DSC 241: Statistical Models, Assignment 2

Winter 2026

Instructions

1. **Deadline.** The submission deadline is *January 23 at 22:00:00*. Late submissions will incur a penalty of 1 point per hour (0.8 from Problems and 0.2 from the Project), applied for up to 50 hours. Submissions received after *January 25 at 23:59:59* will not be accepted under any circumstances.

2. **Use of Large Language Models (LLMs).** The use of LLMs is *permitted and encouraged*. You may use LLMs to help explain the questions, obtain hints, and/or generate draft solutions for reference. *LLMs may produce incorrect or misleading content, and you remain fully responsible for the correctness of your submitted answers.* If you use an LLM, you must explicitly disclose this and clearly describe how it was used. If you do not use an LLM for one or more questions, you must explicitly state this as well. You are not required to provide verbatim transcripts; instead, briefly summarise your LLM usage and the type of assistance it provided. *Failure to provide an accurate and genuine disclosure constitutes a violation of academic integrity and may result in point deductions or a nullified grade.*

3. **Contribution statement.** Include a concise yet complete description of the individual contributions made by each team member. *Failure to provide an accurate and genuine contribution statement may result in a deduction of up to 5 points.*

4. **Clarity and conciseness.** Answers should be clear, concise, and directly address the questions. *Redundant, irrelevant, unclear, or unnecessarily verbose responses may result in point deductions.*

5. **Quality of graphics.** Where applicable, present high-quality graphics, charts, and tables with appropriate axis labels, legends, colors, line styles, and font sizes. Avoid unnecessary duplication of figures when results can be clearly and effectively combined for comparison. *Redundant, irrelevant, unclear, or unnecessary graphics may result in point deductions.*

Problems (80 points)

The purpose of this problem is to illustrate Simpson's paradox and the linear projection of conditional expectation functions (CEFs).

Let two random variables X and Y be bivariate Gaussian with mean (μ_X, μ_Y), common variance σ^2, and correlation coefficient $\rho > 0$. The bivariate mean itself is random and is determined by a third variable Z (referred to as a confounder) according to

$$(\mu_X, \mu_Y) = \begin{cases} (-\mu, \mu) & \text{if } Z = -1, \\ (\mu, -\mu) & \text{if } Z = 1. \end{cases}$$

Here $Z \sim$ Rademacher, meaning that Z takes the values -1 and 1 with equal probability $\frac{1}{2}$.

1. **(25 points)** Conditional expectations:

   (a) **[Math]** Derive mathematical expressions for the three conditional expectation functions $\mathrm{E}[Y \mid X, Z]$, $\mathrm{E}[Y \mid Z]$, and $\mathrm{E}[Y \mid X]$. Comment on whether each CEF is linear in its corresponding predictors X and Z.

(b) **[Coding]** Set $\mu = 1$, $\sigma^2 = 1$, and $\rho = \frac{1}{2}$. Plot the three CEFs derived above as functions of X over the range $[-5, 5]$. Produce a separate plot for each CEF. Confirm their (non)linearity.

2. **(20 points)** Best linear approximations:

   (a) **[Math]** Compute the best linear approximations to the three CEFs obtained above.

   (b) **[Coding]** Using the same parameter settings as in Problem 1(b), overlay the best linear approximations on the corresponding CEF plots. Comment on the quality of these linear approximations.

3. **(25 points)** Conditional residual variances:

   (a) **[Math]** Derive mathematical expressions for the conditional residual variances $\text{Var}[Y - \text{E}[Y \mid X, Z] \mid X, Z]$, $\text{Var}[Y - \text{E}[Y \mid Z] \mid Z]$, and $\text{Var}[Y - \text{E}[Y \mid X] \mid X]$. Show that increasing the number of predictors reduces the conditional residual variance.

   (b) **[Coding]** Using the same parameter settings as in Problem 1(b), plot the three conditional variances as functions of X over the range $[-5, 5]$. Produce a separate plot for each conditional variance. Verify your results from Part (a).

4. **(10 points)** Simpson's paradox:

   (a) **[Research]** Study Simpson's paradox and explain why this setup constitutes an example.

   (b) **[Math]** Determine the conditions on the parameters μ, σ, and ρ under which Simpson's paradox arises.

Project (20 points)

1. **(5 points)** In your dataset, produce appropriate univariate graphical representations of the variables in the dataset according to their type (e.g., histograms, bar plots, etc.). Comment on the distribution of the variables, if they are well balanced, if there are suspected outliers, etc.

2. **(5 points)** Based on the research article associate with your dataset (if there is one), write down the research question or questions (at most 3) that you wish to solve in the context of the data. If the questions involve one or several response variables, identify those variables and the relevant predictors. Be sure to include questions involving inference, not just prediction.

3. **(10 points)** In your dataset, produce a pair plot of bivariate scatter plots. Place the response variable or variables in the upper left corner. Comment on the distribution of the variables, if there are nonlinearities, high correlations between predictors, suspected outliers, etc.