Problem Set 06
Linear Regression
Problems

Instructions:

1. This problem set contains paired programming and individual programming problems. Each problem has a set of deliverables that needs to be submitted. 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 PS06.
3. Compress all deliverables into one zip file named **PS06_yourlogin.zip**. Submit the zip file to the Blackboard drop box for PS06 before the due date.

Problem Set

Item	Type	Deliverable
Problem 1: Excel Regression	Individual	PS06_excel_regression_*yourlogin*.xlsx
Problem 2: MATLAB Regression	Paired	PS06_fuelcost_model_*yourlogin*.m PS06_fuelcost_model_*yourlogin*_report.pdf
Problem 3: Seasonality of Passenger Enplanements	Individual	PS06_enplanements_*yourlogin*.m PS06_enplanements_*yourlogin*_report.pdf

Formatting Reminder

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

- Numerical values must have a reasonable number of decimal places and include units when necessary.
- Displayed text should be descriptive and professional. Use complete sentences.

Problem 1: Regression in Excel

Individual Programming

Problem Setup

You work for a company that wants to provide a fuel-cost estimation tool to US airlines. In Problem 2 of PS02, you created scatter plots and examined airfares and airline fuel costs. In this problem, you will revisit one of those plots and use linear regression to model the relationship.

In PS02, you saw that fuel price appeared to be a good predictor of fuel costs. In this problem, you will use the two-point method and the least squares method in Excel to find a best-fit linear model for fuel price and cost data. Then you will quantify the extent to which each linear model explains the variation that exists in the fuel price and fuel cost data .This resulting model has the potential to be used by US airlines to estimate fuel cost trends.

Problem Set 06
Linear Regression
Problems

Use the data file named **Data_fuelcost.csv** that contains fuel price and cost information for all domestic US flights. You will recreate the Fuel Cost Plot from PS02, Problem 2 and perform linear regression on the data.

Problem Steps

1. Open the Excel template file and fill out the appropriate header information.
2. Save the Excel workbook as **PS06_excel_regression_*yourlogin*.xlsx**. 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 fuel price and fuel cost 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 fuel costs if fuel price is $2.35/gal.

    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.

Problem Set 06
Linear Regression
Problems

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 fuel costs if the fuel price is $2.35/ gal and $4.25/gal. 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.

Problem 2: Regression in MATLAB

Paired Programming

Problem Setup

Your company has decided that a MATLAB model of fuel prices and costs would be a beneficial tool to add to its fuel-cost estimation package. Convert the least squares analysis from Problem 1 into a MATLAB program. Determine a linear model for the same data using MATLAB's built-in functions. Create a user-defined function that determines the linear model using the data provided in Problem 1 and then use the resulting model to make predictions. This user-defined function must accept the model's independent variable as an input argument and return the model's dependent variable as an output argument.

Problem Steps

1. Open *PS06_fuelcost_model_template.m* and complete the header. Save it as **PS06_fuelcost_model_*yourlogin*.m**.

2. Create a user-defined function as described in the Problem Setup. Use programming standards to place code in the appropriate sections within the UDF template. Use a fuel price of $2.35/gal as the test case while you write and debug the function.

a. Perform linear regression on the fuel data using the `polyfit` command. The fuel data should be loaded in the function as to opposed to passed as an input argument.

b. Compute the predicted values of the linear model using the `polyval` command.

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 scatter plot and overlay your linear model on the data.

3. In the `ANALYSIS` section of your code:

   Q1: Compare the Excel and MATLAB least squares models. What observations can you make?

4. Publish your code to a PDF file using a fuel price input of $3.00/gallon. Name the published file **PS06_fuelcost_model_*yourlogin*_report.pdf**.

Problem 3: Seasonality of Airline Passenger Enplanements

Individual Programming

Problem Setup

Your company wants to offer a new tool that examines the best time to fly internationally for both passengers and airlines. You have a set of data, in a file named **Data_airpassengers_seasons.csv**, that contains the passenger enplanements for all international U.S. Air Carrier flights. The US Department of Transportation collected and reported these data by month every two years, from 1996 to 2014.

Your supervisor asks you to answer the following questions:

- For which season does a linear model best explain the variation that exists in the data?
- Which season had the greatest growth rate in number of passenger enplanements per year? Which season had the lowest growth rate?

Write a set of three user-defined functions to perform the least squares analysis on the data and use it to answer the questions above. One UDF will compute the linear model coefficients for a season's data and display that season's best-fit line. A second UDF will compute a season's predicted model values and the model's r^2 value and then will display the model's r^2 value. The third function will be an executive function that calls the two sub-UDFs and allows you to analyze and display the data.

Problem Steps

1. Create a sub-UDF that uses *PS06_enplanements_subUDF_template.m* as the program template and:
   - Is named **PS06_enplanement_coefs_*yourlogin*.m**

Problem Set 06
Linear Regression
Problems

- Accepts a string variable with the season's name and the inputs necessary to calculate the linear model coefficients of a set of a season's x-y data
- Returns the linear model coefficients as two separate variables
- Displays to the Command Window the linear model equation (with appropriately-named variables) and reference the specific season

  You can save character strings to variable names and use them in fprintf statements.
- Has code in the appropriate sections of the template

2. Create a sub-UDF that uses *PS06_enplanements_subUDF_template.m* as the program template and:
   - Is named **PS06_enplanements_predict_*yourlogin*.m**
   - Accepts inputs necessary to calculate the model's predicted values and the r^2 for the linear model
   - Returns the linear model's predicted values as a vector and the r^2 value for the linear model as a scalar
   - Displays to the Command Window the season's r^2 value and refers to the specific season in the statement
   - Has code in the appropriate sections of the template
3. Create an executive (no-input, no-output) UDF that uses *PS06_enplanements_exec_template.m* as the program template and:
   - Is named **PS06_enplanements_exec_*yourlogin*.m**.
   - Calls the two sub-UDFs to perform the regression analysis
   - Plots the data with its least squares regression for each season in the data set.
     - These plots must be displayed in one figure with a 2x2 subplot grid.
     - Each subplot must show one of the season's data overlaid with its linear model.
     - Each subplot must use unique color formatting.
     - The axes of the subplots must allow the plots to be comparable.
4. Run your UDFs. Then, in the `ANALYSIS` section of your executive function code answer these questions:

   Q1: For which season does a linear model best explain the variation that exists in the data? Clearly state the basis of your reasoning.

   Q2: Which season had the greatest growth rate in number of passenger enplanements per year? Which season had the lowest growth rate? Clearly state the basis of your reasoning.
5. Publish your executive function code to a PDF file named **PS06_enplanements_exec_*yourlogin*_report.pdf**.