Math 10 – Homework 6

**Instructions**

* Type your answers and paste images directly into this document.
* You are free to use any regression/scatter-plot tool. Suggestion: <http://www.alcula.com/calculators/statistics/scatter-plot/>
* [You are also free to use any standard deviation calculator. Suggestion: http://www.calculator.net/standard-deviation-calculator.html](http://www.calculator.net/standard-deviation-calculator.html)
* Note : in the standard deviation calculator above, we want the “sample standard deviation” (scroll down on that page for formula)
* Print out and hand in homework in class on Tuesday.
* You may collaborate on the homework but you must write it up yourselves.

**Problem 1 – Regression** *( 24 points )*

We are not given what x and y really are here. But let’s assume that these are all meaningful dataset and not something silly like x = margarine consumption in Maine and y = divorce in the USA.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset 1 | Dataset 2 | Dataset 3 | Dataset 4 |
| x | y | x | y | x | y | x | y |
| 10 | 8.04 | 10.0 | 9.14 | 10.0 | 7.46 | 8.0 | 6.58 |
| 8.0 | 6.95 | 8.0 | 8.14 | 8.0 | 6.77 | 8.0 | 5.76 |
| 13.0 | 7.58 | 13.0 | 8.74 | 13.0 | 12.74 | 8.0 | 7.71 |
| 9.0 | 8.81 | 9.0 | 8.77 | 9.0 | 7.11 | 8.0 | 8.84 |
| 11.0 | 8.33 | 11.0 | 9.26 | 11.0 | 7.81 | 8.0 | 8.47 |
| 14.0 | 9.96 | 14.0 | 8.10 | 14.0 | 8.84 | 8.0 | 7.04 |
| 6.0 | 7.24 | 6.0 | 6.13 | 6.0 | 6.08 | 8.0 | 5.25 |
| 4.0 | 4.26 | 4.0 | 3.10 | 4.0 | 5.39 | 19.0 | 12.50 |
| 12.0 | 10.84 | 12.0 | 9.13 | 12.0 | 8.15 | 8.0 | 5.56 |
| 7.0 | 4.82 | 7.0 | 7.26 | 7.0 | 6.42 | 8.0 | 7.91 |
| 5.0 | 5.68 | 5.0 | 4.74 | 5.0 | 5.73 | 8.0 | 6.89 |

1. Compute the mean of x, and the mean of y for all four datasets to 2 decimal places. *( 4 points )*
2. Compute the (*sample*) standard deviation s of x, and the (*sample*) standard deviation s of y, for all four datasets to 2 decimal places. *( 4 points )*
3. For each of the four datasets: state the equation of the regression line to 2 decimal places (*please round to the nearest 2 decimal places*). Then, plot both the line and the scatter plot together in one diagram and include the diagrams here. *( 4 points )*
4. Consider your regression line and scatter plot for dataset 3. The point (x,y) = (13.0,12.74) is what we call an outlier.

If we remove this outlier, would the slope of the regression line become steeper (*higher slope coefficient, line becomes more vertical*) or less steep (*lower slope coefficient*, *line becomes more horizontal*)? Do not do any calculations. *( 2 points )*

1. Using your regression line in dataset 1, predict what the y value would be for x = 16. *( 2 points )*

For the following questions, consider your regression plots for each of the datasets.

Recall that linear regression produces a line that is meant to be a mathematical model for the data. This model can also be used to try and predict what the Y value will be, for a given X value.

A key assumption of linear regression is that there is a true linear relationship between X and Y. Deviations from this assumption can be checked by plotting the residual errors, which should all come from normal distributions with the same variance and a mean of zero. However, for this question, you can simply look at the scatter plots.

For each of the datasets 1-4, look at the data, regression line and scatter plot. Then, based only on these information, comment on whether linear regression is a good model for the data, based on the explanation above. We don’t really know what the x,y really are, so this is your best guess. Explain your answer.

*Hint: your answers should include one clear “Yes”, two clear “No” and an uncertain “Yes”. In the one uncertain case, your mission is to give a reason for your uncertainty, and why linear regression could still be a good model despite this reason, or what we can do to fix the problem (we are very lenient with grading this case).*

1. Dataset 1 *( 2 points )*
2. Dataset 2 *( 2 points )*
3. Dataset 3 *( 2 points )*
4. Dataset 4 *( 2 points )*