates?
7. Decide what 2-way interactions you want to examine.
8. Use forward selection, backward elimination and stepwise selection procedures to identify potential models for further investigation; include in your exploration also relevant model fit criteria.
9. Evaluate potential confounding effects by forcing all removed predictors that may be confounders, back into the model. Do any of them need to be kept, even though not statistically significant, because they appear to exert a confounding effect?
10. From here on, we will focus on the effect of -ar_sev- on -adg-. Identify the model which best evaluates this effect.
11. Evaluate the reliability of the model. (We will assume that you have already evaluated the (internal) validity of the model using the usual regression diagnostics.) Compute also the PRESS statistic, and explain what it tells you about the fitted model.