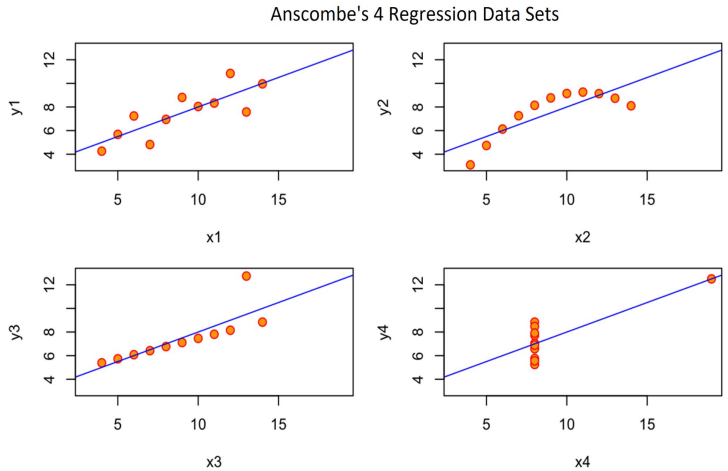
Anscombe's Quartet Research

Axyl Carefoot-Schulz

Supervisors: Dr. La Haye & Dr. Zizler

What is Anscombe’s Quartet?



[2]

- F. J. Anscombe was an English mathematician famous for computerizing statistical analysis [1]
- All data sets share regression line, correlation coefficient and summary statistics
- Currently unknown how he generated these data sets

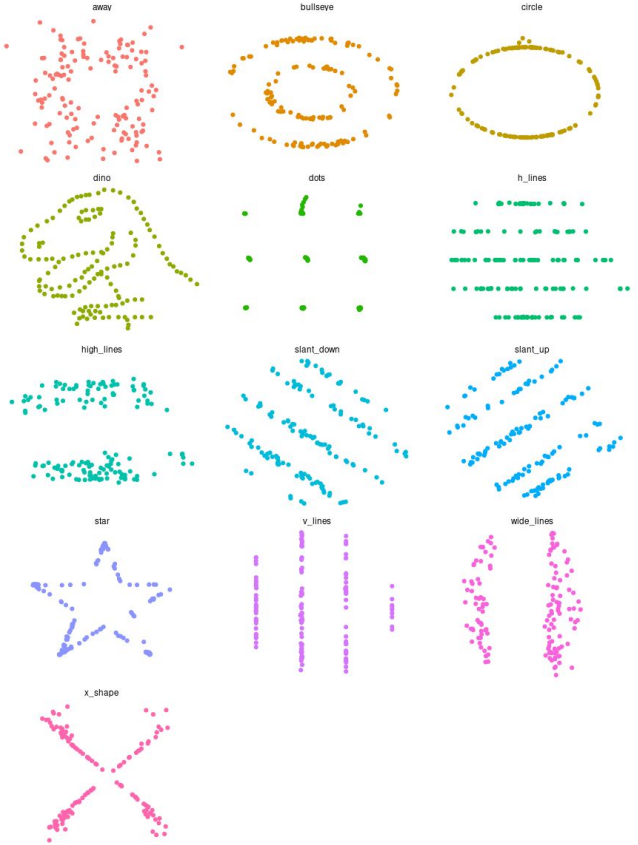
[2]

Anscombe's Quartet (and my own graphs) share:

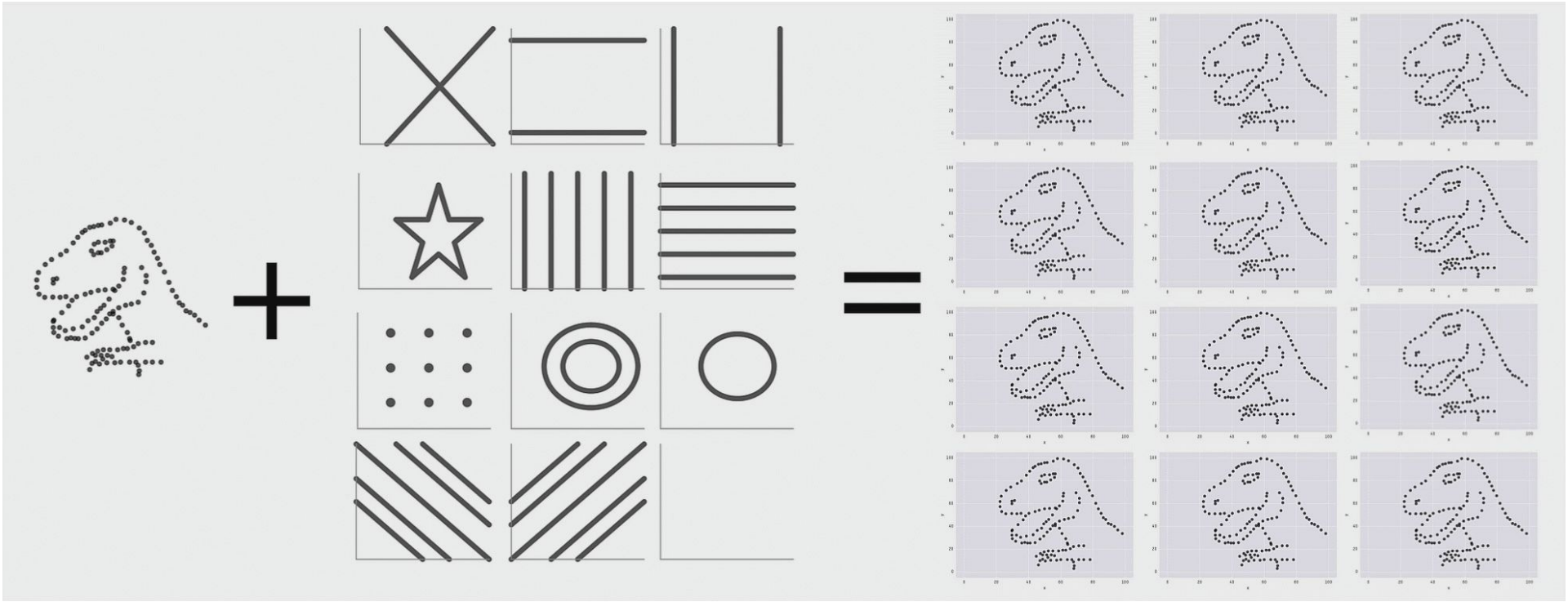
sample size (*11*) sample x-values (*4 .. 14*) y-variance (33/8) mean of y (*7.5*)	x-variance: (*11)* mean of x (*9*) slope of regression line (*½)* intercept of the regression line (*3*) correlation coefficient (*√⅔ ≈ 0.82*)

Other Solutions

- Uses a computing intensive process called “simulated annealing”
- Takes in a set of data points and is able to output data sets with the same best-fit line
- Similar problem being solved, differing methods



[3]



[3]

Solving using Statistics, Linear Algebra and Computer Science

Statistics

$$\bar{y} = b_1\bar{x} + b_0$$

$$s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x^2} = r\frac{s_y}{s_x}$$

Linear Algebra [4]

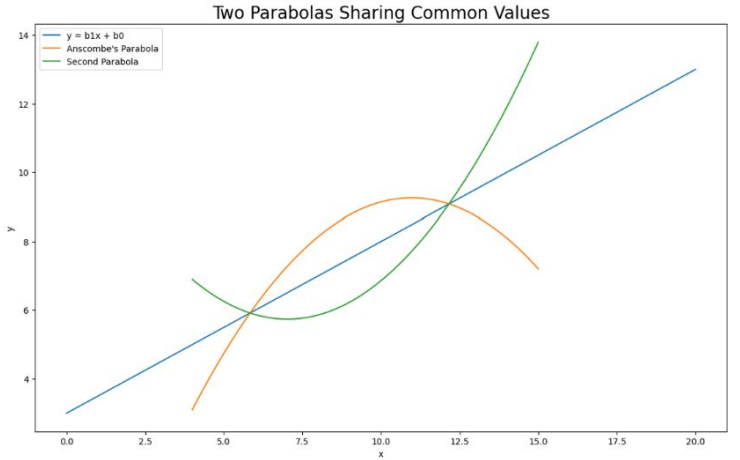
$$<\vec{y}, \vec{1}> = (n-1)\bar{y}$$

$$<\vec{y} - \bar{y}\vec{1}, \vec{y} - \bar{y}\vec{1}> = (n-1)S_y^2$$

$$<\vec{x} - \bar{x}\vec{1}, \vec{y} - \bar{y}\vec{1}> = b_1(n-1)s_x^2$$

$$y = \begin{bmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ x_3^2 & x_3 & 1 \\ & \dots & \\ x_n^2 & x_n & 1 \end{bmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

Computer Science



Quadratic Details

```
x: 4, y: 3.10
x: 5, y: 4.74
x: 6, y: 6.13
x: 7, y: 7.26
x: 8, y: 8.14
x: 9, y: 8.77
x: 10, y: 9.14
x: 11, y: 9.26
x: 12, y: 9.13
x: 13, y: 8.74
x: 14, y: 8.10
Set 1 Regression Line: y = 0.50x + 3.00
Set 1 Correlation Coefficient: 0.82
```

```
x: 4, y: 6.90
x: 5, y: 6.26
x: 6, y: 5.87
x: 7, y: 5.74
x: 8, y: 5.86
x: 9, y: 6.23
x: 10, y: 6.86
x: 11, y: 7.74
x: 12, y: 8.87
x: 13, y: 10.26
x: 14, y: 11.90
Set 2 Regression Line: y = 0.50x + 3.00
Set 2 Correlation Coefficient: 0.82
```

Computing the Quadratic

Quadratic:

- $f(x) = x^2$
- $g(x) = x$
- $h(x) = 1$

Final Function:

- $af(x) + bg(x) + ch(x)$

Target Parameters:

- Regression line slope: ½
- Regression line intercept: 3
- Correlation Coefficient: 0.82

Results from Algorithm:

- a = -0.13 & 0.13
- b = 2.77 & -1.77
- c = -5.99 & 11.99

Generalization of Matrix Method

$$\mathbf{y} = \begin{bmatrix} f(x_1) & g(x_1) & h(x_1) \\ f(x_2) & g(x_2) & h(x_2) \\ f(x_3) & g(x_3) & h(x_3) \\ \vdots & \vdots & \vdots \\ f(x_n) & g(x_n) & h(x_n) \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

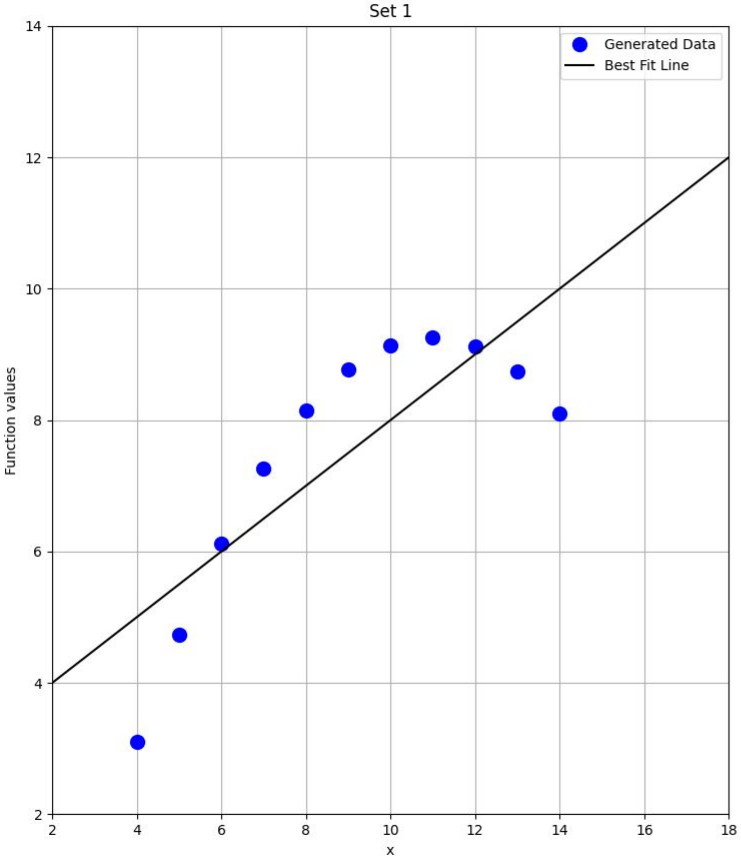
The Quadratic Function

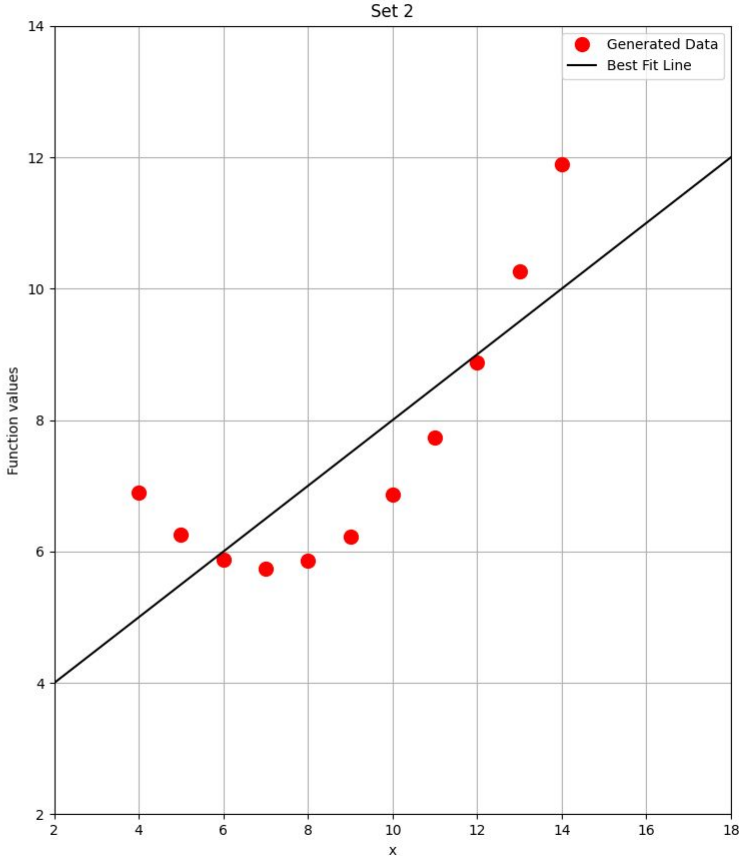
$f(x) = x^2$
$g(x) = x$
$h(x) = 1$

$f(x) = x^2$

$g(x) = x$

$h(x) = 1$

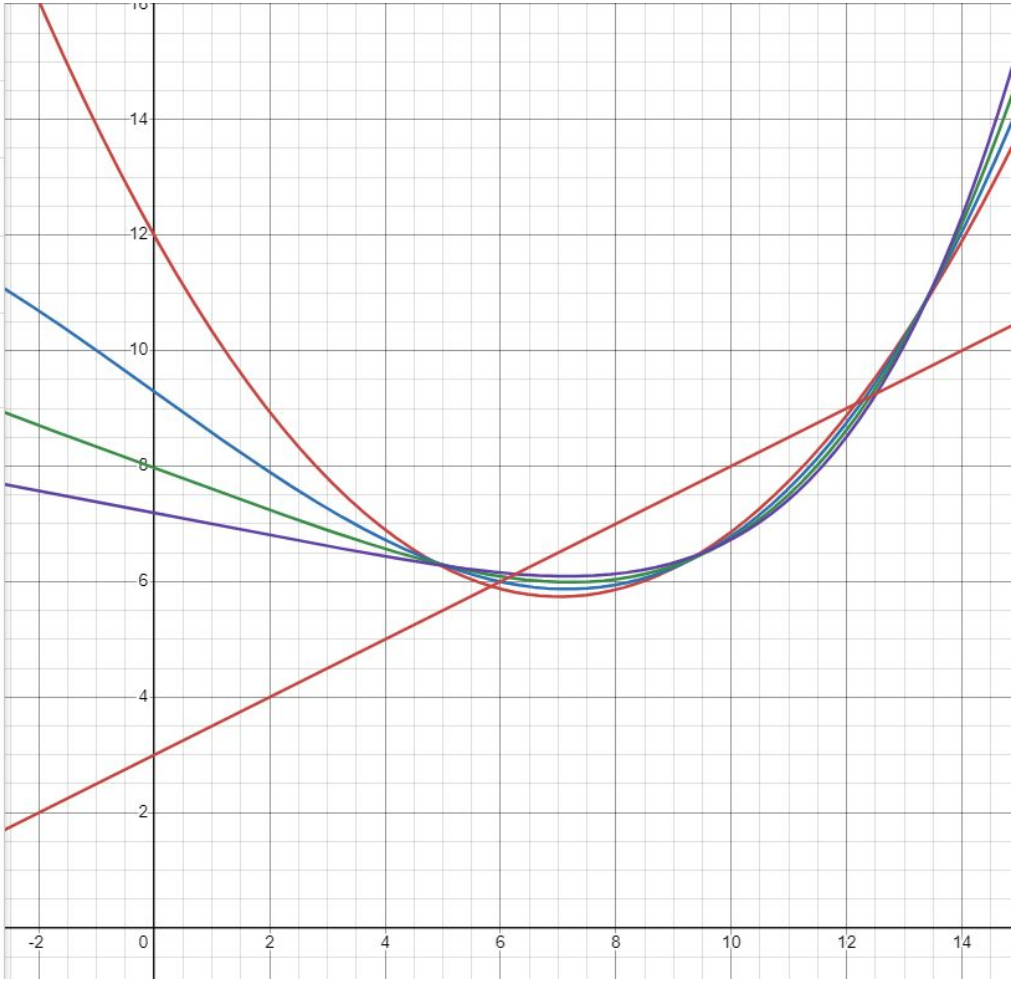

Set 1
Generated Data
Best Fit Line
Function values
x
14
12
10
8
6
4
2
2
4
6
8
10
12
14
16
18


Set 2
Generated Data
Best Fit Line
Function values
x
14
12
10
8
6
4
2
2
4
6
8
10
12
14
16
18

Other Graphs created using the same method

Multiple Polynomial Functions

1. $0.126592420885454x^2 - 1.77866357593815x + 11.9880618828672$
2. $0.00466562466909788x^3 - 0.716794913700729x + 9.29019517887777$
3. $0.000243401869660717x^4 - 0.365731770009239x + 7.96836764388849$
4. $1.45201707031063 \cdot 10^{-5}x^5 - 0.191102044785046x + 7.18768983162936$
5. $\frac{1}{2}x + 3$
6.



Piecewise Functions

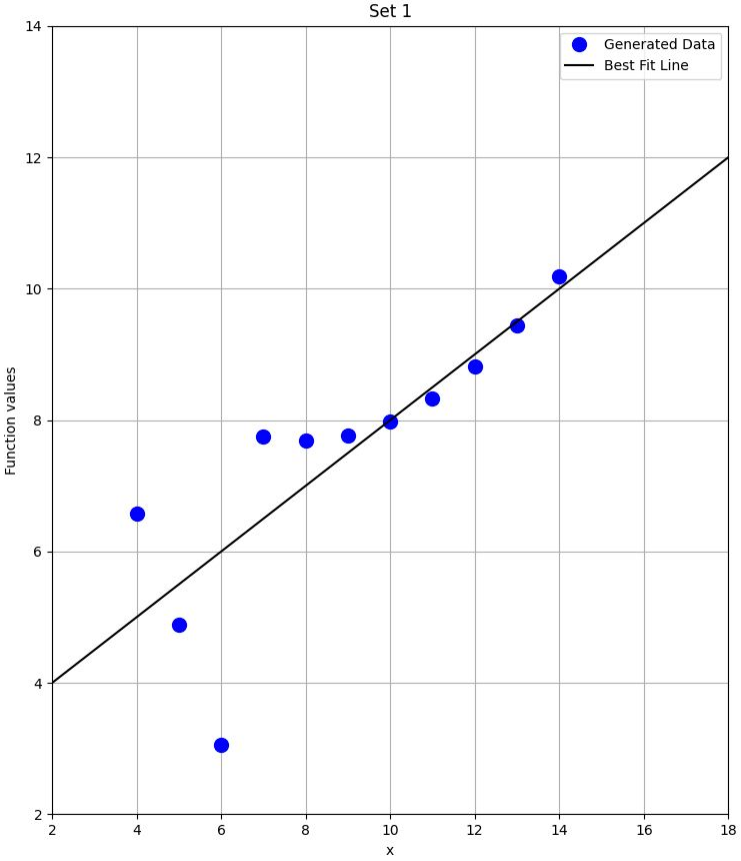
$f(x) = x^2$ if $x < 7$
$= -(x^2)$ if $x >= 7$
$g(x) = x$
$h(x) = 1$

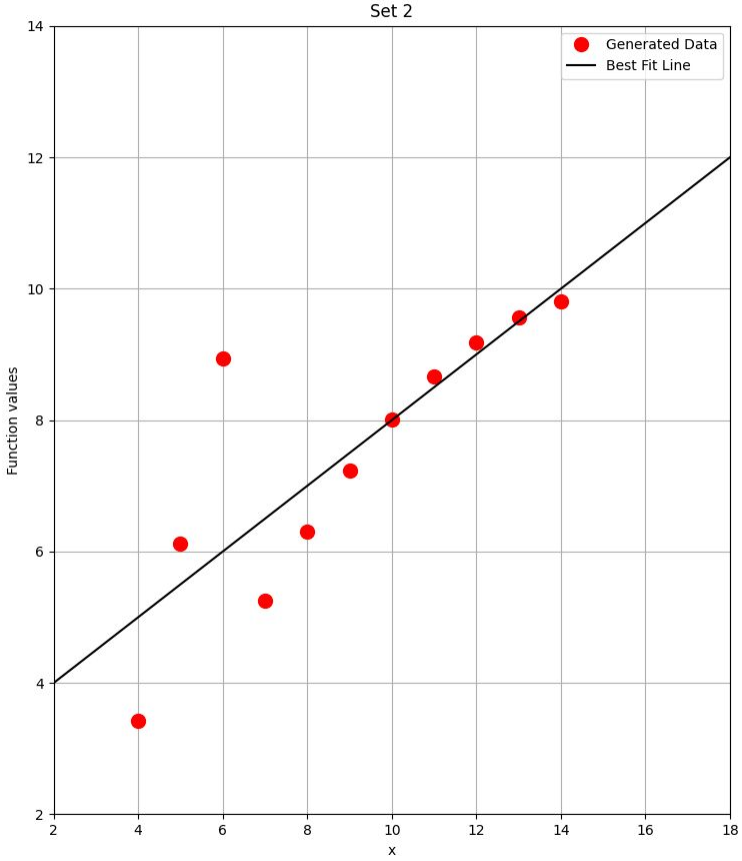
$f(x) = x^2$ if $x < 7$

$= -(x^2)$ if $x >= 7$

$g(x) = x$

$h(x) = 1$





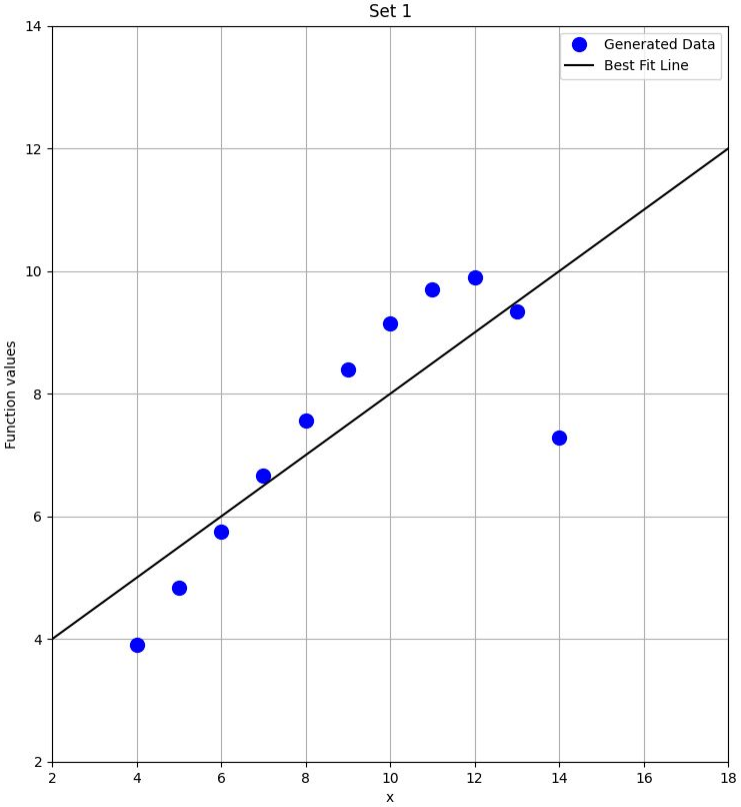
Exponential Function

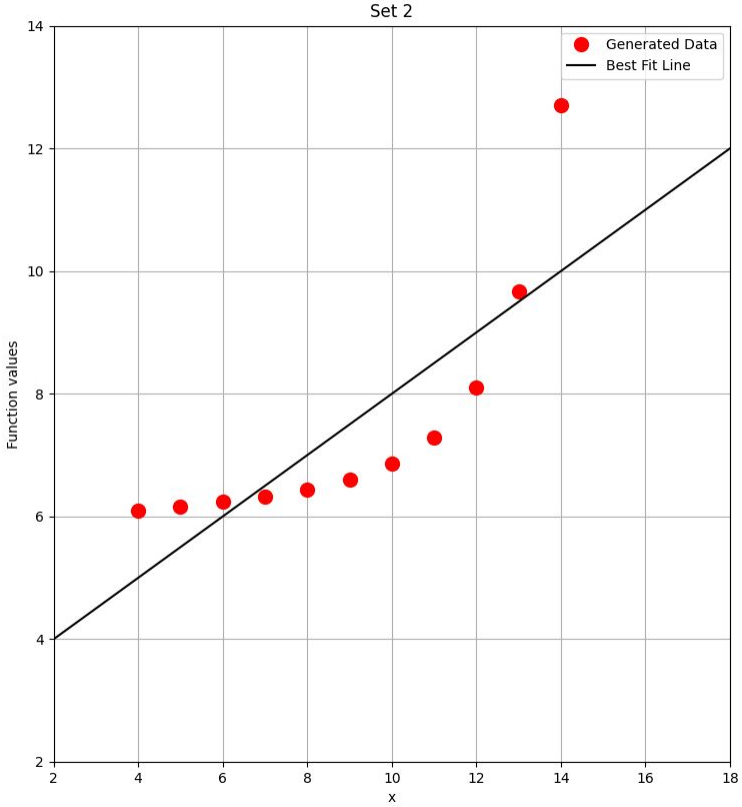
$f(x) = 2^x$
$g(x) = x$
$h(x) = 1$

$f(x) = 2^x$

$g(x) = x$

$h(x) = 1$

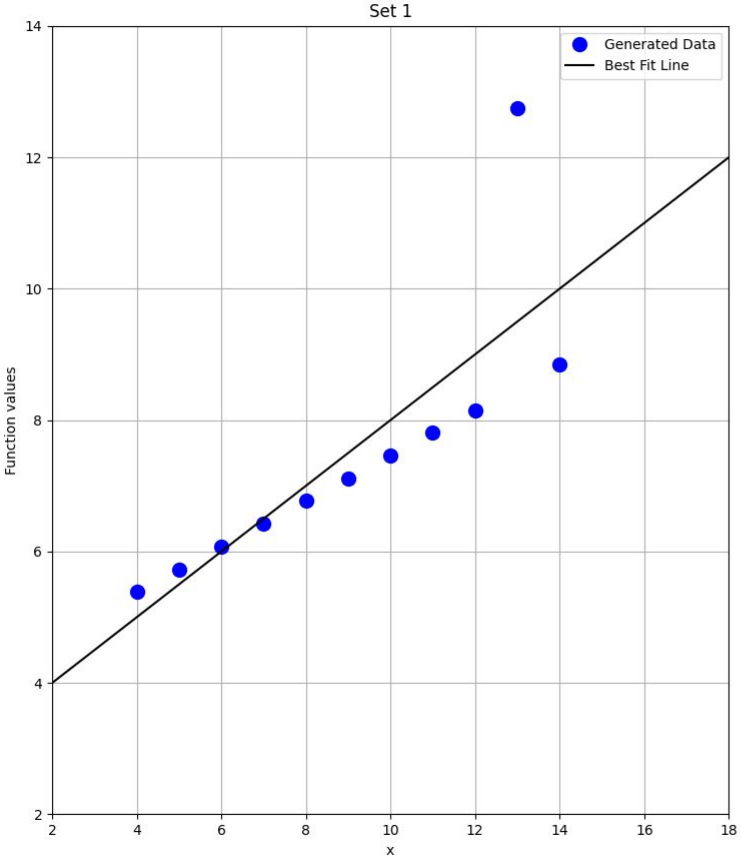

Set 1
Generated Data
Best Fit Line
Function values
x
14
12
10
8
6
4
2
2
4
6
8
10
12
14
16
18

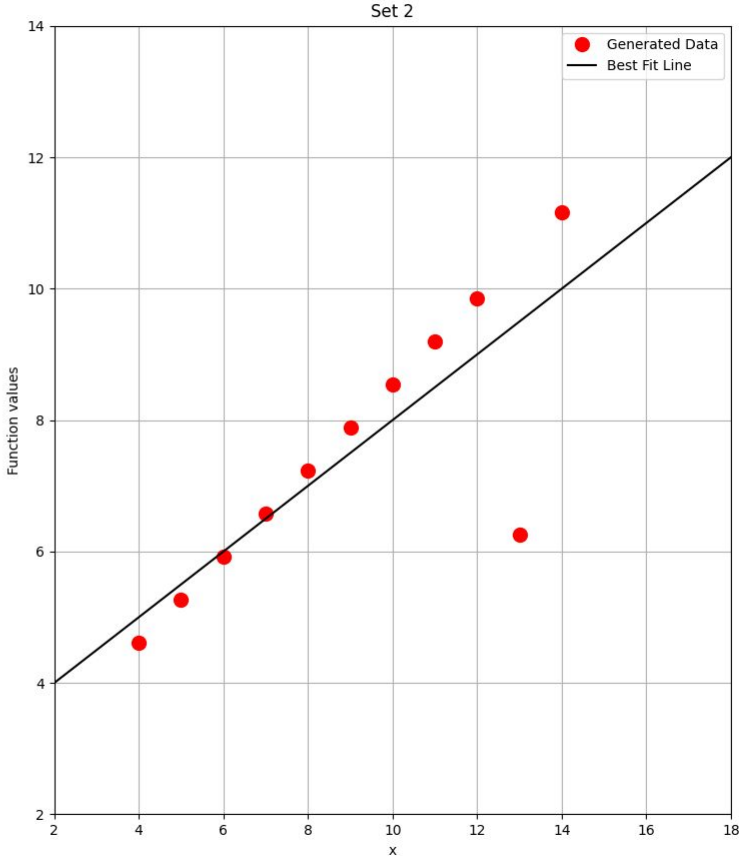

Set 2
Generated Data
Best Fit Line
Function values
x
14
12
10
8
6
4
2
2
4
6
8
10
12
14
16
18

Linear with One Point Off

$f(x) = x$
$g(x) = 1$
$h(x) = 1$ if $x = 13$, otherwise 0

f(x) = x, g(x) = 1, h(x) = 1 if x = 13, otherwise 0





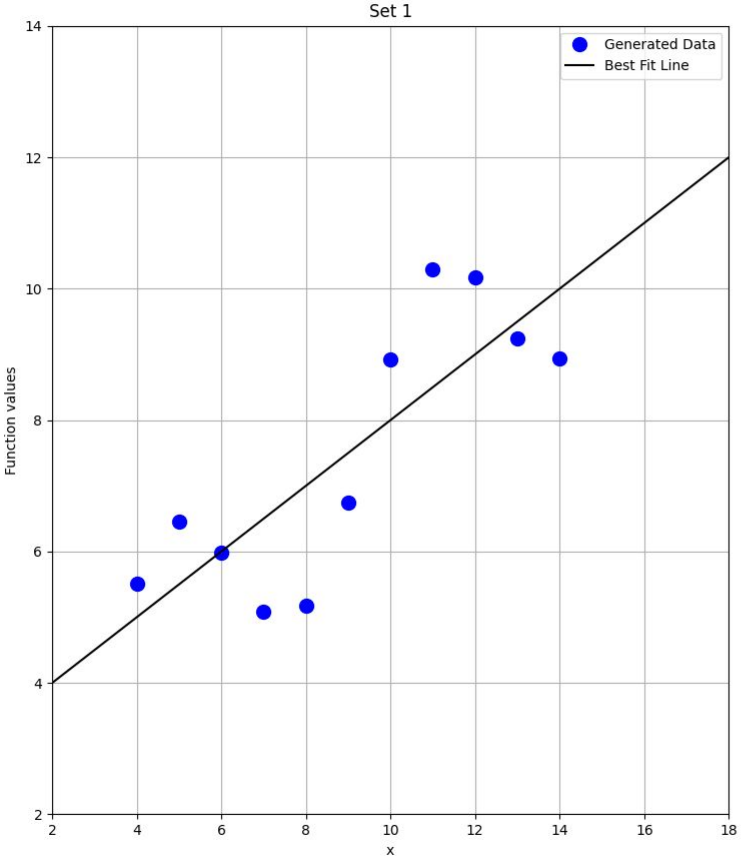
Sine Wave

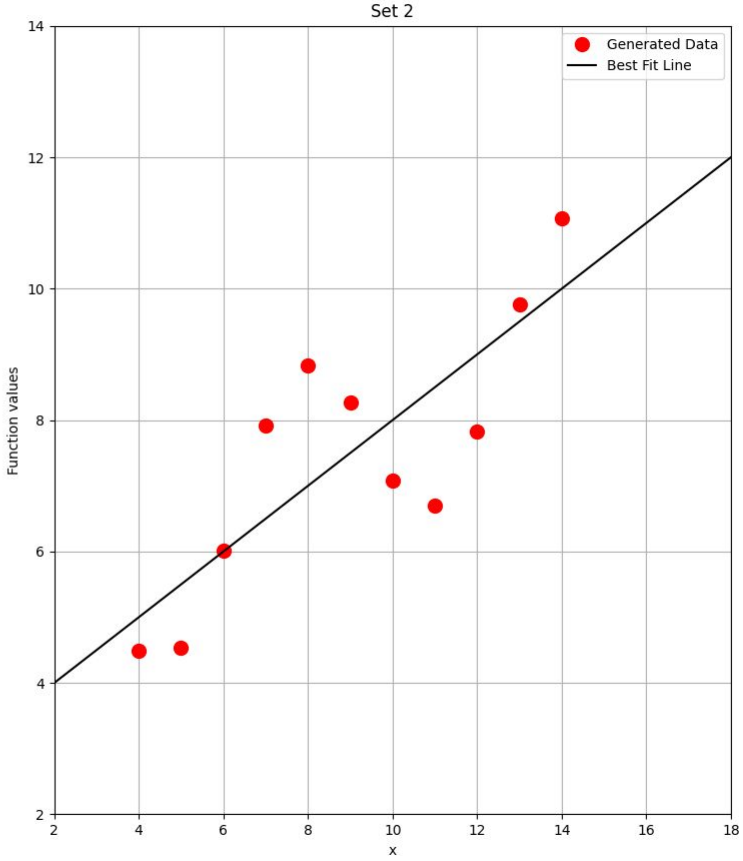
$f(x) = \sin(x)$
$g(x) = x$
$h(x) = 1$

f(x) = sin(x)

g(x) = x

h(x) = 1





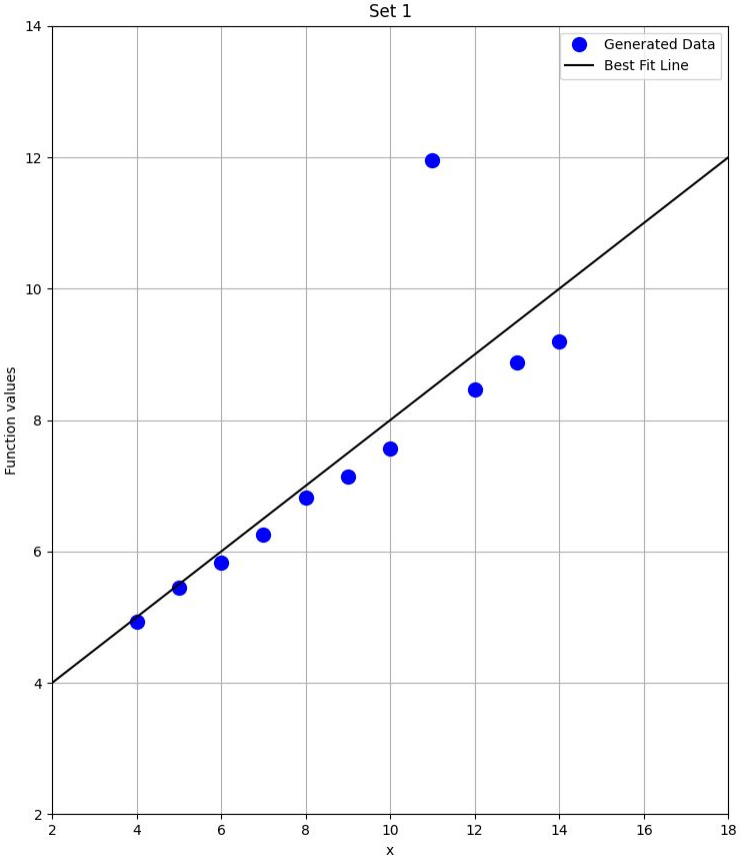
Odd Case: tan

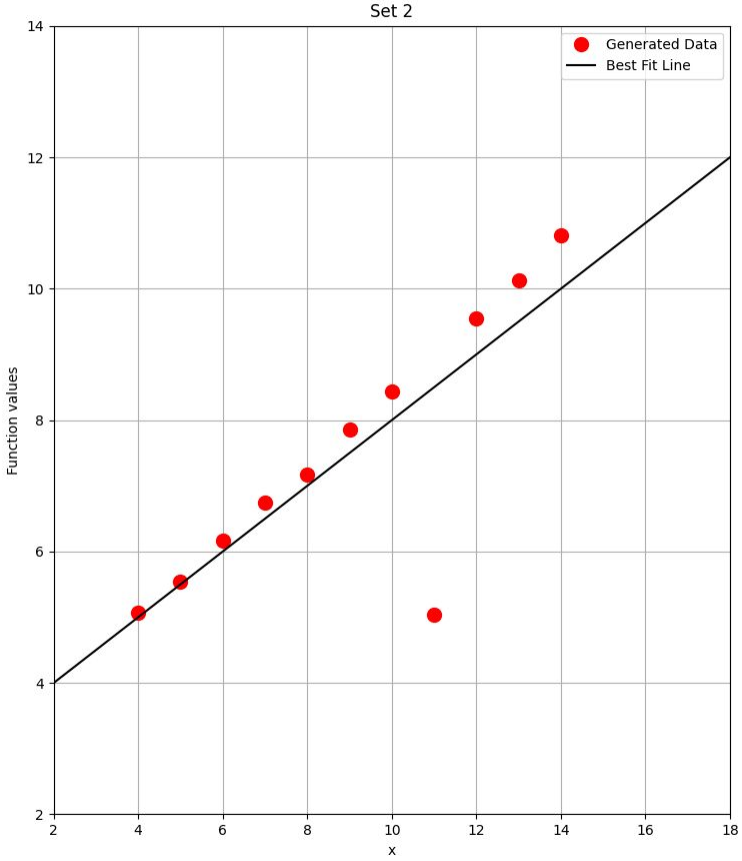
$f(x) = \tan(x)$
$g(x) = x$
$h(x) = 1$

f(x) = tan(x)

g(x) = x

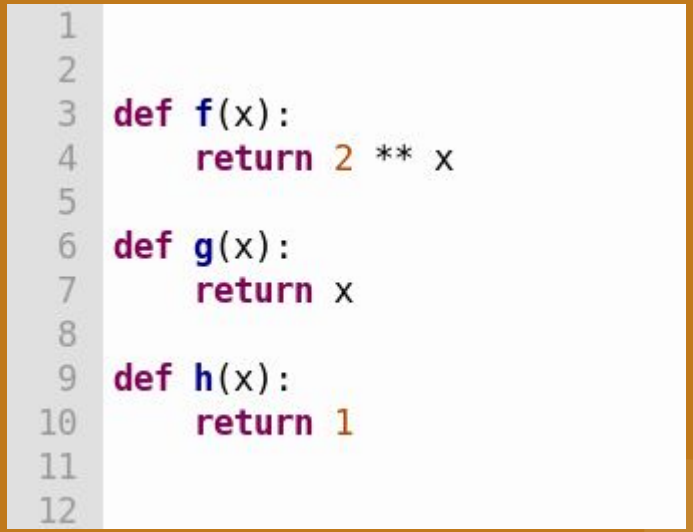
h(x) = 1





Future:

- Currently in Python
- Application will solve based on the three functions defined
- Use in classrooms for statistics and introductory programming



```
1
2
3  def f(x):
4      return 2 ** x
5
6  def g(x):
7      return x
8
9  def h(x):
10     return 1
11
12
```

Roberta La Haye: Using linear algebra to construct Anscombe's quartet

Abstract: Statistician F.J. Anscombe published his now famous quartet of data sets in a 1973 paper. The four data sets share many common summary statistics but have vastly different scatter plots. Anscombe's quartet provides a convincing argument of the importance of visualizing data. It is a mystery how he came up with the data sets.

There has been a recent flurry of computer science related research on generating Anscombe like data sets. Dr. Zizler and I considered the problem from a linear algebra perspective. We believe we have figured out exactly how Anscombe created the four data sets. The method can be understood in terms of first year linear algebra or first year statistics.

Saturday @ 11:05am

Thank you!

Supervisors:
Dr. Roberta La Haye & Dr. Peter Zizler

Sources:

[1] “Credibly curious,” Tidyverse Case Study: Anscombe’s quartet | Credibly Curious, https://www.njtierney.com/post/2020/06/01/tidy-anscombe/ (accessed Mar. 13, 2024).

[2] W. Saxon, “Francis John Anscombe, 83, mathematician and professor,” The New York Times, https://www.nytimes.com/2001/10/25/nyregion/francis-john-anscombe-83-mathematician-and-professor.html (accessed May 1, 2024).

[3] G. Fitzmaurice, “Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing,” Autodesk Research, https://www.research.autodesk.com/publications/same-stats-different-graphs/ (accessed May 1, 2024).

[4] Linear Algebra to Construct Anscombe's Quartet. R. La Haye and P. Zizler, College Journal of Mathematics