ENGR 132 Practice Exam Problems
Linear Regression

Problems

Practice 1.

Velocity data was acquired during an acceleration test of a concept vehicle that is powered by fuel cells. To determine the line of best fit for the data, the following MATLAB script was created to perform linear regression by the least squares method:

```
1  % --- INPUTS ---
2  time = [3 5 7 10]; % Time [s]
3  velocity = [10 25 30 50]; % Velocity [m/s]
4
5  % --- CALCULATIONS ---
6  output_1 = polyfit(time, velocity, 1);
7
```

When the script is executed, the following vector is assigned to the variable **output_1** in the MATLAB Workspace:

```
output_1 = [6 -6]
```

Using the results produced from executing the script,

A. Write the linear model, **y = ax + b,** for the given data.

B. Fill in the values for $f(x_1)$ and $f(x_2)$ in the table below.

i	x_i (Time [s])	y_i (Velocity [m/s])	x_i^2	$x_i * y_i$	$f(x_i)$	e_i	e_i^2	$[y_i - \bar{y}]^2$
1	3	10	9	30		- 2	4	361
2	5	25	25	125		1	1	16
3	7	30	49	210	36	- 6	36	1
4	10	50	100	500	54	- 4	16	441
Sum	25	115	183	865		-11	57	819
Mean	6	29	46	216		-3	14	205

C. Write the line of code for line 7 that will calculate the predicted velocities for the given time values. You must use the variables defined in the MATLAB script above and an appropriate built-in MATLAB command.

D. Set up the equation necessary to determine the coefficient of determination, r^2. This means, fill in the numerical values for SSE and SST in the equation. **You DO NOT need to solve the equation.**

E. What is the meaning of the r^2 for this data set?

a. Error due to problems during the acquisition of raw data measurements

b. Error between the raw data points and the predicted velocity data

c. Variation between the raw data points and the mean of the predicted velocity data

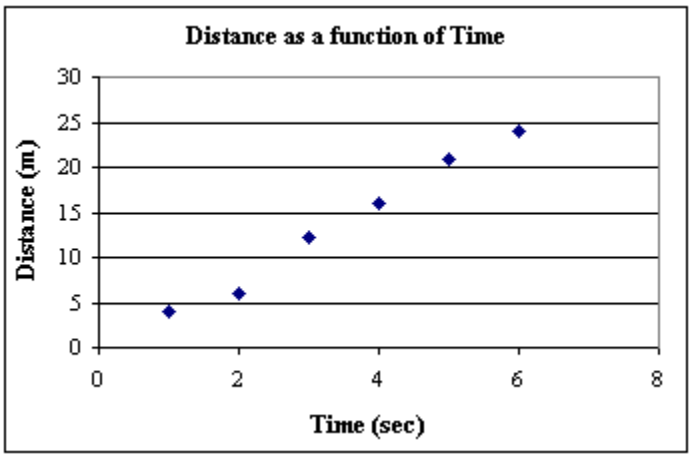
d. Variation of the raw data points around the regression line

e. Can't tell from the data

Solution

Practice 2.

Use the information shown in the graph and summary table to answer the following questions.



	x	y	x^2	x*y	f(x)	e_i	e_i^2	$[y_i - \bar{y}]^2$
	1	4	1	4	3.25	0.75	0.5625	97.51563
	2	6	4	12	7.5	-1.5	2.25	62.01563
	3	12.25	9	36.75	11.75	0.5	0.25	2.640625
	4	16	16	64	16	0	0	4.515625
	5	21	25	105	20.25	0.75	0.5625	50.76563
	6	2	36	144	24.5	-0.5	0.25	102.5156
Sums:	21	83.25	91	365.75	83.25	0	3.875	319.9688
Mean:	3.5	13.875	15.1667	60.9583	13.875	0	0.64583	53.32812

A. Show the equations needed to determine the coefficients for the regression line. Any known quantities should be substituted into equations as appropriate.

B. What percentage of the variance in the data is explained by the data?

Solution

Practice 3.

A MATLAB script has been started that will:

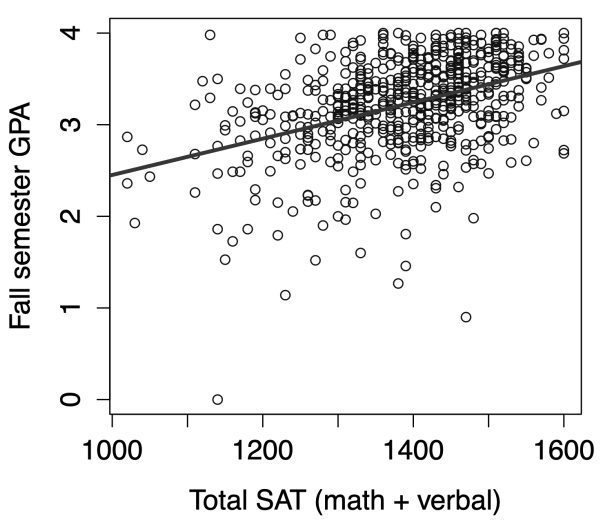
- read in a data file,
- plot the raw data using red squares,
- calculate the linear regression line,
- plot the linear regression line with a black solid line on the same graph and
- display to the screen the value of the slope of that line.

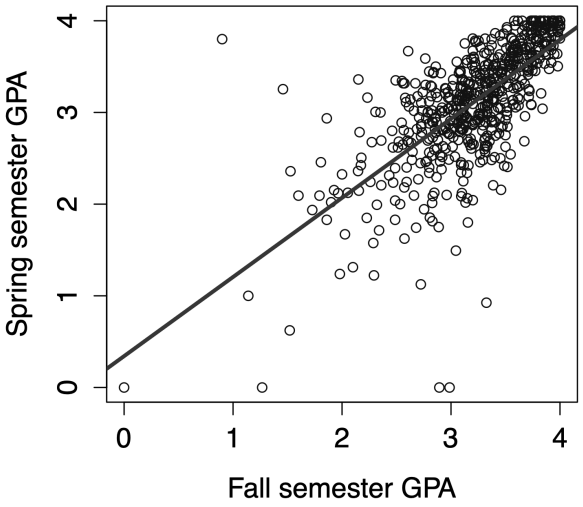
Starting at line 10, complete the code to accomplish these tasks. You do not need to add any additional comment statements and you may end up with blank lines in the table.

Line #	MATLAB Code
1	`% ENGR 126 Spring 2009, Exam 2 MATLAB question`
2	`clear all   % This command clears all variables and functions`
3	`clc         % This command clears the command window`
4	`close all   % This command deletes all figures`
5	`% ---- Input Section ----`
6	`load ('exam2_data.txt');  % Reads in data containing`
7	`resi = exam2_data(:,1);   % creates vector of Resistivity in microohms-cm`
8	`tempe = exam2_data(:,2);  % creates vector of temperature in degrees C`
9	`% ---- Calculation Section ----`
10	
11	
12	
13	
14	
15	`% ---- Output Section ----`
16	`% Generate a plot of the data suitable and regression line for technical presentation`
17	
18	
19	
20	
21	
22	
23	
24	
25	`%  Display the value of the slope predicted by the model`
26	
27	
28	
29	
30	
31	

Solution

Practice 4.





A hypothetical west coast university is interested in developing a model for first-year student performance. They used the two plots above to examine (for a particular cohort of first-year students): (i) the relationship between SAT score and Fall semester GPA, and (ii) the relationship between Fall semester GPA and Spring semester GPA. Their linear regression models for the two datasets (shown as solid black lines on the figures) are:

- Fall GPA-SAT: $m_{fall} = 0.002$, $b_{fall} = 0.5$, SSE = 125, SST = 150
- Fall GPA-Spring GPA: $m_{spring} = 0.9$, $b_{spring} = 0.25$, SSE = 125, SST = 500

.Answer the following questions about this data:

A. Calculate the coefficient of determination (r^2) for each of the two linear models above. Show your work in the space provided on the answer sheet.

B. An in-coming student from the cohort studied here earned a total SAT score of 1250. Using the two linear models above, predict this student's fall semester GPA and spring semester GPA. Please show your work in the space provided on the answer sheet.

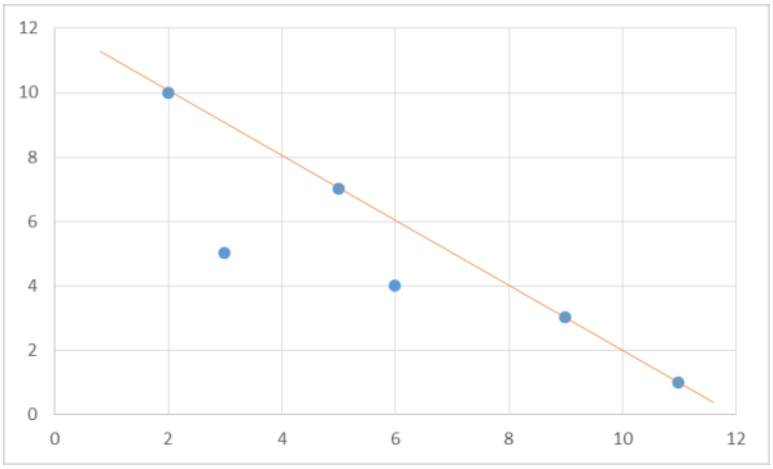
Fall GPA: _____ Spring GPA: _____

C. Briefly explain (1-2 sentences) which of these two models is more predictive of student performance, and provide specific justification.

Solution

Practice 5.

A team of engineers collected some data, included in the table below. They have performed Linear Regression "By Eye" and drawn the line shown on the plot.



x	y
2	10
3	5
5	7
6	4
9	3
11	1

A. Using the two-point method, choose two points on the regression line and calculate the slope (m) and the y-intercept (b) of the model. Declare the (x,y) coordinates of the points you used to calculate m and b and show the equation you used.

Point 1: _____ m = _____

Point 2: _____ b = _____

B. Use the least squares method to set up the simultaneous equations that could be solved for the slope ***a*** and intercept ***b***. Fill in the boxes in the two equations below the table. **You do NOT need to solve these equations.**

	x	y	x^2	x*y	y^2
	2	10	4	20	100
	3	5	9	15	25
	5	7	25	35	49
	6	4	36	24	16
	9	3	81	27	9
	11	1	121	11	1
sums	36	30	276	132	

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$$ a ☐ + b ☐ = ☐

$$a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i$$ a ☐ + b ☐ = ☐

C. One of the engineers has started writing a MATLAB script to perform a linear regression analysis. Read the comments provided and complete lines 5, 7, 9, and 11 accordingly.

```
1  % Load x and y vectors
2  x = [2 3 5 6 9 11];
3  y = [10 5 7 4 3 1];
4  % calculate coefficients of linear model
5  coeff =
6  % calculate fitted y
7  y_modeled =
```

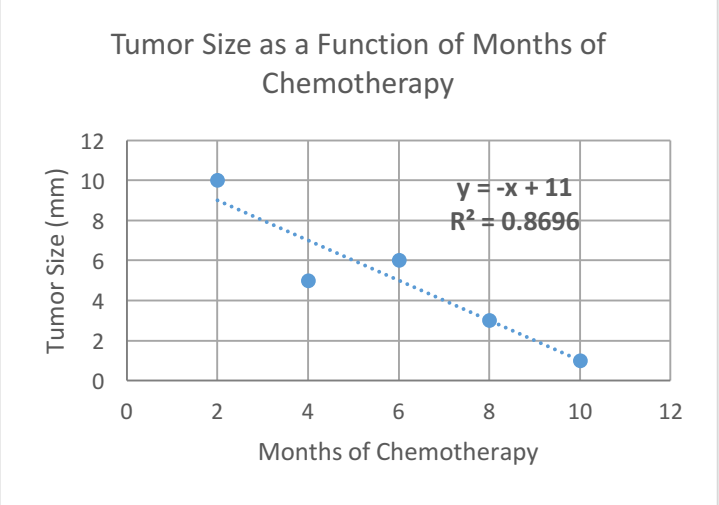
```
8  % calculate SSE
9  SSE =
10 % Calculate SST
11 SST =
12 % Calculate R^2
13 r_squared = 1 - SSE/SST;
```

Solution

Practice 6.

A new method of chemotherapy is being tested to treat small tumors of various types of cancer. Tumors of random patients receiving the new method were measured at various times during their treatment. The data is shown in the table and the scatter plot below.

Months of Chemotherapy	Tumor Size [mm]
2	10
4	5
6	6
8	3
10	1



According to the model, what would be the size of a tumor after

A. 5 months

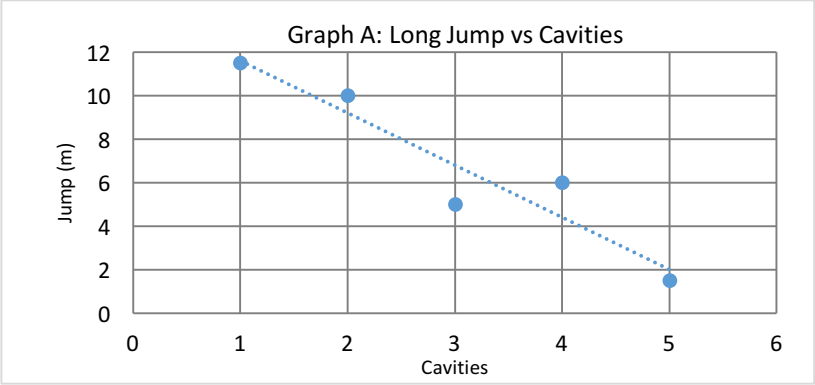
B. 9 months

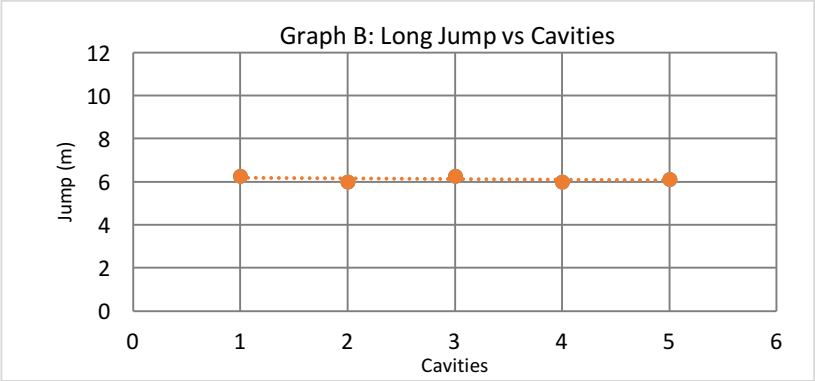
C. 11 months

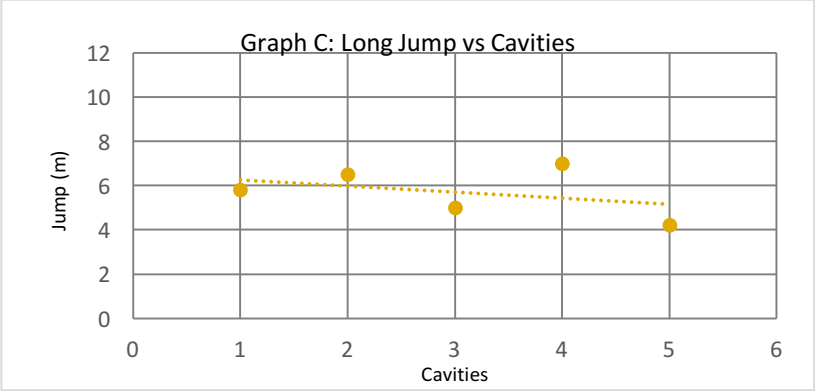
Solution

Practice 7.

This is a problem derived directly from your class slides. ***Therefore, there are no partial credits for this problem.*** The graphs below represent the relationship between long jump and cavities.







Based on the above graphs, answer the following questions.

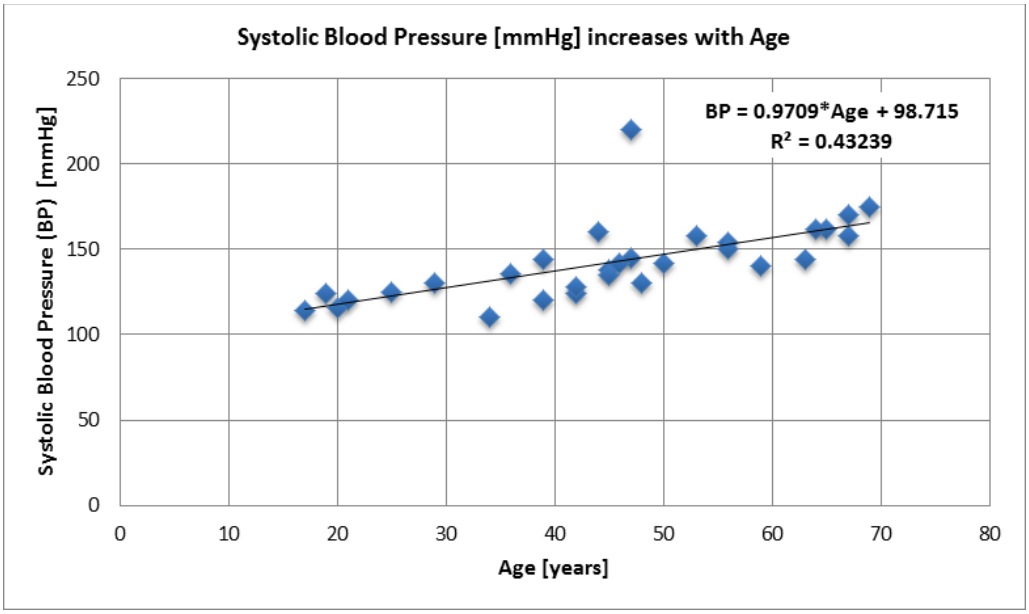
A. Which graph do you estimate will have the highest r^2 value?

B. Describe the observed relationship between long jump distance and cavities based only on the graph you selected in response to question A. What is the rationale for your answer?

Solution

Practice 8.

The following graph contains data and analysis from an experiment that examined the relationship between age and blood pressure. The blood pressure measurements were recorded using a standard off-the-shelf blood pressure monitor purchased at a local pharmacy. The participants were 30 individuals whose ages were 17 – 69. Based on this graph, answer the following questions:



Data source: http://people.sc.fsu.edu/~jburkardt/datasets/regression/x03.txt

A. You receive a memo from your company's Director of Human Resources saying, "We need an estimate, by the end of the day, of the systolic blood pressure of a 75-year-old person. This is critical to determining whether or not one of our employees must go on disability."

   Use your knowledge of the data and its analysis to respond to the Director's request.

B. Based on the data provided to you, what can you say about the accuracy of the blood pressure monitor used to measure the blood pressure? What is the rationale for your answer?

C. Read the following statements regarding the relationship between age and blood pressure shown on the graph. On the answer sheet, circle the letter for each statement that is TRUE.

   a. People in this study with a higher age tend to have a higher blood pressure.

   b. The coefficient of determination is positive.

   c. The coefficient of determination is approximately 0.16.

   d. The coefficient of determination is approximately 0.43.

   e. The coefficient of determination is approximately 0.65.

   f. There is not enough information to determine the coefficient of determination.

   g. The coefficient of determination is calculated to ignore the effect of outliers such as the systolic blood pressure of approximately 225 in a person approximately 47 years old.

Solution

Solutions

Practice Solution 1

A. Write the linear model, **y = ax + b,** for the given data. `y = 6x-6`

B. Fill in the values for $f(x_1)$ and $f(x_2)$ in the table below.

i	x_i (Time [s])	y_i (Velocity [m/s])	x_i^2	$x_i * y_i$	$f(x_i)$	e_i	e_i^2	$[y_i - \bar{y}]^2$
1	3	10	9	30	12	- 2	4	361
2	5	25	25	125	24	1	1	16
3	7	30	49	210		- 6	36	1
4	10	50	100	500		- 4	16	441
Sum	25	115	183	865		-11	57	819
Mean	6	29	46	216		-3	14	205

C. Write the line of code for line 7 that will calculate the predicted velocities for the given time values. You must use the variables defined in the MATLAB script above and an appropriate built-in MATLAB command.

```
velocity_predicted = polyval(output_1, time);
```

```
polyval(output_1, time)
```

D. Set up the equation necessary to determine the coefficient of determination, r^2. This means, fill in the numerical values for SSE and SST in the equation. **You DO NOT need to solve the equation.**

$$r^2 = 1 - \frac{SSE}{SST} = 1 - \frac{57}{819}$$

E. What is the meaning of the r^2 for this data set?

a. Error due to problems during the acquisition of raw data measurements

b. Error between the raw data points and the predicted velocity data

c. Variation between the raw data points and the mean of the predicted velocity data

d. Variation of the raw data points around the regression line

e. Can't tell from the data

Practice Solution 2

A. Show the equations needed to determine the coefficients for the regression line. Any known quantities should be substituted into equations as appropriate.

21a + 6b = 83.25

91a + 21b = 365.75

B. What percentage of the variance in the data is explained by the data? 1- (3.875/319.9688) = .99

Practice Solution 3

Line #	MATLAB Code
1	% ENGR 126 Spring 2009, Exam 2 MATLAB question
2	clear all % This command clears all variables and functions
3	clc % This command clears the command window
4	close all % This command deletes all figures
5	% ---- Input Section ----
6	load 'exam2_data.txt'; % Reads in data containing
7	resi = exam2_data(:,1); % creates vector of Resistivity in microohms-cm
8	tempe = exam2_data(:,2); % creates vector of temperature in degrees C
9	% ---- Calculation Section ----
10	coeffs = polyfit(resi,tempe,1);
11	pred_t = polyval(coeffs,resi);
12	
13	
14	
15	% ---- Output Section ----
16	% Generate a plot of the data suitable and regression line for technical presentation
17	plot(resi,tempe,'rs')
18	title('Temperature as a function of resistivity')
19	ylabel('Temperature (degrees C)')
20	xlabel('Resistivity (microohms-cm)')
21	hold on
22	plot(resi,pred_t,'k')
23	
24	
25	% Display the value of the slope predicted by the model
26	disp('The slope of the predicted line is:')
27	disp(coeffs(1))
28	or
29	disp('The slope of the predicted line is:')
30	coeffs(1)
31	or
	fprintf('The slope of the predicted line is: %4f.1 \n',coeffs(1))

Practice Solution 4

A. Calculate the coefficient of determination (r^2) for each of the two linear models above. Show your work in the space provided on the answer sheet.

Fall GPA-SAT: r2 = 0.17; Fall GPA-Spring GPA: r2 = 0.75.

B. An in-coming student from the cohort studied here earned a total SAT score of 1250. Using the two linear models above, predict this student's fall semester GPA and spring semester GPA. Please show your work in the space provided on the answer sheet.

Fall GPA: 0.002(1250) + 0.5 = 3.0 Spring GPA: 0.9(3.0) + 0.25 = 2.95

C. Briefly explain (1-2 sentences) which of these two models is more predictive of student performance, and provide specific justification.

The Fall-Spring linear regression has a higher r-squared value, therefore the Fall-Spring linear regression model accounts for more of the variance in the data and is of higher predictive value.

Practice Solution 5

A. Point 1: (2,10) m = -1

Point 2: (11, 1) b = 12

B.

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \qquad a\ \boxed{276} + b\ \boxed{36} = \boxed{132}$$

$$a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i \qquad a\ \boxed{36} + b\ \boxed{6} = \boxed{30}$$

C.

```
% Load x and y vectors
x = [2 3 5 6 9 11];
y = [10 5 7 4 3 1];
% calculate coefficients of linear model
coeff = polyfit(x,y,1);
% calculate fitted y
y_modeled = polyval(coeff, x);
% calculate SSE
SSE = sum((y - y_modeled).^2);
% Calculate SST
SST = sum((y - mean(y)).^2);
% Calculate R^2
r_squared = 1 - SSE/SST;
```

Practice Solution 6

A. 6 mm

B. 2 mm

C. Predicting beyond the range of data is risky / inappropriate. Number can be calculated as long as a relevant warning is provided

Practice Solution 7

A. Graph A

B. The variables are negatively correlated based on selecting Graph A

Practice Solution 8

A. Value cannot be predicted because it is out of the age range. A broader range of data must be collected or located

B. NO observations can be made about the accuracy of the measures with the blood pressure monitors; r2 measures the fit between the model and data, it does not account for measurement error

C. a, b, d