

Due: Sunday, Dec 15, 11:59PM

This homework comprises five coding exercises from the ISLP book. Please start early!

Guideline for those new to data analysis using Python:

Review the Lab section within each chapter prior to tackling the programming tasks.

(e.g., p. 489 of Chapter 11 or <https://islp.readthedocs.io/en/latest/labs/Ch11-surv-lab.html>)

Find datasets and Jupyter notebooks at https://github.com/intro-stat-learning/ISLP_labs/.

Deliverables: Submit a PDF of your homework, with an appendix listing all your code, to the Gradescope assignment entitled “HW4 Write-Up”. You may typeset your homework in LaTeX or Word or submit neatly handwritten and scanned solutions. Make sure each solution is on a new page, and graphs are included in the correct sections. We need each solution to be self-contained on pages of its own.

Guideline:

1. On the first page of your write-up, please sign next to the integrity statement. We want to make extra clear the consequences of cheating.
2. On the first page of your write-up, please list students who helped you or whom you helped on the homework. (Note that sending each other code is not allowed.)
3. Please write your answers in English. Korean is not allowed. Non-Korean staff members may not be able to grade responses in Korean.
4. Please don't forget to select all the pages that are related to each question during the Gradescope submission! (Submissions that do not clearly reference the exact pages containing the solution may not be graded.)

For staff use only

Honor Code	Q1	Q2	Q3	Q4	Q5	Total
/ 4	/ 24	/ 18	/ 18	/ 16	/ 20	/ 100

Honor Code [4 pts]

Declare and sign the following statement:

"I certify that all solutions in this document are entirely my own and that I have not looked at anyone else's solution. I have given credit to all external sources I consulted."

Signature:

We welcome group discussions, but the work you submit should be entirely your own. If you use any information or pictures not from our lectures or readings, make sure to say where they came from. Please note that breaking academic rules can lead to severe penalties.

- (a) Did you receive any help whatsoever from anyone in solving this assignment? If your answer is 'yes', give full details (e.g. "Junho explained to me what is asked in Q2-a")

- (b) Did you give any help whatsoever to anyone in solving this assignment? If your answer is 'yes', give full details (e.g. "I pointed Josh to Ch. 2.3 since he didn't know how to proceed with Q2")

- (c) Did you find or come across code that implements any part of this assignment? If your answer is 'yes', give full details (book & page, URL & location within the page, etc.).

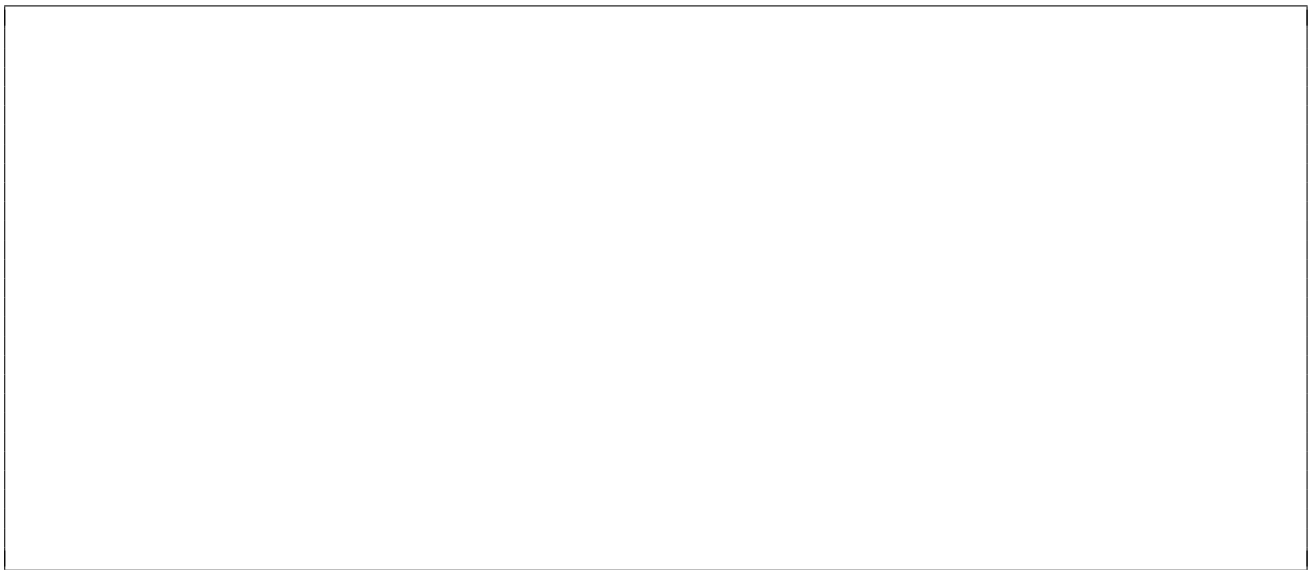
Q1. Multiple Testing 📊 [24 pts]

In this problem, we will simulate data from $m = 100$ fund managers.

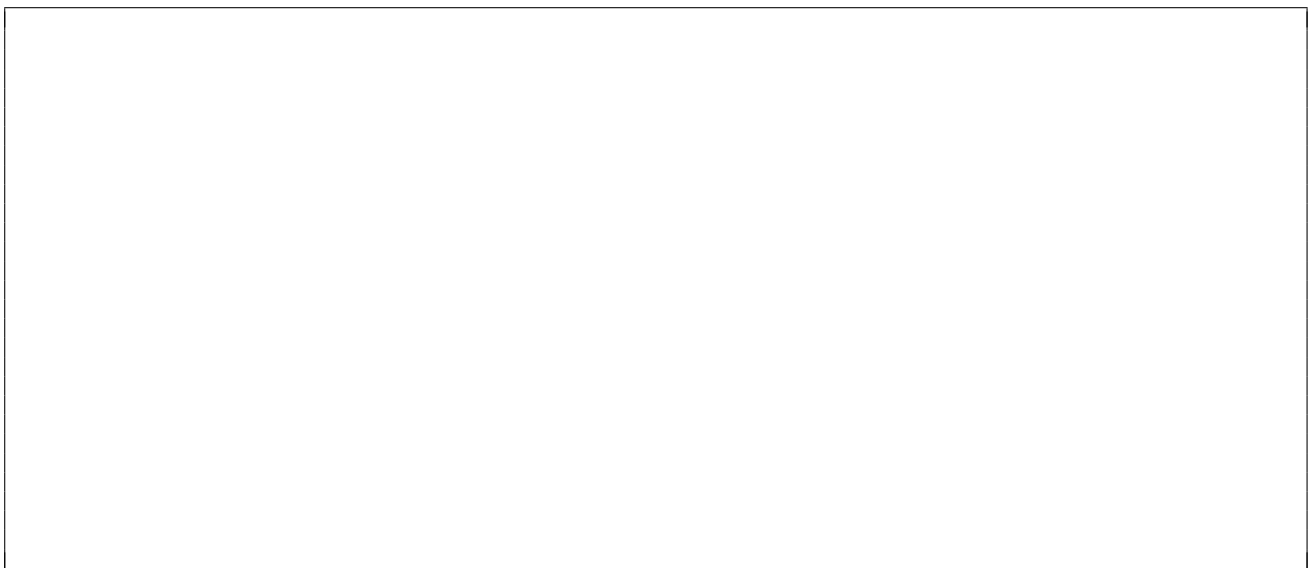
```
rng = np.random.default_rng(1)
n, m = 20, 100
X = rng.normal(size=(n, m))
```

These data represent each fund manager's percentage returns for each of $n = 20$ months. We wish to test the null hypothesis that each fund manager's percentage returns have population mean equal to zero. Notice that we simulated the data in such a way that each fund manager's percentage returns do have population mean zero; in other words, all m null hypotheses are true.

- (a) Conduct a one-sample t -test for each fund manager, and plot a histogram of the p -values obtained. [4 pts]



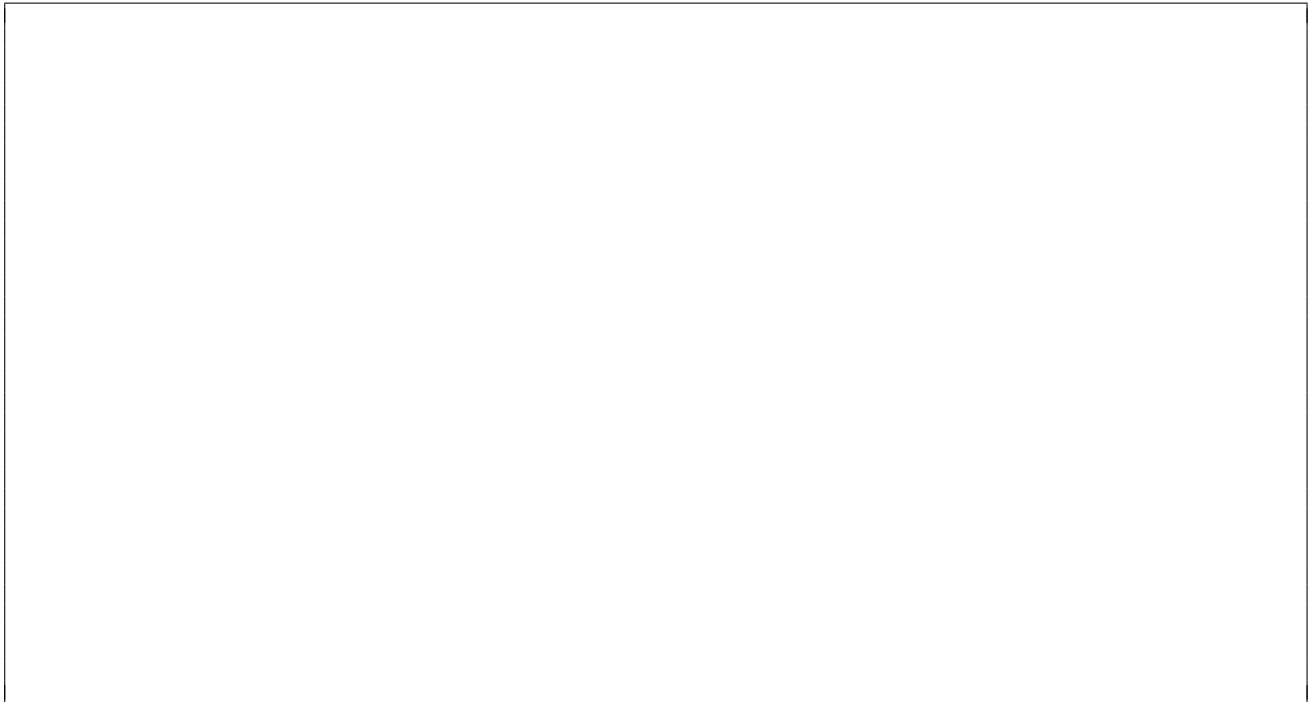
- (b) If we control Type I error for each null hypothesis at level $\alpha = 0.05$, then how many null hypotheses do we reject? [4 pts]



(c) If we control the FWER at level 0.05, then how many null hypotheses do we reject? [4 pts]

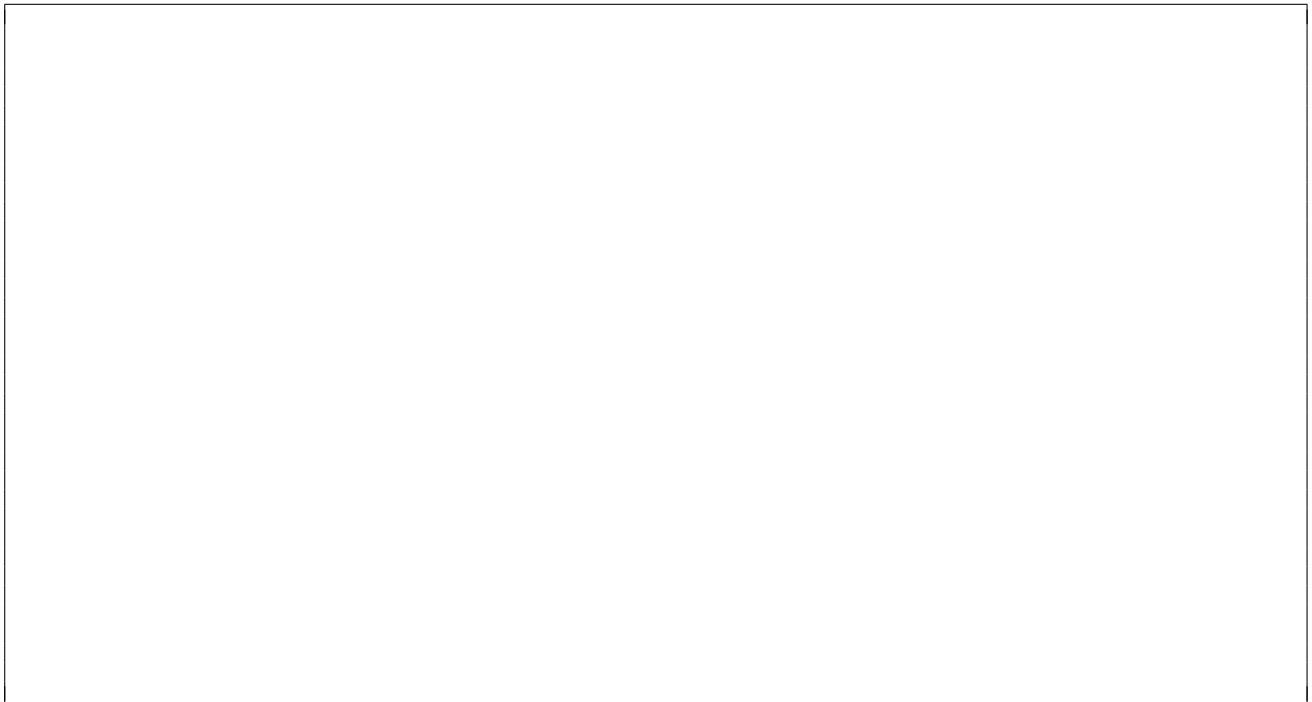
(d) If we control the FDR at level 0.05, then how many null hypotheses do we reject? [4 pts]

- (e) Now suppose we “cherry-pick” the 10 fund managers who perform the best in our data. If we control the FWER for just these 10 fund managers at level 0.05, then how many null hypotheses do we reject? If we control the FDR for just these 10 fund managers at level 0.05, then how many null hypotheses do we reject? [4 pts]



- (f) Explain why the analysis in (e) is misleading. [4 pts]

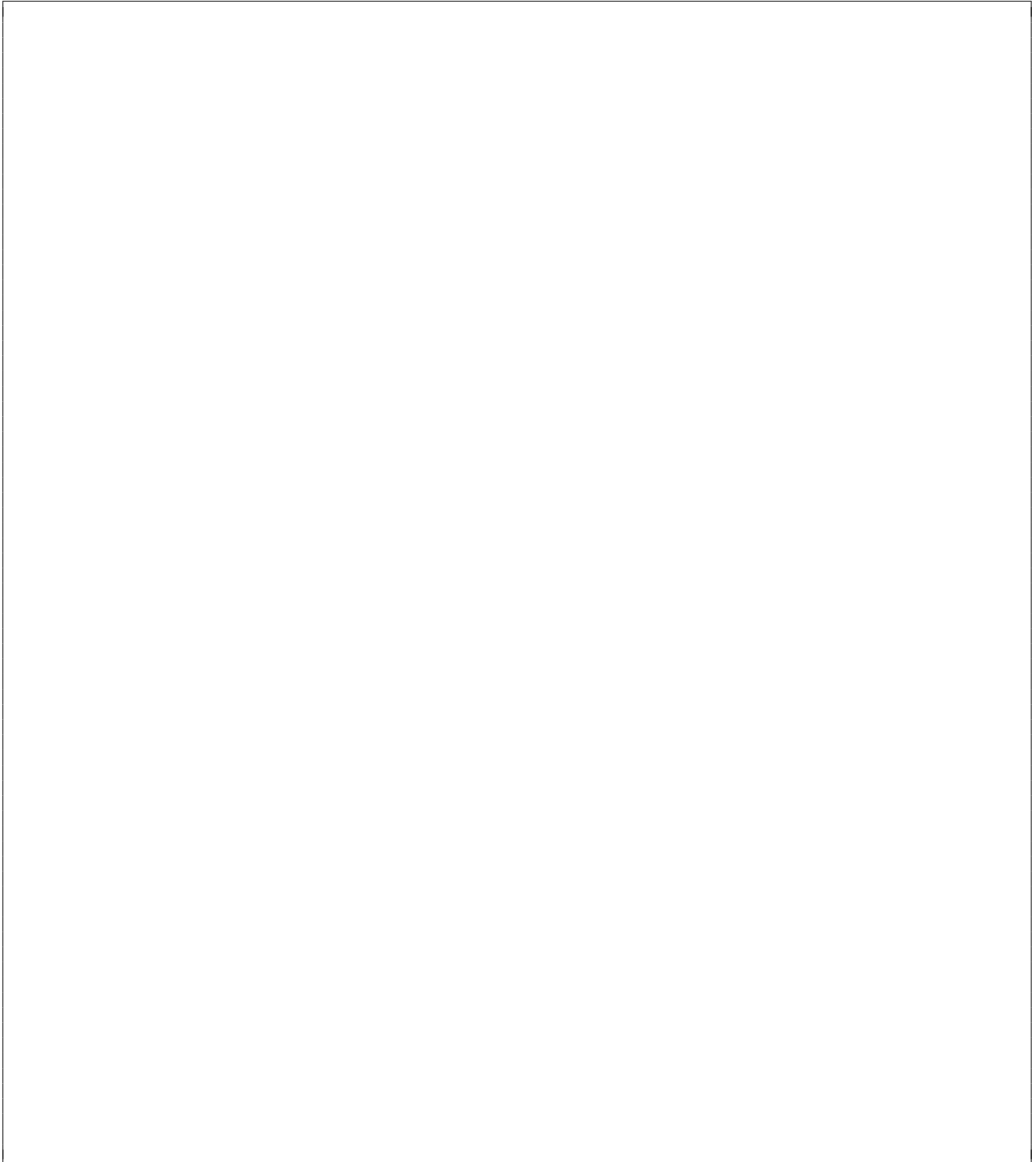
Hint: The standard approaches for controlling the FWER and FDR assume that all tested null hypotheses are adjusted for multiplicity, and that no “cherry-picking” of the smallest p -values has occurred. What goes wrong if we cherry-pick?



Q2. Support Vector Machine [18 pts]

At the end of Section 9.6.1, it is claimed that in the case of data that is just barely linearly separable, a support vector classifier with a small value of C that misclassifies a couple of training observations may perform better on test data than one with a huge value of C that does not misclassify any training observations. You will now investigate this claim.

- (a) Generate two-class data with $p = 2$ in such a way that the classes are just barely linearly separable. [6 pts]



- (b) Compute the cross-validation error rates for support vector classifiers with a range of C values. How many training observations are misclassified for each value of C considered, and how does this relate to the cross-validation errors obtained? [6 pts]

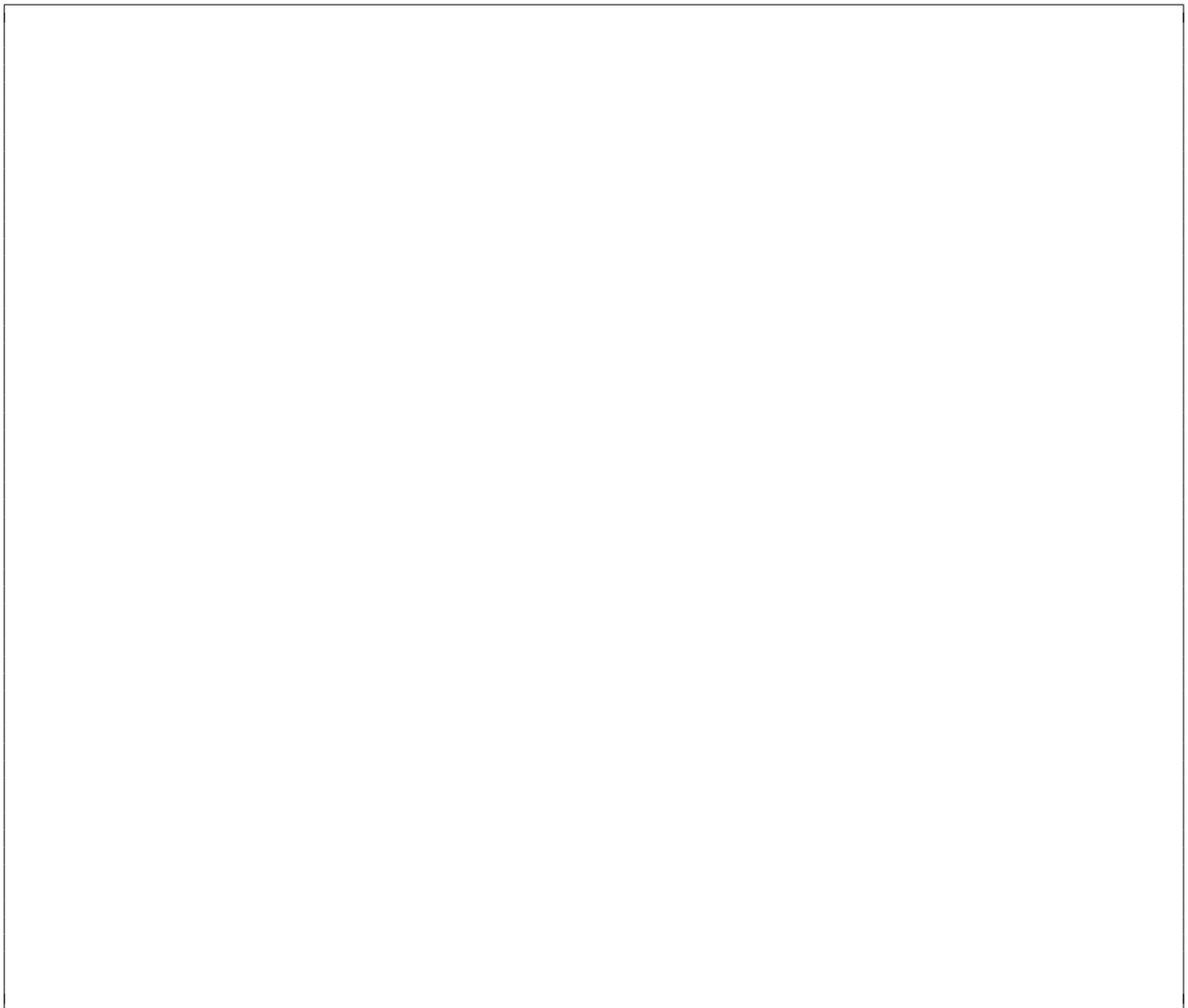
- (c) Generate an appropriate test data set, and compute the test errors corresponding to each of the values of C considered. Which value of C leads to the fewest test errors, and how does this compare to the values of C that yield the fewest training errors and the fewest cross-validation errors? [6 pts]

Q3. Survival Analysis 📊 [18 pts]

This exercise makes use of the data in the below table.

Observation (Y)	Censoring Indicator (δ)	Covariate (X)
26.5	1	0.1
37.2	1	11
57.3	1	-0.3
90.8	0	2.8
20.2	0	1.8
89.8	0	0.4

- (a) Create two groups of observations. In Group 1, $X < 2$, whereas in Group 2, $X > 2$. Plot the Kaplan-Meier survival curves corresponding to the two groups. Be sure to label the curves so that it is clear which curve corresponds to which group. By eye, does there appear to be difference between the two groups' survival curves? [6 pts]



- (b) Fit Cox's proportional hazards model, using the group indicator as a covariate. What is the estimated coefficient? Write a sentence providing the interpretation of this coefficient, in terms of the hazard or the instantaneous probability of the event. Is there evidence that the true coefficient value is non-zero? [6 pts]

- (c) Recall that in the case of a single binary covariate, the log-rank test statistic should be identical to the score statistic for the Cox model. Conduct a log-rank test to determine whether there is a difference between the survival curves for the two groups. How does the p -value for the log-rank test statistic compare to the p -value for the score statistic for the Cox model from (b)? [6 pts]

Q4. Hierarchical Clustering 📖 [16 pts]

On the book website, there is a gene expression data set (<https://www.statlearning.com/s/Ch12Ex13.csv>) that consists of 40 tissue samples with measurements on 1,000 genes. The first 20 samples are from healthy patients, while the second 20 are from a diseased group.

- (a) Load in the data using `pd.read_csv()` with `header = None`. Then, apply hierarchical clustering to the samples using correlation-based distance, and plot the dendrogram. Do the genes separate the samples into the two groups? Do your results depend on the type of linkage used? [9 pts]

(b) Your collaborator wants to know which genes differ the most across the two groups. Suggest a way to answer this question, and apply it here. [7 pts]

Q5. RNN 📖 [20 pts]

In ISLP Chapter 10.9.6 (page 458), we showed how to fit a linear AR model to the NYSE data using the `LinearRegression()` function. However, we also mentioned that we can “flatten” the short sequences produced for the RNN model in order to fit a linear AR model. Use this latter approach to fit a linear AR model to the NYSE data.

- (a) Compare the test R^2 of this linear AR model to that of the linear AR model that we fit in the lab (ISLP Chapter 10.9.6). What are the advantages/disadvantages of each approach? [8 pts]

- (b) Repeat the previous exercise, but now fit a nonlinear AR model by “flattening” the short sequences produced for the RNN model. [6 pts]

(c) Modify the code with a 12-level factor representing the month, and the variable `day_of_week`. Do these factors improve the performance of the model? Compute the test R^2 . [6 pts]