DSC 241: Statistical Models, Assignment 4

Winter 2026

Instructions

The important updates are highlighted for your convenience.

1. **Deadline.** The submission deadline is *February 6 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 8 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]** The goal of this question is to better understand how multicollinearity affects ordinary least squares (OLS) estimation.

   Consider the linear regression setting with response vector $\boldsymbol{y} \in \mathbb{R}^n$ and design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$.

   (i) Suppose the columns of $\boldsymbol{X}$ are linearly dependent. Prove that the OLS coefficient vector $\hat{\boldsymbol{\beta}}$ is not uniquely defined. Find the form of the solution set.

(ii) Suppose the columns of $\boldsymbol{X}$ are linearly independent, and the linear model

$$y = \boldsymbol{x}^\top \boldsymbol{\beta} + e, \qquad \boldsymbol{x}, \boldsymbol{\beta} \in \mathbb{R}^p, \ e \sim \mathcal{N}(0, \sigma^2),$$

holds. Let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ be the singular value decomposition (SVD) of $\boldsymbol{X}$. Compute the maximum variance over all unit-norm linear combinations of the fitted coefficients:

$$\max_{\|\boldsymbol{a}\|_2 = 1} \ \mathrm{Var}[\boldsymbol{a}^\top \hat{\boldsymbol{\beta}}].$$

[*Hint:* Recall that the variance of the fitted coefficients in this simplified case is $\mathrm{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$.]

(iii) Suppose the first $p-1$ columns of $\boldsymbol{X}$ are linearly independent, but the final column is nearly a linear combination of the first $p-1$ columns:

$$\boldsymbol{x}_p = \sum_{i=1}^{p-1} c_i \boldsymbol{x}_i + \boldsymbol{z},$$

where the coefficients $\{c_i\}$ are not all zero and $\boldsymbol{z}$ is a perturbation term. Show that the maximum variance from part (ii) grows without bound as $\|\boldsymbol{z}\|_2 \to 0$. Provide an interpretation of this phenomenon in terms of multicollinearity.

2. **(40 points)** [**Coding**] The Capital Asset Pricing Model (CAPM) is a linear regression model often used in finance. In this question we will compare the historical daily closing prices of NVIDIA (ticker symbol: `NVDA`) and compare it to the S&P 500 Index (ticker symbol: `SPY`) as a market benchmark.

To finish this question, you need to download the `NVDA` and `SPY` adjusted daily closing prices from Yahoo Finance for the last five years.

- In `R` this can be done by the following code (you may install `quantmod` first):

```
library(quantmod)
getSymbols(c("NVDA", "SPY"), src = "yahoo", from = Sys.Date() - 1825, to = Sys.Date())
NVDA <- NVDA$NVDA.Adjusted
SPY <- SPY$SPY.Adjusted
```

- In `Python` this can be done by the following code (you may install `yfinance` first):

```
from datetime import datetime, timedelta
import yfinance as yf
data = yf.download(["NVDA", "SPY"], start = datetime.today() - timedelta(days = 1825),
end = datetime.today(), auto_adjust = True)["Close"]
NVDA = data["NVDA"]
SPY = data["SPY"]
```

(i) Plot the adjusted stock prices

(a) as a function of time for both `NVDA` and `SPY`, and

(b) as a scatterplot of `NVDA` versus `SPY` on the same trading day.

Comment on any relationship between the two stocks. Fit a linear regression of `NVDA` on `SPY` and use regression diagnostics discussed in class to assess nonlinearities.

(ii) The daily log return on day t is defined as

$$r_t = \log\left(\frac{P_t}{P_{t-1}}\right),$$

where P_t denotes the price on day t. Compute daily log returns

(a) as a function of time for both `NVDA` and `SPY`, and

(b) as a scatterplot of NVDA versus SPY on the same trading day.

Comment on any relationship between the two stocks. Fit a CAPM regression

$$r_t^{\texttt{NVDA}} = \alpha + \beta\, r_t^{\texttt{SPY}} + e_t,$$

and use regression diagnostics discussed in class to assess nonlinearities.

(iii) Using the log-return regression in part (ii), identify the three most influential observations and report their trading dates. Quantify influence by reporting how much the fitted coefficients change. Briefly investigate whether each date corresponds to a noteworthy news or market event (cite your sources).

(iv) Interpret the fitted coefficients from the log-return model in part (ii) in terms of how NVDA changes relative to SPY on the original price scale. Provide a clear mathematical justification for your interpretation.

(v) Fit the log-return CAPM regression in a moving 90-trading-day window. Plot the estimated coefficients as functions of time. Comment on how NVDA's relationship to SPY appears to have changed over the last 5 years.

Project (20 points)

1. (5 points) Consider a linear regression of a continuous response as a function of continuous predictors in your data. Use diagnostic methods discussed in class to assess whether there may be nonlinearities. Justify your assessment both numerically and using appropriate graphics. Offer brief comments on what you observe.

2. (10 points) Fix the diagnostic issues identified above. Consider univariate and multivariate variable transformations, polynomial regression, PCA, and other feature engineering methods if desired. Re-evaluate the presence of nonlinearities using diagnostic tools and appropriate graphics.

3. (5 points) Compare the old model and the transformed model in terms of R-squared, variance of the coefficients and interpretation of the coefficients.