DSC 241: Statistical Models, Assignment 7

Winter 2026

Instructions

The important updates are highlighted for your convenience.

1. **Deadline.** The submission deadline is *February 27 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 *March 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. **Code presentation.** 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)

1. **(50 points)** A flood warning system uses n ocean buoys spread out in the ocean. Buoy i records the wave height X_i in metres, and the system triggers an alert if

$$Y \;=\; \max\{X_1, \ldots, X_n\} \;>\; c,$$

where c is a detection threshold.

Note: in this question $\mathrm{Exp}(\mu)$ means an exponential distribution with mean μ.

(a) **[Math]** Under the null hypothesis that there is no surge, assume $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F$ with a cumulative density function (CDF) F. Derive the CDF of Y under the null hypothesis. Derive an expression for c that controls the false positive rate (FPR) at level α where F is either:
i. $\mathcal{N}(\mu, \sigma)$, or ii. $\mathrm{Exp}(\mu)$.
[*Hint:* Here the FPR is $\Pr[Y > c \mid H_0]$ where $\Pr[\cdot \mid H_0]$ is the probability under the null hypothesis.]

(b) **[Coding]** Suppose the designer wishes to set c to control the FPR at $\alpha = 0.01$ under a specific distribution, which may or may not be true. Use Monte Carlo simulation to estimate the empirical FPR $\hat{\alpha}$ as a function of $n \in \{2, 5, 10, 20, 50, 100\}$ in each scenario:

i. True distribution is $\mathcal{N}(\mu, \sigma)$; assumed distribution is $\mathcal{N}(\mu, \sigma)$.
ii. True distribution is $\mathcal{N}(\mu, \sigma)$; assumed distribution is $\mathrm{Exp}(\mu)$.
iii. True distribution is $\mathrm{Exp}(\mu)$; assumed distribution is $\mathcal{N}(\mu, \sigma)$.
iv. True distribution is $\mathrm{Exp}(\mu)$; assumed distribution is $\mathrm{Exp}(\mu)$.

Here let $\mu = 1.5,\ \sigma = 0.5$. Conduct a simulation of $N = 5000$ independent repetitions and create one plot for each scenario. Briefly explain your observation.

(c) **[Coding]** Suppose the designer instead does not wish to assume a parametric form for F. On a calm day with no surge, they observe n i.i.d. measurements and use the bootstrap to approximate the distribution of Y and choose c to control the FPR at $\alpha = 0.01$. Use Monte Carlo simulation to estimate the empirical FPR $\hat{\alpha}$ as a function of $n \in \{2, 5, 10, 20, 50, 100\}$ in each scenario:

i. True distribution is $\mathcal{N}(\mu, \sigma)$.
ii. True distribution is $\mathrm{Exp}(\mu)$.

Here also let $\mu = 1.5, \sigma = 0.5$. Conduct a simulation of $N = 5000$ independent repetitions and create one plot for each scenario. Describe and explain how the bootstrap thresholds compare to the theoretical thresholds from part (b). Describe and explain how well the bootstrap threshold control the FPR as n varies.

2. **(30 points)** Let X and Y be bivariate Gaussian with marginal standard Gaussian distribution and correlation ρ. Suppose n i.i.d. samples $\{(X_i, Y_i)\}_{i=1}^n$ are observed. Define the binary response

$$\tilde{Y}_i \;=\; \mathbf{1}\{Y_i \geq c\}, \qquad c \in \mathbb{R}.$$

where c is the threshold.

(a) **[Math]** Derive the conditional distributions for $Y \mid X$ and $\tilde{Y} \mid X$.

(b) **[Coding]** Let $\rho = 0.5,\ c = 0.5$ and simulate a sample of size $n = 1000$. Create a single scatter plot for both (x_i, y_i) and $(x_i, \tilde{y}_i)$, and overlay both CEFs $\mathrm{E}[Y \mid X]$ and $\mathrm{E}[\tilde{Y} \mid X]$. Describe your observation.

(c) **[Math]** Suppose only $\tilde{Y}$ is observed. If you fit a generalised linear model (GLM) for $\tilde{Y}$, what link function is most appropriate? Briefly justify.

(d) **[Coding]** Suppose only $\tilde{Y}$ is observed, but you fit a linear model (with an intercept)

$$\tilde{Y} = \beta_0 + \beta_1 X + e$$

by OLS anyway. Here also let $\rho = 0.5,\ c = 0.5$. Use Monte Carlo simulation of $N = 100$ independent repetitions to estimate $\mathrm{E}[\hat{\beta}_0]$ and $\mathrm{E}[\hat{\beta}_1]$ as a function of $n \in \{2, 5, 10, 20, 50, 100\}$. Compare them to the coefficients from the original model.

(e) **[Math]** Follwing part (d), derive $\mathrm{E}[\hat{\beta}_0]$ and $\mathrm{E}[\hat{\beta}_1]$ in terms of (ρ, c). Argue whether they are biased.

Project (20 points)

1. **(6 points)** Consider the features and response that you identified or created via transformations in previous assignments. Binarise the continuous outcome by setting an appropriate threshold. Fit a new model with the same predictors to the binarised outcome using logistic regression. Include standard errors and an interpretation of the coefficients.

2. **(8 points)** For the logistic regression in Project Question 1, use an appropriate method to perform variable selection on the predictors. Report the reduced model and the standard errors of its coefficients using the bootstrap. Make sure to include the model selection step into the bootstrap procedure.

3. **(6 points)** Compare the selected model using the binary outcome with the selected model using the continuous outcome obtained in Assignment 6. Report the differences between the two models in terms of model interpretation.