Problem Set 04 · Linear Regression

Instructions

1. This assignment requires both Excel and MATLAB.
2. Read each problem carefully before starting your work. You are responsible for following all instructions within each problem. Remember that all code submissions must follow the course programming standards.
3. Below are the expected deliverables for each problem.
    - Name your files to match the format in the table below.
    - Publish your code for each problem. See PS02 for more information.
    - Do not forget to include any data files loaded into your code.

Item	Type	Deliverable to include in Submission
Problem 1: Power Plant Output (Excel)	Individual	□ PS04_power_plant_excel_*yourlogin*.xlsx
Problem 2: Power Plant Output (MATLAB)	Individual	□ PS04_power_plant_*yourlogin*.m □ PS04_power_plant_*yourlogin*_report.pdf □ Data file loaded into your m-file
Problem 3: Global Mean Sea Level	Individual	□ PS04_GMSL_*yourlogin*.m □ PS04_GMSL_*yourlogin*_report.pdf □ Data files loaded into your m-file

4. Save all files to your Purdue career account in a folder specific to PS04.
5. When you are ready to submit your assignment,
    - Compress all the deliverables into one zip file and name it **PS04_yourlogin.zip**. Be sure that you
        i. Only compress files using **.zip** format. No other compression format will be accepted.
        ii. Only include deliverables. Do **not** include the problem document, blank templates, etc.
    - Submit the zip file to the Blackboard drop box for PS04 before the due date.
6. After grades are released for this assignment, access your feedback via the assignment rubric in the My Grades section of Blackboard.

Notes Before You Start

Formatting Reminder

Always format your text, plots, and numerical outputs in a professional manner.

- Numerical values must have a reasonable number of decimal places. Include units when necessary.
- Displayed text should be descriptive and professional. Use complete sentences.
- Plots require a title, x- and y-axis labels, and gridlines. Multiple data sets on a single plot require a legend.

Problem 1: Power Plant Output (Excel)

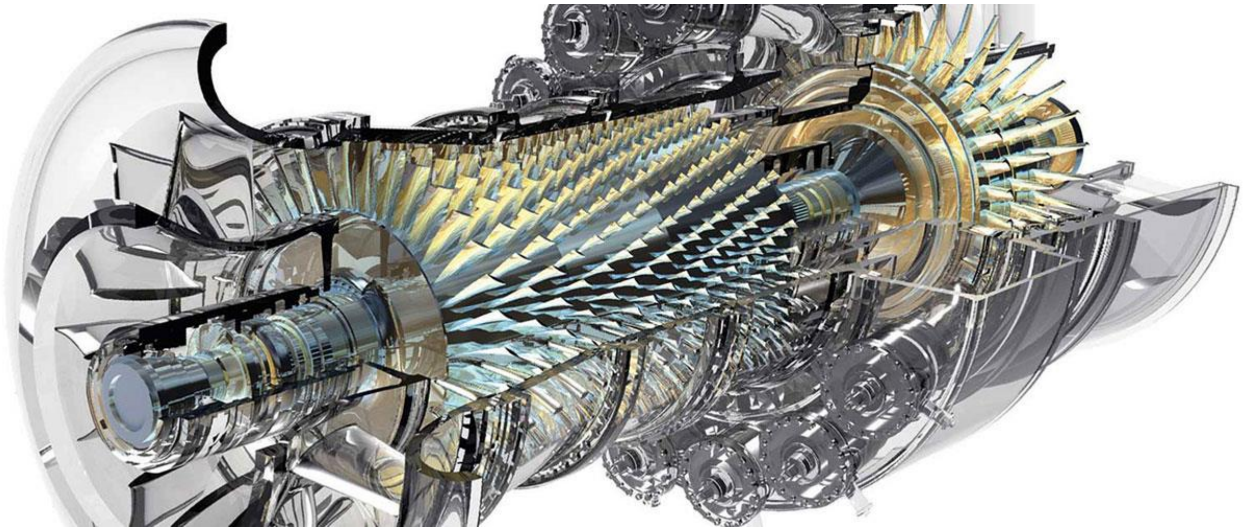
Individual

Learning Objectives

Below are learning objectives that may be used to assess your work on this problem. Learning objectives from past assignments may also be used to assess your work. Use the links to find the full evidence lists for each topic.

Linear Regression	12.02 Compute and present in equation form the coefficients of a best-fit linear model using visual approximation and the two-point method
	12.03 Manually compute the SSE
	12.04 Manually compute the SST
	12.05 Manually compute the r-squared value from SSE and SST
	12.06 Add a trendline to a scatter plot of raw x-y data (Excel)
	12.07 Display the equation and r-squared value of a trendline added to a scatter plot (Excel)
	12.08 Manually compute and present in equation form the coefficients of a best-fit linear model using least-squares method
	12.14 Interpret the slope and intercept of a best-fit linear model
	12.15 Interpret the r-squared value
	12.17 Use the best-fit linear model to make predictions only when appropriate

Problem Setup



Natural gas power plants rely on gas combustion turbines to produce electricity. Ambient air temperature can have a significant influence on the power output of a gas turbine. The main effect is due to the inverse relationship between air temperature and air density. As temperature increases, density decreases. Decreased air density reduces the air mass flow entering the compressor, given constant volumetric flow.

Your task is to quantify how the ambient air temperature affects a specific power plant. The plant operators provided you with a data file, named **Data_power_measurements.csv**, of ambient air temperatures and the corresponding hourly energy output by the plant. The data are representative of the plant's performance at full capacity.

Using the data file provided, you will perform linear regression to determine how the ambient air temperature affects the plant's net electrical power output.

Problem Steps

1. Open the Excel template file and fill out the appropriate header information. Save the workbook with the name format required by the deliverables list.
2. Use this workbook to complete all of your computational work for this problem.
   a. Complete your work in the appropriate section of the sheets. Plots should be in the Output Section. You can add extra columns to a section as needed, but do not change the order of the sections

A. Two-Point Method of Regression

3. Load the data into the **Two Point** worksheet in the Excel workbook.
4. Use the two-point method to determine a linear model of the data
   a. Create a scatter plot of the data.
   b. Use Excel draw tools (INSERT>Illustrations>Shapes) to draw a reasonable best-fit line over the data in the scatter plot.
   c. Use the two-point method to determine the linear model (in the form $y = ax + b$). Show your work in the calculations section of the worksheet.

   Hint: Your two points need to be on the line you drew, not necessarily two actual data points from the data set.

   d. Calculate the SSE, SST, and r^2 values for the linear models. Show your work in the calculations section of the worksheet.
5. On the **Analysis** worksheet:

   Q1: Report the equation (using clear, appropriate variable names in place of x and y in the equation) and the SSE, SST, and r^2 for your linear model.

   Q2: Explain how well your model represents the relationship between the data. Justify your answer.

   Q3: Use your model to predict the power output when the ambient temperature is 20 deg C.

   Q4: What is the meaning of the slope of your model?

B. Manual Least Squares Regression

6. Load the data into the **Least Squares** worksheet in the Excel workbook
7. Use the <u>**manual**</u> least squares method to determine a linear model of the data
   a. Solve for coefficients *a* and *b* in the linear model $y = ax + b$. Show your work in the Calculations section of the worksheet.
   b. Calculate the SSE, SST, and r^2 for the linear model.
8. On the **Analysis** worksheet:

Q5: Report the linear model (in form *y = ax + b*). Define and use appropriate variable names in place of x and y in the equation. Report SSE, SST, and r^2 for the model.

Q6: Use your model to predict the power output for an ambient temperature of 20 deg C and of 40 deg C. Justify each prediction using your knowledge of the original data set and your linear model.

Q7: Compare the two point method model to the least squares model. Which model is the better fit to the data? Justify your answer using r^2.

C. Excel Least Squares Regression

9. Continue working with the data in the **Least Squares** worksheet.

10. Use the Excel built-in linear regression method:

    a. Create a scatter plot of the data.

    b. Add a linear trendline.

    c. Display the equation and the r^2 values on the plot. Replace x and y in the trendline equation with clear, appropriate variable names.

Reference: Combine-cycle gas & steam turbine power plants

Image: https://www.ge.com/power/resources/knowledge-base/what-is-a-gas-turbine

Problem 2: Power Plant Output (MATLAB)

Individual

Learning Objectives

Below are learning objectives that may be used to assess your work on this problem. Learning objectives from past assignments may also be used to assess your work. Use the links to find the full evidence lists for each topic.

Scripts	04.00 Create and execute a script
Variables	02.00 Assign and manage variables
Arrays	03.00 Manipulate arrays (vectors or matrices)
Text Display	05.00 Manage text output
Import Data	06.00 Import numeric data stored in .csv and .txt files
Plotting	07.00 Create and evaluate x-y plots suitable for technical presentation
Linear Regression	12.03 Manually compute the SSE
	12.04 Manually compute the SST
	12.05 Manually compute the r-squared value from SSE and SST
	12.09 Compute the coefficients of a best-fit linear model using least-squares method (MATLAB)
	12.10 Compute predicted values using the best-fit linear model (MATLAB)
	12.11 Plot the best-fit linear regression line on a plot of raw x-y data (MATLAB)
	12.12 Display the results of linear regression (MATLAB)
	12.16 Compare data sets based on their best fit linear models and r-squared values

Problem Setup

Convert the least squares analysis from Problem 1 into a MATLAB program. Determine a linear model for the same data using MATLAB's functionality. Create a script that determines the linear model using the data provided in Problem 1 and then use the resulting model to make predictions.

Problem Steps

1. Open **PS04_power_plant_template.m** and complete the header. Save it using the name format given in this problem's deliverables list. Use programming standards to place code in the appropriate sections within the template.
   a. Perform linear regression on the power plant data using the `polyfit` command.
   b. Compute the predicted values of the linear model.
   c. Calculate the SSE, SST, and r^2 values of the model.
   d. Display the linear model equation (with clear variable names), SSE, SST, and r^2 to the Command Window.
   e. Generate a data plot and overlay your linear model on the data.
2. In the `ANALYSIS` section of your code:

   Q1: Compare the Excel and MATLAB least squares models. What observations can you make?
3. Publish your code to a PDF file and save it using the name format given in the deliverables list for this problem.

Problem 3: Global Mean Sea Level

Individual

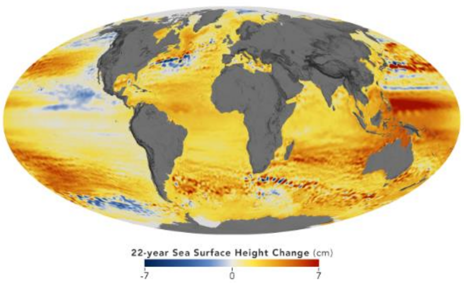
Learning Objectives

Below are learning objectives that may be used to assess your work on this problem. Learning objectives from past assignments may also be used to assess your work. Use the links to find the full evidence lists for each topic.

Scripts	04.00 Create and execute a script
Variables	02.00 Assign and manage variables
Arrays	03.00 Manipulate arrays (vectors or matrices)
Text Display	05.00 Manage text output
Import Data	06.00 Import numeric data stored in .csv and .txt files
Plotting	07.00 Create and evaluate x-y plots suitable for technical presentation
Linear Regression	12.03 Manually compute the SSE
	12.04 Manually compute the SST
	12.05 Manually compute the r-squared value from SSE and SST
	12.09 Compute the coefficients of a best-fit linear model using least-squares method (MATLAB)
	12.10 Compute predicted values using the best-fit linear model (MATLAB)
	12.11 Plot the best-fit linear regression line on a plot of raw x-y data (MATLAB)
	12.12 Display the results of linear regression (MATLAB)
	12.14 Interpret the slope and intercept of a best-fit linear model
	12.15 Interpret the r-squared value
	12.16 Compare data sets based on their best fit linear models and r-squared values
	12.17 Use the best-fit linear model to make predictions only when appropriate
Relational & Logical Operators	14.00 Perform and evaluate relational and logical operations

Problem Setup

As sea ice melts, ocean temperatures increase, and as the sea floor changes due to geologic shifts, the sea level along coastlines around the world will change. People have been monitoring sea levels for a long time since many population and economic centers are on or near coastlines. One way to measure sea level is to use tide gauges; another is to use satellite-based altimeters. Each method has benefits and drawbacks. Both methods can be used to determine the global mean sea level (GMSL) change over time.



As an enviromental engineer who works with ocean infrastructure, you need to study two GMSL measurement types over a common time span. You have two data files:

- **Data_CSIRO_gmsl_mo_2013.csv**, which contains tide gauge data from 1880-2013. All GMSL values are the difference in sea level compared to the midpoint measurement in 1990.
- **Data_NASA_altimeter_gmsl_meas.txt**, which contains satellite altimeter data from 1993 to the beginning of 2018. All GMSL values are the difference in sea level compared to the 20-year mean from 1996-2016.

You will examine the two data sets by performing linear regression on each data set over a common time span of 1993 – 2013.

Problem Steps

1. Open the script **PS04_GMSL_template.m** file. Complete the header information. Save your script with the name format required by the deliverables list.
2. Open the data files and review the information they contain.
3. Write the code to perform least-squares linear regression to model the relationship between sea level and year for each data set.
    a. Import both data sets and copy each relevant data column into a separate variable.

    b. Use relational and logical operators to find the data that correspond to the 1993 – 2013 time span. Include all 1993 entries and all 2013 entries in the data.

    c. Compute the best-fit line equation for each data set's relationship.

    d. Compute the coefficient of determination for each linear model.

    e. Display to the Command Window each linear model equation (with appropriately named variables) and the coefficient of determination with reference to the data type represented in the model.

    f. Plot the data with its regression model for each data set.

    - Display the plots in one figure with a 2x1 subplot grid.
    - Display the tide gauge data overlaid by its linear model on the bottom subplot.
    - Display the satellite altimeter data overlaid by its linear model on the top subplot.
    - Format the plots for technical presentation.
4. Run your script. Then, in the `ANALYSIS` section of the script, answer these questions:

    Q1: What do you know about the accuracy of the two data collection methods (i.e., satellite altimeter and tide gauge) in measuring global sea level? Explain your answer.

    Q2: For which data collection method does a linear model best explain the variation that exists in the data? Clearly state the basis of your reasoning.

    Q3: Which model shows the fastest global mean sea level rise?

    Q4: Your supervisor has asked you about predicting the GMSL in 2019 using both the satellite and the tide models. Respond to your supervisor with a short paragraph explaining your results.
5. Publish your script as a PDF and name it as required in the Deliverables List.

References

GSFC. 2017. Global Mean Sea Level Trend from Integrated Multi-Mission Ocean Altimeters TOPEX/Poseidon, Jason-1, OSTM/Jason-2 Version 4.2 Ver. 4.2 PO.DAAC, CA, USA.

https://climate.nasa.gov/vital-signs/sea-level/

http://www.cmar.csiro.au/sealevel/sl_data_cmar.html

image: https://earthobservatory.nasa.gov/IOTD/view.php?id=91746