DSC 241: Statistical Models, Assignment 6

Winter 2026

Instructions

The important updates are highlighted for your convenience.

1. **Deadline.** The submission deadline is *February 20 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 22 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)

The following setting applies to Problems 1 and 2.

Let (X_1, X_2) be jointly Gaussian with

$$\mathrm{E}[X_1] = \mathrm{E}[X_2] = 0, \qquad \mathrm{Var}[X_1] = \mathrm{Var}[X_2] = 1, \qquad \mathrm{Corr}[X_1, X_2] = \rho \geq 0.$$

Then, let

$$Y \;=\; \beta_1 X_1 + \beta_2 X_2 + e, \qquad e \perp (X_1, X_2), \qquad e \sim \mathcal{N}(0, \sigma^2).$$

Suppose we observe n i.i.d. samples $\{(Y_i, X_{1,i}, X_{2,i})\}_{i=1}^n$ from the above settings.
We also have the following two model candidates (whether to neglect X_2):

- **Model S (partial):** $Y = \beta_1 X_1 + e$,
- **Model 3 (full):** $Y = \beta_1 X_1 + \beta_2 X_2 + e$.

1. **(50 points) [Coding]** The goal of this problem is to show that model selection may bias inference.

   We wish to perform a model selection and obtain inference for β_1 using the following algorithm:

Algorithm 1 Inference for β_1 with model selection

Require: Data $\{(Y_i, X_{1,i}, X_{2,i})\}_{i=1}^n$, significance level $\alpha \in (0, 1)$

Fit Model S: regress Y on X_1; record $\hat{\beta}_1^{\mathsf{S}}$ and $(1-\alpha)$ confidence interval $\mathrm{CI}_1^{\mathsf{S}}$.
Fit Model 3: regress Y on (X_1, X_2); record $\hat{\beta}_1^{\mathsf{3}}$ and $(1-\alpha)$ confidence interval $\mathrm{CI}_1^{\mathsf{3}}$ ($\mathsf{3}$ is not the cube here).
In Model 3, perform the two-sided t-test on $\beta_2 = 0$; obtain the p-value p.
if $p < \alpha$ **then**
    X_2 cannot be neglected, select Model 3: $\hat{\beta}_1 \leftarrow \hat{\beta}_1^{\mathsf{3}}$, $\mathrm{CI}_1 \leftarrow \mathrm{CI}_1^{\mathsf{3}}$ ($\mathsf{3}$ is also not the cube here).
else
    X_2 can be neglected, select Model S: $\hat{\beta}_1 \leftarrow \hat{\beta}_1^{\mathsf{S}}$, $\mathrm{CI}_1 \leftarrow \mathrm{CI}_1^{\mathsf{S}}$.
end if
return $\hat{\beta}_1$, CI_1

Consider a Monte Carlo simulation. Fix

$$\beta_1 = 1, \qquad \sigma^2 = 0.25, \qquad \alpha = 0.05$$

and consider

$$\rho \in \{0,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 0.99\}, \qquad \beta_2 \in \{0,\ 0.1,\ 0.2,\ \ldots,\ 1.2\}.$$

For each pair (ρ, β_2), conduct an experiment with $N = 20000$ independent replications with sample size $n = 60$. In each replication, generate a dataset, apply the selection algorithm, and record the resulting point estimate $\hat{\beta}_1$ and $(1-\alpha)$ confidence interval CI_1. Answer the following questions based on your simulation results.

(a) For each (ρ, β_2), estimate the bias of $\hat{\beta}_1$

$$\widehat{\mathrm{Bias}}[\hat{\beta}_1](\rho, \beta_2) = \frac{1}{N} \sum_{r=1}^{N} \left(\hat{\beta}_1^{(r)}(\rho, \beta_2) - \beta_1 \right).$$

Here $\hat{\beta}_1^{(r)}(\rho, \beta_2)$ means $\hat{\beta}_1$ in the r-th simulation under the given setting of ρ and β_2.
Plot $\widehat{\mathrm{Bias}}[\hat{\beta}_1](\rho, \beta_2)$ versus β_2 for each ρ (multiple labelled curves on a single plot) with a pointwise 95% Monte Carlo confidence interval. Assess under which situations the bias is statistically distinguishable from 0.

[*Hint:* The Monte Carlo standard error for $\widehat{\mathrm{Bias}}[\hat{\beta}_1]$ is $\widehat{\mathrm{se}}_{\mathrm{MC}}[\widehat{\mathrm{Bias}}[\hat{\beta}_1]] = \sqrt{\frac{\widehat{\mathrm{Var}}[\hat{\beta}_1]}{N}}$.]

(b) For each (ρ, β_2), estimate the empirical coverage probability of CI_1 on the true value β_1

$$\widehat{\mathrm{CP}}_1(\rho, \beta_2) = \frac{1}{N} \sum_{r=1}^{N} \mathbf{1}\left\{\beta_1 \in \mathrm{CI}_1^{(r)}(\rho, \beta_2)\right\}.$$

Here $\mathrm{CI}_1^{(r)}(\rho, \beta_2)$ means CI_1 in the r-th simulation under the given setting of ρ and β_2.

Plot $\widehat{\mathrm{CP}}_1(\rho, \beta_2)$ versus β_2 for each ρ (multiple labelled curves on a single plot) with a pointwise 95% Monte Carlo confidence interval. Interpret the pattern of coverage in light of the bias behavior in part (a).

[*Hint:* The Monte Carlo standard error for $\widehat{\mathrm{CP}}_1$ is $\widehat{\mathrm{se}}_{\mathrm{MC}}[\widehat{\mathrm{CP}}_1] = \sqrt{\frac{\widehat{\mathrm{CP}}_1(1-\widehat{\mathrm{CP}}_1)}{N}}$.]

(c) Further fix $\rho = 0.8$ and select $\beta_2 \in \{0.0, 0.4, 0.8\}$. For each β_2, plot the empirical distribution of $\hat{\beta}_1$. Argue that the distribution of $\hat{\beta}_1$ is a mixture of two components and explain what determines the weight of each component in the mixture.

(d) It may seem counterintuitive that the decision to include or exclude X_2 in the model should produce bias in the estimation of the coefficient of X_1. Use the mixture argument from part (c) to explain the bias of $\hat{\beta}_1$ and how it changes as a function of β_2.

2. **(30 points) [Coding]**

This problem introduces the mean squared prediction error (MSPE) and its estimation.

Let D be a training set and let $\hat{f}_D$ be a predictor fitted on D. The (conditional) MSPE under a test distribution $\mathcal{T}$ is

$$R(\hat{f}_D) \triangleq \mathrm{E}_{(\boldsymbol{X},Y)\sim\mathcal{T}}\left[\left(Y - \hat{f}_D(\boldsymbol{X})\right)^2 \middle| D\right].$$

Given an i.i.d. test set $T = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{|T|}$ from $\mathcal{T}$, its empirical estimator is

$$\hat{R}(\hat{f}_D) \triangleq \frac{1}{|T|} \sum_{i=1}^{|T|} \left(y_i - \hat{f}_D(\boldsymbol{x}_i)\right)^2$$

To finish the question, consider Monte Carlo simulations of $N = 20000$ independent replications with sample size $n = 60$. Fix

$$\beta_1 = 1, \qquad \beta_2 = 0.25, \qquad \sigma^2 = 0.25.$$

(a) Conduct a simulation for each $\rho \in \{0, 0.2, 0.4, 0.6, 0.8, 0.99\}$. In each replication, fit **Model S** on a training set with a holdout ratio of $r = 0.2$, and compute the empirical estimator of MSPE $\hat{R}(\hat{f}_D)$ on the test set. Plot the Monte Carlo mean of $\hat{R}(\hat{f}_D)$ with its pointwise 95% Monte Carlo confidence interval as a function of ρ. Overlay the theoretical value and briefly interpret the trend.

[*Hint:* The theoretical MSPE for **Model S** is $R(\hat{f}_D) = \left(1 + \frac{1}{|D|-2}\right)\left(\sigma^2 + \beta_2^2\left(1-\rho^2\right)\right)$.]

(b) Based on the setting in part (a), fix $\rho = 0.5$ but let r vary. Conduction a simulation for each $r \in \{0.1, 0.2, \ldots, 0.9\}$ to obtain $\hat{R}(\hat{f}_D)$. Plot the Monte Carlo mean of $\hat{R}(\hat{f}_D)$ with its pointwise 95% Monte Carlo confidence interval as a function of r, and overlay the theoretical value. Based on the plot, give a practical rationale for choosing r.

(c) Based on the setting in part (b), replace holdout by K-fold cross-validation. Conduct a simulation for each $K \in \{2, 3, 6, 10, 20, 30\}$ to obtain $\hat{R}(\hat{f}_D)$. Plot the Monte Carlo mean of $\hat{R}(\hat{f}_D)$ with its pointwise 95% Monte Carlo confidence interval as a function of K, and overlay the theoretical value. Based on the plot, give a practical rationale for choosing K.

Project (20 points)

Choose one of the multiple regression fits you performed before on your data with several predictors.

1. **(2 points)** Report the standard errors of the model coefficients as automatically provided by the linear model fit under the standard assumptions.
2. **(8 points)** Estimate the standard errors of the coefficients by bootstrapping. Decide and explain whether it is appropriate to resample cases or residuals. Compare to the classical estimates from part (a).
3. **(10 points)** 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.