DSC 241: Statistical Models, Assignment 5

Winter 2026

Instructions

The important updates are highlighted for your convenience.

1. **Deadline.** The submission deadline is *February 13 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 15 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. **(40 points)** **[Math]** A standard method for comparing two groups (often called A/B testing) is the two-sample t-test. In this problem you will show that the pooled-variance two-sample t-test is equivalent to the usual OLS t-test from a linear regression with a binary group indicator. This is interesting because it shows that the t-test is not a separate statistical procedure but just a special case of linear regression.

   Let $X_i \in \{0, 1\}$ indicate the group membership (0 means Control group; 1 means Treatment group), and let Y_i be the outcome for observation $i = 1, \ldots, n$. Let $n_0 = \sum_{i=1}^{n} \mathbf{1}\{X_i = 0\}$ and $n_1 = \sum_{i=1}^{n} \mathbf{1}\{X_i = 1\}$.

Denote by $\bar{Y}^{(0)}$ and $\bar{Y}^{(1)}$ the sample means in the Control and Treatment groups, and by $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$ the corresponding within-group sample variances, which are defined as

$$\hat{\sigma}_g^2 \triangleq \frac{1}{n_g - 1} \sum_{i:X_i=g} \left(Y_i - \bar{Y}^{(g)}\right)^2, \; g \in \{0, 1\}. \tag{1}$$

The pooled-variance two-sample t-statistic is defined as

$$t \triangleq \frac{\bar{Y}^{(1)} - \bar{Y}^{(0)}}{s_p \sqrt{\frac{1}{n_0} + \frac{1}{n_1}}}, \tag{2}$$

where the pooled standard deviation is defined as

$$s_p \triangleq \sqrt{\frac{(n_0 - 1)\hat{\sigma}_0^2 + (n_1 - 1)\hat{\sigma}_1^2}{n - 2}}. \tag{3}$$

Now consider the linear regression model

$$Y = \beta_0 + \beta_1 X + e.$$

(a) Show that the OLS slope estimate equals the difference between the sample mean of the two groups.

(b) Show that the OLS intercept estimate equals the sample mean of the Control group.

(c) Show that the homoscedastic OLS estimator of the error standard deviation equals the pooled standard deviation.
[*Hint:* Recall that the homoscedastic OLS estimator of the error variance is $\hat{\sigma}^2 \triangleq \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$.]

(d) Show that the t-statistic for the slope estimate equals the two-sample t-statistic.
[*Hint:* Recall that the t-statistic for the null hypothesis $H_0 : \hat{\beta}_1 = 0$ is $t_0 \triangleq \frac{\hat{\beta}_1}{\hat{\text{se}}(\hat{\beta}_1)}$.]

(e) Briefly discuss what distributional assumptions are needed for the equalities in parts (a) – (d) to hold?

2. **(40 points) [Coding]** In this problem you will practice working with discrete predictors on the Boston Housing Dataset. To access this dataset, you may either

- call `Boston` in `MASS` if you use `R`, or
- run in shell
  `curl -o boston.txt https://lib.stat.cmu.edu/datasets/boston`
  to download the original dataset as `boston.txt`, and run the following code to load it into `df` if you use `Python` (you may install `pandas` and `numpy` first):

```
import pandas as pd
import numpy as np
raw = pd.read_csv("boston.txt", sep=r"\s+", skiprows=22, header=None)
X = np.hstack([raw.values[::2, :], raw.values[1::2, :2]])
y = raw.values[1::2, 2]
columns = [ "crim", "zn", "indus", "chas", "nox", "rm", "age",
"dis", "rad", "tax", "ptratio", "b", "lstat"]
df = pd.DataFrame(X, columns=columns)
df["medv"] = y
```

Take the median property value `medv` as the response variable, and use the percentage of lower-status population `lstat`, the indicator for bordering the Charles River `chas`, and the index of accessibility to radial highways `rad` as predictors.

(a) Model `medv` as a function of `lstat` and `chas`. Use pair plots to assess the relationship and suggest transformations that may reduce nonlinearity. Fit a linear model and interpret the estimated coefficients.

(b) Extend the model in part (a) by including interaction terms. Assess whether interactions improve model fit and interpret the resulting coefficients.

(c) Fit two separate models of `medv` as a function of `lstat`, one for observations with `chas` $= 0$ and one for `chas` $= 1$. Compare these to the joint interaction model from part (b) using appropriate quantitative criteria, and determine which approach is preferred.

(d) Restrict the sample to observations with `rad` ≤ 8. Model `medv` as a function of `lstat` and `rad`. Consider `rad` both as a numerical and as a categorical variable, fit both models, and use appropriate quantitative criteria to decide which specification is more appropriate.

(e) Using the categorical specification of `rad` from part (d), estimate the difference in the expected value of `medv` between houses with `rad` $= 3$ and `rad` $= 4$, and compute the corresponding estimated standard error.

Project (20 points)

1. **(10 points)** Fit a linear model using the response and a combination of continuous and discrete predictor variables. If you have no categorical predictors, you may split a numerical variable into a categories by thresholding. Report and interpret the estimated coefficients and their associated standard errors and p-values. Use graphics to evaluate assumptions and offer brief comments on what you observe.

2. **(10 points)** Split your data according to a categorical variable and fit separate linear models to each of the data subsets including at least one numerical variable of interest.

   (a) Compare the slope coefficient of that numerical variable across the models. Do the results suggest the inclusion of interaction terms? Justify your answer numerically and graphically.

   (b) Fit a single model with interactions. Compare your results to the models fitted separately.

   (c) Give a specific interpretation of the interaction terms.