Bivariate Statistics Exploration

Jacob Bruner

September 26, 2022

1 Introduction

In the real-world, statistics are often employed as objective measures of data. Because of this, we often draw important, life-changing decisions on their basis alone. For instance, pharmacists employ statistics when researching new treatments for disease, or economists might draw on statistics to understand trends in consumer spending. Luckily for us, statistics often *are* helpful and insightful in disseminating patterns in data. But, again, it is all too easy to forget their short-comings.

In 1973, British mathematician Francis Anscombe formulated four datasets, each with 11 points, in order to demonstrate the pitfalls of many common statistical measures. In this paper, I will explore the ways in which these statistical measures can be misleading in interpreting the significance of, trends in, or validity of data using Anscombe's 'quartet' of points to highlight their oversimplification.

2 The Statistical Mean

2.1 Computation

The mean is a commonly used measure of center for a given statistic. To calculate the mean, we can use the following formula: for a given indexed set of values, the mean is written:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Where N is the length of the data set. In other words, the mean is the sum of each value divided by the number of values for a given parameter.

Calculating these values for each dataset we obtain:

Data set	**I**		**II**		**III**		**IV**	
Parameter	x	y	x	y	x	y	x	y
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5

Figure 1: Computed Means of Dataset I-IV

2.2 Interpretation

After computing the mean for each, it's clear that the means for each x, y of the respective datasets are equal. Without any other prior knowledge, one might infer that these datasets have a similar center, save variation or spread. This is because the mean doesn't provide information about the variation or spread of the data. Despite this, however, one might assume it to be a reasonable assumption that these datasets and parameters might be highly similar, especially because they match in two variables.

	$\bar{x}$	$\bar{y}$	σ_x^2	σ_y^2	σ_x	σ_y	r^2	$cov(x, y)$
set 1	9.0	7.5	11.0	4.13	3.32	2.03	0.67	5.5
set 2	9.0	7.5	11.0	4.13	3.32	2.03	0.67	5.5
set 3	9.0	7.5	11.0	4.12	3.32	2.03	0.67	5.5
set 4	9.0	7.5	11.0	4.12	3.32	2.03	0.67	5.5

Figure 2: Comparison of different Statistical Measures of Datasets I-IV

3 Exploring Other Statistical Measures

3.1 Computation

In addition to computing the mean, there are a number of other tools that provide insight into the properties of a dataset. For instance, the ***variance*** (σ^2) and ***standard deviation*** (σ) convey a measure of the 'spread' or variation in a dataset. Similarly, ***'Peterson's Correlation Coefficient'*** (r) often denoted with its square, r^2, is measure of the *linear* coorelation of two parameters. (Unsquared) values of r range from -1 to 1, with numbers farther from zero denoting stronger coorelation. A closely related notion is the ***covariance***, which measures the linear coorelation, but with a magnitude not necessarily between $(-1, 1)$. These measures are related like so: $r = \frac{cov(x,y)}{\sigma_x \sigma_y}$ where σ denotes the standard deviation.

3.2 Plotting Data

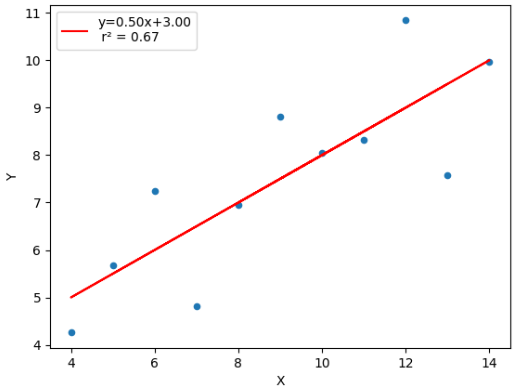
Using the python packages `matplotlib`, `scipy`, `numpy` and `pandas`, we can plot each dataset (as a dataframe contained in the array `dfs`). A linear regression is fit to the data using the `np.polyfit()` method with an exponent of 1, and the Pearson's coefficent is calculated using the `stats.pearsonr()` method. This is accomplished like so:

```
for i in range(4):
  dfs[i].plot.scatter(x='x', y='y')
  reg = np.polyfit(dfs[i]['x'], dfs[i]['y'], 1)
  rval = stats.pearsonr(dfs[i]['x'], dfs[i]['y'])[0]
  plt.plot(dfs[i]['x'], np.polyval(reg, dfs[i]['x']), 'r', label='y={:.2f}x+{:.2f}\n r^2 =
      {:.2f}'.format(reg[0], reg[1], rval**2))
  plt.xlabel('x')
  plt.ylabel('y')
  plt.legend()
  plt.savefig('dataset{}.png'.format(i+1))
```

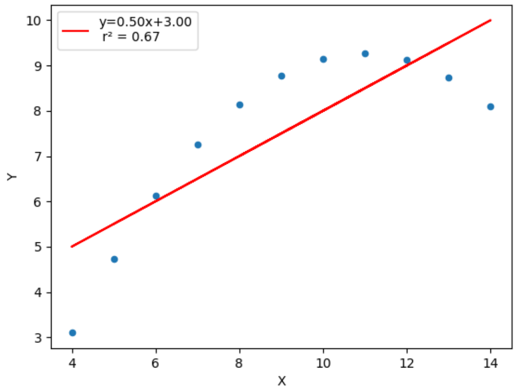
In so doing, we get the following interesting results (Figure 3):

3.3 Interpretation

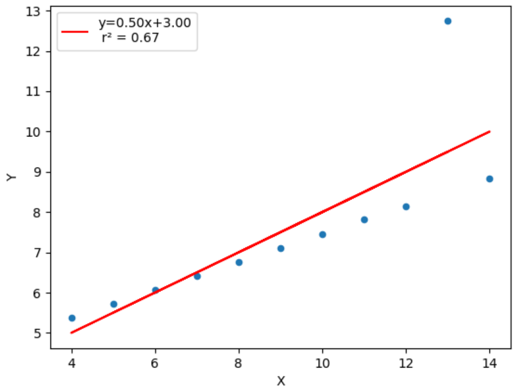
Upon examination of the computed statistical measures (std, variance, cov, etc.), the naive conclusion would be that each dataset contains extremely similar data and that similar conclusions could be drawn from each. For instance, the r^2 value for each, 0.67, indicates that there is a weak-to-moderate, positive correlation between the variables, and might imply that linear fits to each dataset would have a similar degree of inaccuracy. However, this notion is *categorically refuted* on even a cursory glance at the graphs. In Figure 3, we see that each dataset is very distinct in its scatterplot pattern. In Figure 3a, its clear that the linear fit to the data indicates weak-moderate, positive correlation, with a healthy amount of deviation from the trendline. This graph's, the most straightforward of the four, statistical measures indicate helpful information about the data. For instance, the r^2 value of 0.67 reinforces the weak, positive spread outlined above. Similarly, the covariance of 5.5 and the variance $\sigma_x^2 = 11.0$, $\sigma_y^2 = 4.13$ reinforce the visual intuition of this dataset having a healthy amount of spread from the means $\bar{x}, \bar{y}$. As we transition to Figure 3b, we see a markedly different trend in the data. Visually, the scatterplot indicates an quadratic trend in x, indicating that the y values may vary as the negative square of x. Despite this, our linear trendlines (from a least squares regression) provides a line of 'best fit' equal to Figure 3a. Additionally, the other statistical measures of dataset II are exactly equivalent to that of dataset I, despite the visual



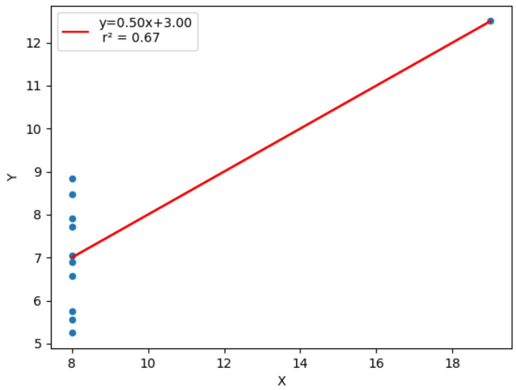
(a) Dataset I



(b) Dataset II



(c) Dataset III



(d) Dataset IV

Figure 3: Plot of Datasets I-IV — X against Y with Lin. Reg.

interpretation being different. For instance, one might interpret the r^2 value in Figure 3a as indicating the moderate (seemingly random) variation from the trendline, whereas in Figure 3b, we see the datapoints follow exactly the (not random) curve of a quadratic, despite still having the same r^2. If one were to perform a polynomial regression (perhaps taking the logarithm to linearize the polynomial exponent, then performing a least squares regression), the resulting trendcurve would likely be an extremely close fit to the data—a fact not illustrated by these statistical measures. In Figure 3c, we see a different outcome where, despite displaying a very linear trend, the computed statistics and trendline are greatly affected by a single outlier. Because of this, we see the trendline having the incorrect slope to match the (linear) trend in the rest of the data. Despite this outlier not being indicated in the computed statistical measures, viewing the graph demonstrates the highly linear relationship between x and y, which would be easily illustrated by removing the outlier or otherwise explaining it in the methodology. This fact highlights the difference between dataset III and dataset I, where, despite having the same statistical measures, Figure 3a displays almost uniformly random deviation from the line of best fit compared to Figure 3c which has a nicely behaved deviation given by the intersection of another trendline if the sole outlier were to be removed. In Figure 3d, dataset IV, we see how the statistical measures fail to indicate that most of the measured independent 'y' values coorespond to a single dependent 'x' value, namely 8. Notwithstanding this, we can still obtain a 'regression line' of sorts, which attempts to interpret this data as being linear. Clearly, from the scatterplot, this regression is wishful thinking *at best*, and complete nonsense at worst. Despite the other datasets indicating some relationship/coorelation between x and y, the grouping of ys at $x = 8$ indicates no such obvious relationship. Despite this, the numerical methods of a linear regression can still be performed, wrongfully indicating much similarity between dataset IV and datasets I-III.

4 Additional Areas to Explore

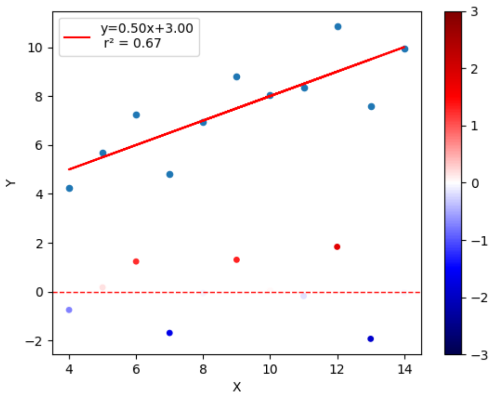
4.1 Residual Plots

Other tools are available in statistics to help better draw conclusions about unintuitive datasets like Ascombe’s Quartet. One such method is the use of residuals when plotting a regression. By slightly modifing the code above, one can run:

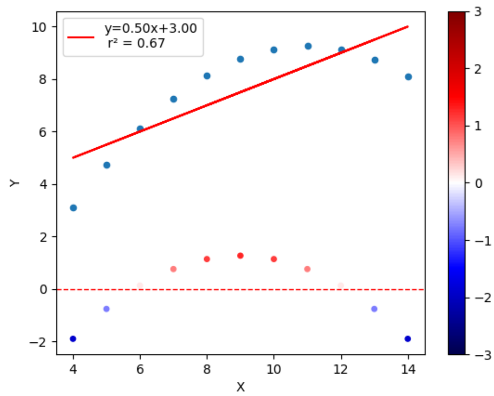
```
for i in range(4):
  # same code from before

  # add points corresponding to difference from reg with color given by magnitude
  plt.scatter(dfs[i][’x’], dfs[i][’y’] - np.polyval(reg, dfs[i][’x’]), c=(dfs[i][’y’] -
      np.polyval(reg, dfs[i][’x’])), cmap=’seismic’, vmin=-3, vmax=3,s=10)
  # add red dashed line at 0
  plt.axhline(y=0, color=’r’, linestyle=’--’, linewidth=1)
  # add colorbar
  plt.colorbar()
```

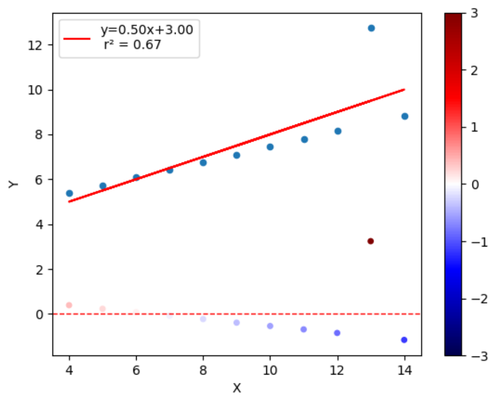
Obtaining:



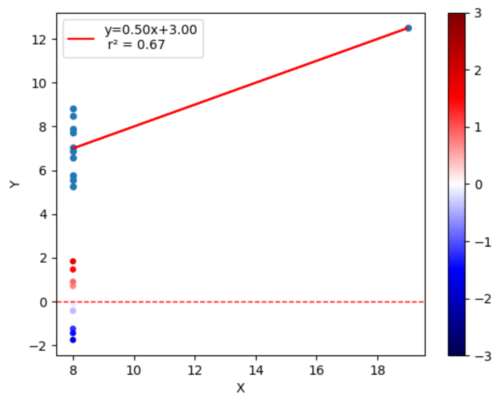
(a) Dataset I



(b) Dataset II



(c) Dataset III



(d) Dataset IV

Figure 4: Residuals Plot of Datasets I-IV — X against Y with Lin. Reg.

After viewing the residual plots, it’s clear that scrutinizing the pattern of residuals could be an insightful method at determining the reliability of a regression. For instance, one could compare the residuals of datasets II and II with that of dataset I to determine that the computed regressions aren’t appropriate to datasets II and III because there are clear patterns in the residuals. In Figure 4a, there seems to be a healthy randomness in the distribution of residuals, corresponding to the natural variation in the parameters from the trendline. Whereas in Figures 4b-4c, there is a quadratic and linear pattern respectively in the residuals. Hence, one could infer that there would be a better fit to each dataset by considering other factors (a non-linear regression and outlier influence here respectively, for instance). Similarly the residuals in Figure 4d confirms that this regression line is a poor model for the data, since almost every datapoint occurs at $x = 8$, creating a vertical column of residuals.

4.2 Arguments for Graphing Before Interpretation

To expand this investigation, another important question is whether one must graph data before analyzing it. From the results of this exploration, one could argue that graphs provide visual insight that cannot be captured within measures like a mean, standard deviation, or r^2 value. We see that all four datasets evaluate to the same statistical values in Table **??** despite having markedly different graphs. For this reason, for datasets where graphing is an option, one would conclude that graphing should absolutely be employed before interpreting statistics of a dataset. However, in certain applications where graphing isn't as easy (for instance multi-variate data), it might be out of necessity that statistical measures are interpreted without visual aid. However, as evidenced by this exploration, this may lead to misinterpretation of the results.

5 Conclusion