DSC 241: Statistical Models, Assignment 3

Winter 2026

Instructions

The important updates are highlighted for your convenience.

1. **Deadline.** The submission deadline is *January 30 at 22:00:00*. Late submissions will incur a hourly penalty of 0.8 point for Problems section and 0.2 point for Project section, applied for up to 50 hours. Submissions received after *February 1 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 can 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. You still need to do so even if you are on your own. *Failure to provide an accurate and genuine contribution statement may result in a deduction of up to 4 points for Problems sections and 1 point for Project section.*

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, illegible, or unnecessary graphics may result in point deductions.*

6. **Present your code.** Although it may not be explicitly required in every question, you should provide the code used wherever applicable. *Failure to include the relevant code may result in point deductions.*

7. **Gradescope page assignment.** You need to assign pages for each part of the question correctly when you submit on Gradescope. *Failure to provide an accurate page assignment may result in point deductions.*

Problems (80 points)

The goal of this problem is to better understand the theory behind the sandwich formula.

1. **(30 points) [Math]** Let random variables $X \sim \text{Unif}[-1, 1]$ and $Z \sim \mathcal{N}(0, 1)$ be independent, and $Y = XZ^2$.

   (i) Compute and plot the conditional expectation function (CEF) $E[Y \mid X]$ and the conditional variance function (CVF) $\text{Var}[Y \mid X]$. Present your code and plots.

(ii) Compute and plot the best linear projection of the CEF $\beta_0 + \beta_1 X$ as a function of X. Present your code and plots.

(iii) Suppose we observe n i.i.d. samples (X_i, Y_i). Compute the asymptotic variance of the fitted coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$.

(iv) Compare whether the asymptotic variances computed above with the asymptotic variances computed under the standard assumptions. Explain the discrepancy if there is one.

The goal of the next two problems is to check the validity of the confidence bands.

2. **(20 points)** **[Coding]** Write the following functions: a function `confBand(x, y, conf)` that takes as inputs a predictor vector $\vec{x} = (x_1, \ldots, x_n)$ and a response vector $\vec{y} = (y_1, \ldots, y_n)$, and returns a plot. The plot should include the points $(x_1, y_1), \ldots, (x_n, y_n)$, the OLS population regression line, and four types of uncertainty bands:

- Pointwise confidence bands, under standard assumptions
- Pointwise confidence bands, using the sandwich formula
- Simultaneous confidence bands using the Bonferroni correction, under standard assumptions
- Simultaneous confidence bands using the Bonferroni correction, using the sandwich formula

The confidence level is given by the input parameter `conf`.

Test your function on the following settings: $n = 10$, $x_i = i$, $y_i = 2^i$, and `conf = 0.95`.

(i) Present your code and plot.

(ii) Briefly describe and explain your observation.

3. **(30 points)** **[Coding]** Perform a Monte Carlo simulation to evaluate the coverage of each of the four types of bands. Repeat the following experiment $N = 1000$ times:

- Generate a bivariate dataset of $n = 100$ pairs (X_i, Y_i) where $X_i \sim \text{Unif}[0, 1]$, $Y_i = X_i^3 + 0.5 X_i Z_i$ where $Z_i \sim \mathcal{N}(0, 1)$ is an independent random variable.
- Compute the four type of confidence bands for the OLS population regression line as described in Problem 2 at 95% level, and record whether they respectively cover the true regression line at all X_i.

(i) Plot one of your experiment as an example. Describe your observation.

(ii) Summarise the result by aggregating the proportion of times that each type of confidence bands contain their corresponding true regression line. Briefly describe and explain the results.

(iii) Now check their respective coverage for the CEF $E[Y_i \mid X_i]$ at all X_i. Compare the results with Problem 3(ii), and explain it.

Project (20 points)

1. **(10 points)** Fit a linear model using the response and the continuous predictor variables identified previously in your dataset. Report and interpret the estimated coefficients and their associated standard errors and p-values. Report and interpret the adjusted R-squared of the model.

2. **(5 points)** Recompute the standard errors and p-values of the fitted coefficients using the sandwich formula. Notice any differences.

3. **(5 points)** Fit a linear model using the response and one of the continuous predictor variables identified previously in your dataset. Plot the four types of confidence bands as described in Problem 2.