Problem Set 04 · Linear Regression

Instructions:

1. Each problem in this problem set has a set of deliverables for you to submit. You are responsible for following the appropriate guidelines and instructions below. Create appropriately-named files as instructed.

2. Save all files to your Purdue career account in a folder specific to PS04.

3. Compress all deliverables into one zip folder named **PS04_yourlogin.zip**. Submit the zip file to the Blackboard drop box for PS04 before the due date. *REMEMBER*:
   - Only include deliverables. Do not include the problem document, blank templates, etc.
   - Only compress files into a .zip folder. No other compression format will be accepted.

Deliverables List

Item	Type	Deliverable
Problem 1: Global Temperature Anomaly (Excel)	Individual	PS04_global_temp_excel_*yourlogin*.xlsx
Problem 2: Global Temperature Anomaly (MATLAB)	Paired	PS04_global_temp_*yourlogin_yourlogin2*.m PS04_global_temp_*yourlogin_yourlogin2*_report.pdf All data files that are loaded into your m-file
Problem 3: Combined Cycle Power Plant	Individual	PS04_ccpp_regression_*yourlogin*.m PS04_ccpp_regression_*yourlogin*_report.pdf All data files that are loaded into your m-file

Problem Set 04 · Linear Regression

Problem 1: Excel - Global Temperature Anomaly

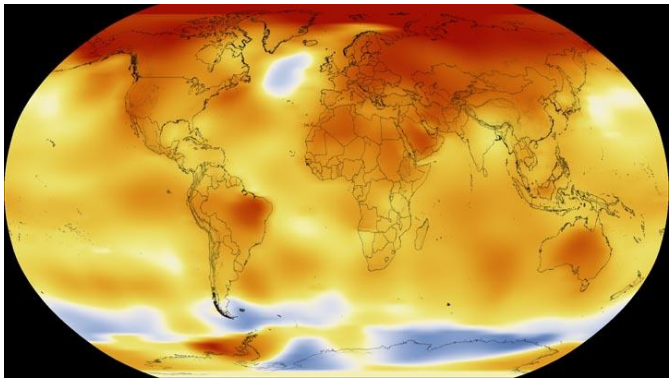
Individual Analysis

Learning Objectives

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

In recent years, there has been a lot of discussion about whether or not the world's temperatures are rising. A few factors can be used to assess climate change, such as change in carbon dioxide levels in the atmosphere, change in global temperatures, percentage change in arctic ice as well as land ice, as well as change in sea level.



NASA/GISS Global Temperature Anomaly, 2016

You work as a civil engineer for a global engineering firm that works with cities that are experiencing climate-related issues. Your firm is investigating infrastructure needs of these cities and needs to better understand the climate to anticipate their clients' needs.

You have been asked to contribute to your firm's climate-change analysis. You will examine global temperature anomaly as a potential indicator of climate change. Temperature anomaly is how much a temperature deviates from a specified base temperature. NASA's Goddard Institute for Space Studies (GISS) uses the mean global temperature over the 30-year period 1951-1980 as the base temperature.

Problem Set 04 · Linear Regression

You have been provided with a GISS data set that shows global temperature anomaly for each year since 1980, with a few years missing from the data. You will use this data file, named **Data_global_temp_anomalies.txt**, to perform linear regression.

Problem Steps

1. Open the Excel template file and fill out the appropriate header information.
2. Save the Excel workbook **as instructed** in the Deliverables list. 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 temperature anomaly 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 temperature anomaly in the year 1995.

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

B. Manual Least Squares Regression

6. Load the fuel price and fuel cost data into the **Least Squares** worksheet in the Excel workbook
7. Use the **manual** 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:

Problem Set 04 · Linear Regression

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 temperature anomaly for the years 1995 and 2020. 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.

References

Data: https://data.giss.nasa.gov/gistemp/

Image: https://climate.nasa.gov/vital-signs/global-temperature/

Problem 2: MATLAB – Global Temperature Anomaly

Paired Programming

Learning Objectives

The solution to this problem requires knowledge of the learning objectives below.

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

Problem Set 04 · Linear Regression

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_global_temp_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 fuel 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 Set 04 · Linear Regression

Problem 3: Combined Cycle Power Plant

Individual Programming

Learning Objectives

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

Problem Setup

A combine cycle power plant generates electricity using gas and steam turbines to produce electricity. A gas turbine uses combustion to produce electricity. The heat generated from the combustion is recovered and used to create steam, which then runs a steam turbine to produce additional electricity.

Problem Set 04 · Linear Regression

Your task is to quantify how different factors affect a specific plant's power output. You have data that has been collected from the plant when the plant was set to work at full capacity. The data set contains three atmospheric conditions, one plant condition, and the net hourly electrical output from the plant in those conditions.

Using the data file provided, **Data_CCPP_measurements.csv**, you will write a script that will perform linear regression to determine how each of the four conditions affects the plant's net electrical power output.

Problem Steps

1. Open **PS04_ccpp_template.m**. Complete the header information. Save your script with the name format required by the deliverables list.
2. Write the code to perform linear regression on the data for each condition and the plant power output. The script must:
    a. Compute the linear coefficients and display to the Command Window the best-fit line equation for each condition's relationship to power output. Display each linear model equation (with appropriately-named variables) and reference the specific condition that is being related to power output.
    b. Compute the coefficient of determination for each condition's relationship to power output and display it to the Command Window. Display each relationship's coefficient of determination with a reference to the specific condition that is being related to power output.
    c. Plot the data with its least squares regression model for each condition's relationship with power output.
        - The plots must be displayed in one figure with a 2x2 subplot grid.
        - Each subplot must show one of the relationship's data overlaid with its linear model.
        - Each subplot must use unique color formatting.

        Depending on your version of MATLAB, you may have the `suptitle` built-in function. Using this command will allow you to put one overarching title on the figure window, thus permitting you to use short titles on each subplot.

        Hint: You can also control the font size of all the text associated with a subplot (i.e., labels, title, tick mark numbers) by using the FontSize property. For example, to change a subplot's font size to 8-point, place the following command after the subplot's formatting commands:

        ```
        set(gca,'FontSize',8)
        ```

        If you chose to use the `set` command above to manage the font size, then you will need that line of code for each subplot.
3. Run your script. Then, in the `ANALYSIS` section, answer these questions from the Problem Setup:
    Q1: For which condition does a linear model best explain the variation that exists in the data? Clearly state the basis of your reasoning.
    Q2: Which of the four conditions has the largest effect on the plant's power output? Which condition has the smallest effect? Clearly state the basis of your reasoning.
4. Publish your code to a PDF file and save it using the name format given in this problem's deliverables list.

Problem Set 04 · Linear Regression

References

https://www.tva.gov/Energy/Our-Power-System/Natural-Gas/How-a-Combined-Cycle-Power-Plant-Works

Image: https://powergen.gepower.com/resources/knowledge-base/combined-cycle-power-plant-how-it-works.html

Data adapted from https://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant

Pınar Tüfekci, Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods, International Journal of Electrical Power & Energy Systems, Volume 60, September 2014, Pages 126-140, ISSN 0142-0615.

Heysem Kaya, Pınar Tüfekci , Sadık Fikret Gürgen: Local and Global Learning Methods for Predicting Power of a Combined Gas & Steam Turbine, Proceedings of the International Conference on Emerging Trends in Computer and Electronics Engineering ICETCEE 2012, pp. 13-18 (Mar. 2012, Dubai).