Chapter 8: Linear models in the real world: overfitting, collinearity, confounding

1. In this exercise, you will explore data from the AMATUS dataset, which includes psychological and educational measures from a sample of German university students. The study aimed to investigate how individual differences in anxiety, self-concept, and personality are associated with arithmetic performance. The dataset includes a range of validated psychological scales[1].

   Your goal is to test whether *math anxiety* (as measured by the Abbreviated Math Anxiety Scale, AMAS) predicts performance on an arithmetic task. You will begin by fitting a simple regression model with only math anxiety as a predictor. Then, you'll extend the model to include other psychological and demographic predictors and reflect on how and why the results change.

   The dependent variable is:

   - `sum_arith_perf`: Total number of correct answers on an arithmetic test (max = 40)

   Predictors include:

   - `score_AMAS_total`: Math anxiety (Abbreviated Math Anxiety Scale)
   - `sex`: Participant sex (coded as "m" or "f")
   - `age`: Age in years
   - `math_grade`: Final school math grade (1 = best, 6 = worst)
   - `score_GAD`: General anxiety (GAD-7)
   - `score_STAI_state_short`: State anxiety (Kurz STAI)
   - `score_TAI_short`: Test anxiety (Short Test Anxiety Inventory)
   - `score_SDQ_M`: Math self-concept (Self Description Questionnaire III)
   - `score_SDQ_L`: Language self-concept (Self Description Questionnaire III)
   - `score_PISA_ME`: Math self-efficacy (from PISA items)
   - `score_BFI_N`: Neuroticism (BFI-K short version)

   - Fit a simple linear regression model predicting `sum_arith_perf` from `score_AMAS_total`. Is math anxiety associated with arithmetic performance in this model? What does the slope estimate suggest about the nature of this relationship?
   - Now fit a multiple regression model predicting `sum_arith_perf` using all the predictors listed above. How does the coefficient for `score_AMAS_total` compare to the one in the simpler model – in terms of both magnitude and statistical significance?
   - Compare the R^2 and adjusted R^2 values from the two models. How do they differ? What might the adjusted R^2 tell you about the added value of the additional predictors?

[1] Cipora, K., Lunardon, M., Masson, N., Georges, C., Nuerk, H.-C., & Artemenko, C. (2024). 'The AMATUS Dataset: Arithmetic Performance, Mathematics Anxiety and Attitudes in Primary School Teachers and University Students.' *Journal of Open Psychology Data*, 12 (1), 10. https://doi.org/10.5334/jopd.115.

- Consider your results. What changes most noticeably when moving from the simple to the full model? What might account for these changes?

2. In Chapter 7, you analysed historical data from mid-19th century British counties to examine John Clay's claims about the relationship between beerhouses, school attendance, and crime. You found that the number of beerhouses per capita was positively associated with recorded crime rates, and that this relationship varied depending on school attendance. But does this imply that beerhouses cause crime?

    - Think critically about the nature of the dataset. What kinds of confounding factors might be influencing both the number of beerhouses and the level of crime in each county? Consider social, economic, or geographic factors that could plausibly affect both.
    - One possibility is that *urbanisation* (i.e., whether a county is more rural or urbanised) could influence both the number of pubs and the level of crime. In the 1850s, more populous or industrialised areas might have had more drinking establishments and more opportunities for crime – independent of any direct causal link. Could this variable be a *confounder*?
    - Sketch a Directed Acyclic Graph (DAG) representing a possible causal structure involving:
        - Beerhouses per capita.
        - School attendance.
        - Crime rates.
        - Urbanisation (or another plausible confounding variable of your choice)

      (There are no right or wrong answers here – we do not know the true causal model. However, try to draw a DAG that is plausible based on your background knowledge, and that you would feel confident justifying.)

3. A simple linear regression shows a positive association between the number of hours students report studying and their final exam mark. However, you later obtain information about each student's A-level tariff points (a measure of prior academic attainment) and notice that tariff points are positively correlated with both study hours and final marks.

    - Explain qualitatively how omitting tariff points from the regression model is likely to bias the coefficient for study hours. In what direction would you expect the bias, and why?
    - Draw a DAG to represent the relationships between study hours, final mark, and A-level tariff points. Use your diagram to identify any backdoor paths, and explain how including tariff points in the model would help block them.

4. Reflecting on the Bechdel Test example, now consider the release year of the movie. Create one or more potential DAGs that illustrate how the year might causally relate to the movie's budget, gross, and its likelihood of passing the

Bechdel Test. Think about how trends over time, such as changing cultural norms or shifts in the film industry, could influence these relationships.

5. A researcher is interested in the relationship between attentional control and cognitive test performance. They recruit participants from a highly selective university subject pool, and restrict their analysis to individuals who scored above the 80th percentile on either an attention-control task or a working memory test (i.e., participants were selected if they did well on at least one of these measures).

    - The researcher finds a negative correlation between attention-control scores and working memory scores in their sample. However, previous research shows these abilities are typically positively correlated in the general population. How could this discrepancy arise?
    - Consider whether the selection procedure might have introduced *collider bias*. What variable is being conditioned on, and how could this open a spurious path between attention and working memory?
    - Draw a DAG representing the relationships among:
        - Attention control.
        - Working memory.
        - Selection into the sample (i.e., high score on at least one task).
    - Use your DAG to explain how conditioning on selection can induce a negative association between two otherwise unrelated or positively related variables.
    - Reflect: What broader lesson does this example illustrate about interpreting correlations in non-random samples?