NHANES project

Feedback

General feedback

- Confusing terms confounding, correlation, collinearity and interaction
- Some variables are just proxies of each other (e.g. physical activity variables, waist and bmi). You don’t need all these to go into your model selection necessarily.
- Don’t just discard some of the explanatory variables because they don’t occur in the research question. Including them will increase precision.
- Beware of obvious confounders which should be included
- Model parsimony
- Include some summary statistics!
- Be careful not to make statements implying causality

Confounder

A confounder is a variable linked to both the exposure (E) and the response (Y), but is not on the causal pathway.

Omitting a confounder from the analysis causes bias: it may introduce a spurious relationship between exposure and response or may increase, decrease or cancel out the relationship.

Confounding

- Confounding occurs where additional variables are associated with both the response and the variable of interest.
- The conditions for considering a factor C to be a potential confounder of the relationship between the explanatory variable E and the response/disease Y are
- C has a causal effect on Y
- C is associated with E, but not a proxy/surrogate for E
- C is not an intermediate step in the causal path between E and Y

Confusing interaction with association between predictors.

Interaction:linear models allow us to include the effects of several predictors of mixed type (metric and/or factor) in an additive way — with the effects of the variables simply added to each other.

But what about the case in which the effects of variables do not simply add up, but rather the effect of one variable is to modify the effect of another variable. In statistical modelling this is known as an interaction (or effect modification).

Interaction is **not** about the relationship of one predictor to another. It is about their combined effect on the response.

Miscellaneous

- You should not define factor levels into separate variables.
- In a linear model you should talk about effects rather than correlations.
- Backward selection is applied using the function step(). Better say, backward selection using the AIC criterion was applied.
- Clearly signpost section
- Describe pre-processing in a coherent manner so that someone else could repeat it.
- Bi-variate relationships don’t tell the whole story. So don’t decide which variables will have an effect just based on this. Some variables might only show a relationship with the response once you have adjusted for another variable.
- Backward selection is preferred to forward selection. Why?

Format

- Formating of markdown: need to make sure that warnings etc are not printed.
- Section headings are very important
- Executive summary needs to be there!
- Conclusions

Present some summary statistics! e.g.

There are 924 observations, with 426 men and 498 women, and the average participants are aged 39. Almost half of them are Non-Hispanic Whites, and more than half are married. 79% of the participants have at least graduated high school, and those with high income (>$20K) make up about 84% of the sample. The average BMI of the participants is 29.15, which is in the overweight range , nearing the obese range (BMI>30.0), while the average apoliprotein level of 103.6 falls in the normal range .

The mean step counts per day is 342.88. On average, participants spend around 55% of the time sedentary, around 36% on light intensity, 11% on lifestyle intensity and 3% of the time on moderate-to-vigorous physical activities. Also, the average sedentary minutes accumulated in bouts of length greater than 30 minutes is around 105 minutes, and note that there are people who do not spend in sedentary for more than 30 minutes. Lastly, the mean number of MVPA bouts per day is 0.1764 which can be interpreted as on average, people do not spend more than 10 minutes of moderate-to-vigorous activities at all (since the mean of this variable is not even close to 1!).