WiDS Kalman Filtered Trend Trader
Assignment 1

Aarav Malde and Krishang Krishna

December 2025

Question 1: Linear Regression

Consider a dataset of n observations with response variable y_i and p predictors $x_{i1}, x_{i2}, \ldots, x_{ip}$. Let

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top, \qquad \mathbf{X} = [\mathbf{1}\ \ \mathbf{x}_1\ \ \cdots\ \ \mathbf{x}_p],$$

denote the vector of regression coefficients and the $n \times (p+1)$ design matrix, respectively. Use this notation in all your answers.

1. Write the multiple linear regression model with p predictors. Define all variables, parameters, and assumptions.

2. State what ordinary least squares (OLS) minimizes and write the full mean-squared error (MSE) objective function.

3. Starting from the MSE objective, derive: $\hat{\beta} = (X^T X)^{-1} X^T y$.

4. State the conditions under which $X^T X$ is invertible and explain why multicollinearity makes it singular.

5. Show that $\hat{y}$ is the orthogonal projection of y onto the column space of X. Prove that: $X^T(y - \hat{y}) = 0$.

6. Given $J(\beta) = \frac{1}{2n}\|X\beta - y\|^2$, derive $\nabla_\beta J(\beta)$ and the batch gradient descent update rule.

Use the dataset **linear_regression_dataset.csv** for the upcoming questions.

Download the dataset from the following link: linear_regression_dataset.csv It contains 300 samples, 12 predictor variables $(x_1, \ldots, x_{12})$, and a continuous response variable y.

7. Using NumPy, compute:$\hat{\beta} = (X^T X)^{-1} X^T y$. Compare your result with `sklearn.linear_model.LinearRe`

8. Plot residuals vs fitted values. Comment on homoscedasticity.

9. Plot a Q–Q plot of residuals. Comment on normality.

10. Explain what violations might indicate about model assumptions.

11. Compute: $H = X(X^TX)^{-1}X^T$ Identify high-leverage and influential points using leverage values and Cook's distance. Comment on their impact.

12. Derive the decomposition: $\mathbb{E}\left[(y - \hat{f}(x))^2\right] = \text{Bias}^2 + \text{Var} + \sigma^2$.

13. Suppose: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, but you fit: $y = \alpha_0 + \alpha_1 x_1 + u$.

a) Derive $\mathbb{E}[\alpha_1]$ in terms of β_1, β_2, and $\text{Cov}(x_1, x_2)$.

b) Show how omitting x_2 biases α_1.

c) State conditions under which the bias disappears.

14. Generate data where:

$$x_2 = x_1 + 0.9z, \qquad z \sim \mathcal{N}(0, 1).$$

a) Compute the condition number of X^TX.

b) Show how variance of $\hat{\beta}$ increases with correlation.

c) Explain why multicollinearity causes instability but not bias.

Question 2: Salary Prediction & Bias Detection

A company wants to build an ML model to predict employee salaries based on demographic and skill-related features. However, they are concerned that the model may unintentionally introduce or amplify social biases.

Your task is to build a predictive model, analyze biases, and propose mitigation strategies, following responsible AI principles.

You are provided a dataset containing 12,000 employee records with the following fields:

Feature	Type	Description
age	numeric	Age of employee
gender	categorical	Male, Female, Other
education_level	categorical	HighSchool, Bachelors, Masters, PhD
years_experience	numeric	Total years of industry experience
job_title	categorical	One of 15 job roles

performance_score	numeric	Annual rating 1–5
industry	categorical	Tech, Finance, Healthcare, Retail
city	categorical	12 possible cities
previous_companies	numeric	Number of companies worked at
remote_worker	categorical	Yes / No
salary	numeric	Annual salary in USD (target variable)

Download the dataset from the following link: salary_dataset.csv

1. Perform EDA: univariate statistics & histograms, correlation heatmaps, distribution of salary across gender, education, and industry.
2. Handle missing values and outliers; justify your decisions.
3. Encode categorical features (explain why you chose the method you did).
4. Split the dataset using a stratified sampling strategy. Specify the variable used for stratification and justify it.
5. Train a baseline OLS Linear Regression model.
6. Report coefficients, standard errors, p-values, confidence intervals.
7. Interpret five of the most influential coefficients.
8. Evaluate using RMSE, MAE, and R^2.
9. Check whether the linearity assumptions hold (plots + explanation).
10. Consider gender as the protected attribute. Compute the following fairness metrics comparing Male vs Female, and Male vs Other:
    (a) Mean Salary Prediction Difference
    (b) Mean Absolute Error per group
    (c) Demographic Parity Difference (DPD)
    (d) Equal Opportunity Difference (EOD)
    (e) Predictive Equality
    (f) Disparate Impact Ratio (DIR)

    Clearly explain what each metric means, what the results imply for your model, and whether the model is biased and why. Also plot the residual distribution by gender.
11. Conduct a statistical test to see if mean residuals differ significantly between groups (e.g., t-test).
12. Identify whether the model systematically overestimates or underestimates for any group.
13. Use SHAP values (or LIME if preferred) to understand feature impact and produce a SHAP summary plot, also create SHAP dependence plots for top 3 features

Question 3: Deep Neural Network Classifier for Handwritten Digits

You are provided with a dataset of grayscale handwritten-digit images (like MNIST), each of size 28 × 28, flattened into a 784-dimensional vector. Using **PyTorch**, you must design and train a deep neural network for digit classification.

Construct a PyTorch model named `DigitClassifier` with the following architecture:

- Input: 784-dimensional vector
- Hidden Layer 1: 256 neurons, ReLU
- Hidden Layer 2: 128 neurons, ReLU
- Output Layer: 10 logits (digits 0–9)

Write the full class definition implementing the forward pass.

Now, train the network for 5 epochs using:

- Loss: CrossEntropyLoss
- Optimizer: Adam (learning rate 0.001)
- Batch size: 64

Assume you are given `train_loader` and `val_loader`. Write the full training loop using forward pass, loss computation, backward pass, and optimizer updates.

1. Why is ReLU preferred over Sigmoid and Tanh in deep networks? Provide two reasons.
2. Explain the role of PyTorch's autograd engine, including how it builds computation graphs and performs backpropagation.
3. Submit the python code