Instructions:

- **Due November 15, 11:59 PM uploaded to Blackboard as a single pdf report and a zipped file with all Matlab code.**
- Each problem is worth $\frac{100}{\text{number of problems}}$ points.
- Start each problem on a new page and put a box around your final answer to help the grader.
- The project submission should concisely summarize your work and include only relevant figures and your conclusions in a single PDF file. All code you used to generate the figures should be in a zipped file. **Both the pdf and the zipped file MUST be submitted**. If either is missing, the score will be marked as zero. There will be no extension or resubmission via email.
- Label your files as `ENME303_F24_MP01_yourname.pdf` and `ENME303_F24_MP01_yourname.zip`.
- A Matlab-published document that prints out the code, variables, and figures is NOT an acceptable report format. If the graders have to wade through a long document to find useful content, such as figures and a discussion of results, the submission will be penalized, even if all the correct results are somewhere in the document. For example, if it takes me more than 10 seconds to find a relevant figure in your report, I will not look for it.
- Do NOT include unnecessary Matlab outputs such as the entire Y or Phi matrices in your report. You are doing it wrong if your report is more than 7-8 pages.
- The graders are NOT expected to read your code or look for your conclusion in a sea of comments and code. I have instructed the graders to mark these "annoying" submissions with a zero.
- Double-check your submitted files after uploading them, and make sure that they are opening in Blackboard. Emailed submissions will not be graded.

Problem 1. Linear regression. In this problem, we will fit a linear model to the measured data. In particular, we will develop a model

$$y = mx + c \tag{1}$$

that best describes the measurements of x and y. Note that (1) can be written as

$$y = \phi\theta, \tag{2}$$

where

$$\phi = \begin{bmatrix} x & 1 \end{bmatrix}, \quad \theta = \begin{bmatrix} m \\ c \end{bmatrix}. \tag{3}$$

The objective is to find θ given several measurements of x and y.

1. Download the Matlab data file `data.mat` and load it in Matlab.
2. Use scatter function to plot the measurements of x and y. Does it look like that a straight line might fit the data?
3. Let x_i and y_i denote the ith measurement. Define

$$Y \triangleq \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \Phi \triangleq \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}. \tag{4}$$

   We would like to find θ such that

$$Y = \Phi\theta. \tag{5}$$

   Show that θ that exactly satisfies (5) does not exist.

4. The least-squares solution provides an estimate of θ that minimizes the error between the measured and predicted y. Let θ_{LS} denote the least-squares solution. Recall that θ_{LS} is given by

$$\theta_{\text{LS}} = (\Phi^{\text{T}}\Phi)^{-1}\Phi^{\text{T}}Y. \tag{6}$$

5. With the computed least-squares solution, compute

$$\hat{y}_i = \phi_i\theta_{\text{LS}} \tag{7}$$

   for $i = 1, \ldots, n$. Plot the predicted data $\hat{y}$ and the measured data y with respect to the measurements of x. Does it look like the linear model developed using the least-squares solution fits the data?

Problem 2. Nonlinear regression. In this problem, we will fit a linearly parameterized nonlinear model to the measured data. In particular, we will develop a model

$$y = \sum_{i=0}^{N} c_i x^i \tag{8}$$

that best describes the measurements of the nonlinear function $\sin x$. Note that (8) can be written as

$$y = \phi\theta, \tag{9}$$

where

$$\phi = \begin{bmatrix} 1 & x & x^2 & \cdots & x^N \end{bmatrix}, \quad \theta = \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_N \end{bmatrix}. \tag{10}$$

The objective is to find θ given several measurements of x and y.

1. In Matlab, define `x=-pi:0.01:pi`. Then, define `y = sin(x)`. Plot the measurements of x and y.
2. Let $N = 1$. Let x_i and y_i denote the ith measurement. Define

$$Y \triangleq \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \Phi \triangleq \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_n \end{bmatrix}. \tag{11}$$

   We would like to find θ such that

$$Y = \Phi\theta. \tag{12}$$

   Show that θ that exactly satisfies (12) does not exist.

3. The least-squares solution provides an estimate of θ that minimizes the error between the measured and predicted y. Let θ_{LS} denote the least-squares solution. Recall that θ_{LS} is given by

$$\theta_{\text{LS}} = (\Phi^{\text{T}}\Phi)^{-1}\Phi^{\text{T}}Y. \tag{13}$$

4. With the computed least-squares solution, compute

$$\hat{y}_i = \phi_i\theta_{\text{LS}} \tag{14}$$

   for $i = 1, \ldots, n$. Plot the predicted data $\hat{y}$ and the measured data y with respect to the measurements of x. Does it look like the linear model developed using the least-squares solution fits the data?

5. Compute the cumulative prediction error

$$e = \|Y - \hat{Y}\| \tag{15}$$

   using the norm function.

6. Repeat steps 2 to 5 for $N = 2, 3, \ldots, 20$. **Hint:** Write a for loop. Does the prediction improve as N increases? If yes, why? If no, why?

Problem 3. Auto-regressive Modeling. An auto-regressive (AR) model is a time-domain model of the form

$$y_{k+1} = \sum_{i=0}^{\ell-1} \beta_i y_{k-i}, \tag{16}$$

where $\ell > 0$ is the model memory and β_i are the model parameters. AR models are routinely used in various applications that require prediction such as signal processing and economics.

In this problem, you will develop an AR model to predict the new COVID cases. Download the Howard county COVID data here. You will use the first half of the data to *train* the AR model and the second half of the data to *validate/test* the trained model. Specifically, you will use linear regression to find the AR model parameters $\beta_0, \beta_1, \ldots, \beta_{\ell-1}$ using the training data.

1. Plot the new cases per day against the day.
2. Note that

$$y_{\ell+1} = \beta_0 y_\ell + \beta_1 y_{\ell-1} + \cdots + \beta_{\ell-1} y_1, \tag{17}$$
$$y_{\ell+2} = \beta_0 y_{\ell+1} + \beta_1 y_\ell + \cdots + \beta_{\ell-1} y_2, \tag{18}$$
$$\vdots \tag{19}$$
$$y_{\ell+N} = \beta_0 y_{\ell+N-1} + \beta_1 y_{\ell+N-2} + \cdots + \beta_{\ell-1} y_N, \tag{20}$$

   where N is the length of the training data. Write these equations in the $\Phi\theta = Y$ form.
3. Choose a value of ℓ and use the least-square solution to find the values of β_i that minimize the *training error*. **Hint:** try $\ell = 10$.
4. With the trained model, predict the new cases for the second half of the pandemic, that is, $\hat{y}_{\ell+N+1}, \ldots, \hat{y}_M$, where M is the total length of the data. Plot the true value and the predicted value on the same plot.
5. For each predicted value, compute the relative error, that is,

$$e_{k,\text{rel}} = \frac{y_k - \hat{y}_k}{y_k} \tag{21}$$

   and plot it.
6. Repeat the steps above for five different choices of ℓ.
7. Comment on the AR model's ability to predict the new cases as well as its limitations.