Introduction to Machine Learning
[Fall 2022]

Linear Regression (Part 2)

September 15, 2022

Lerrel Pinto

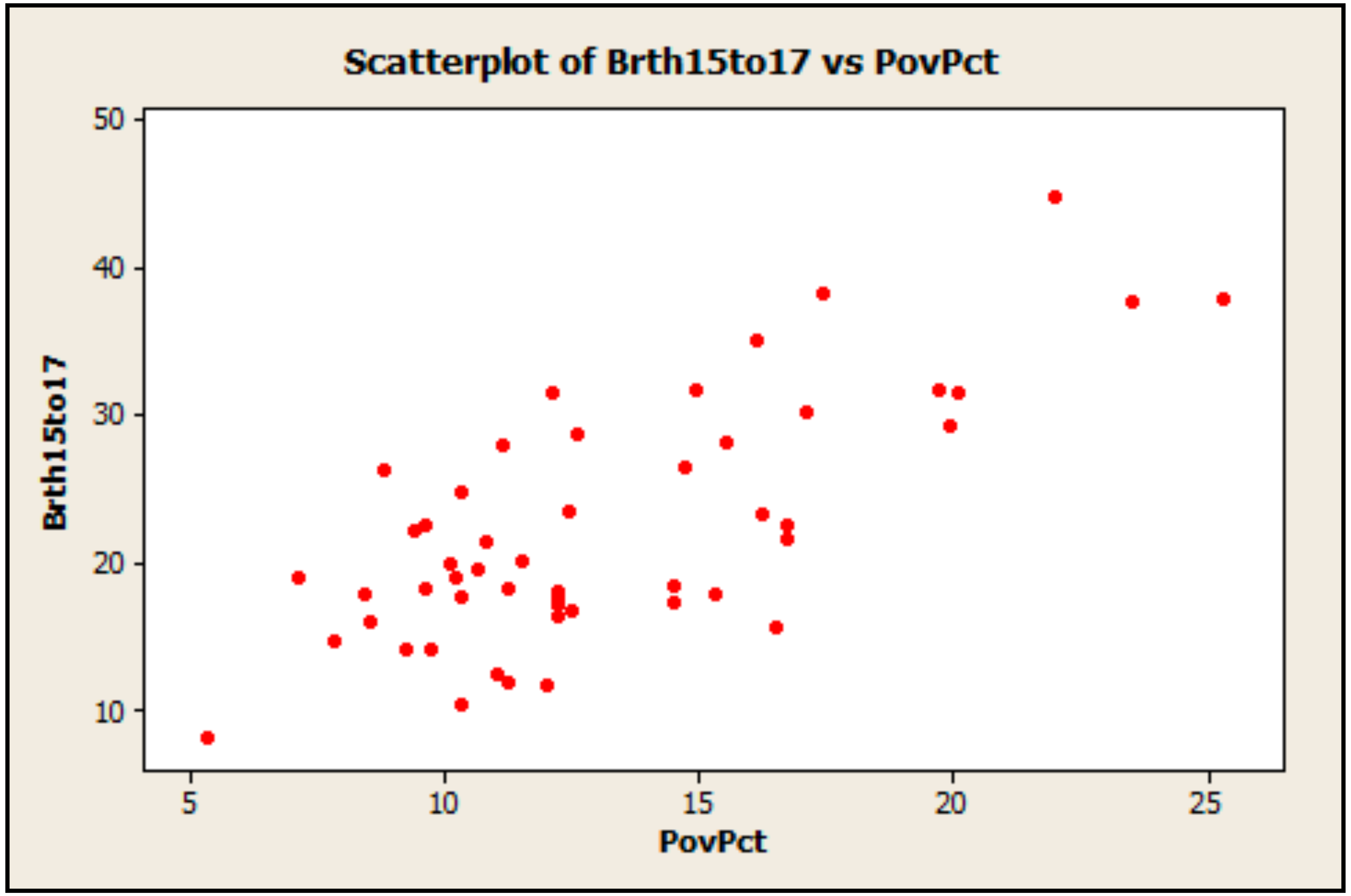
Topics for today

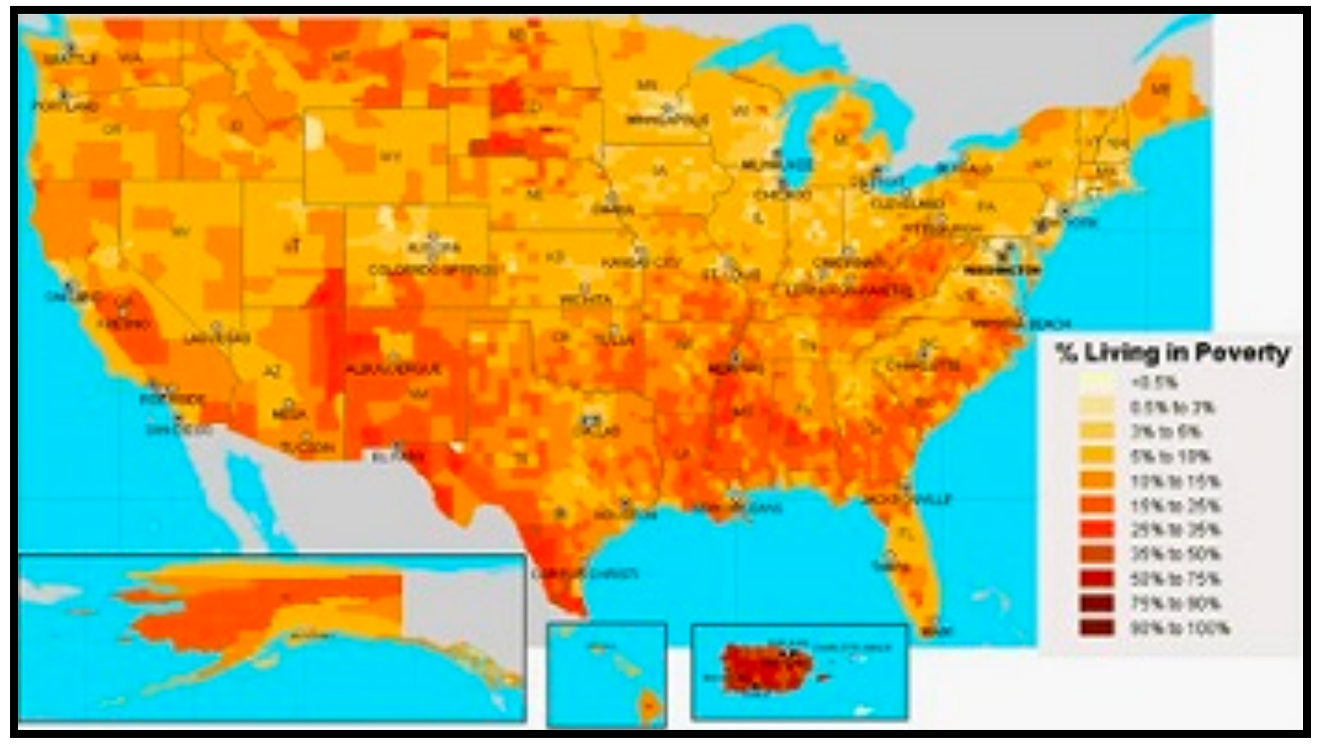
- More advanced linear regression.

Linear Regression (example)

https://online.stat.psu.edu/stat462/node/101/

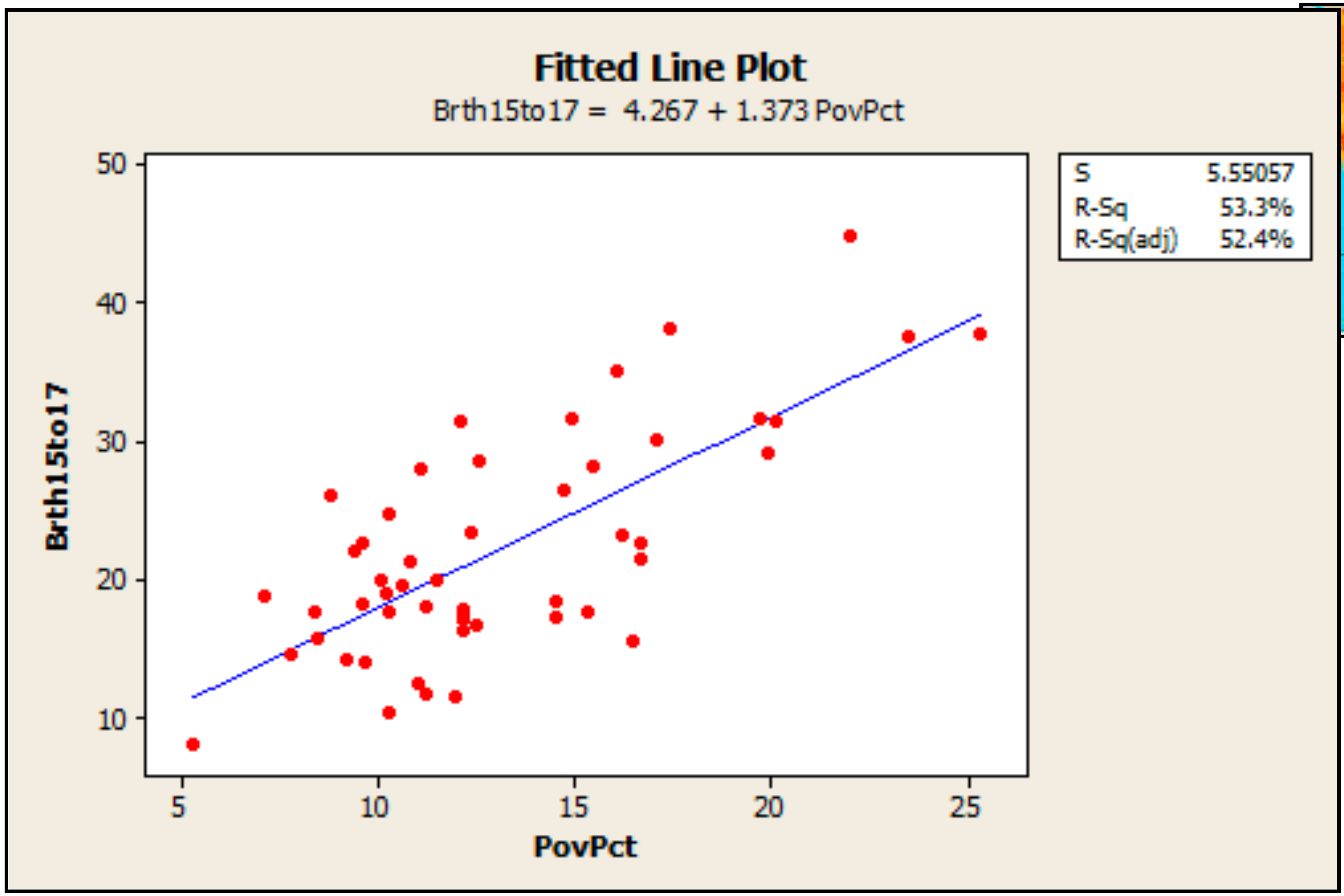
Linear Regression (example)

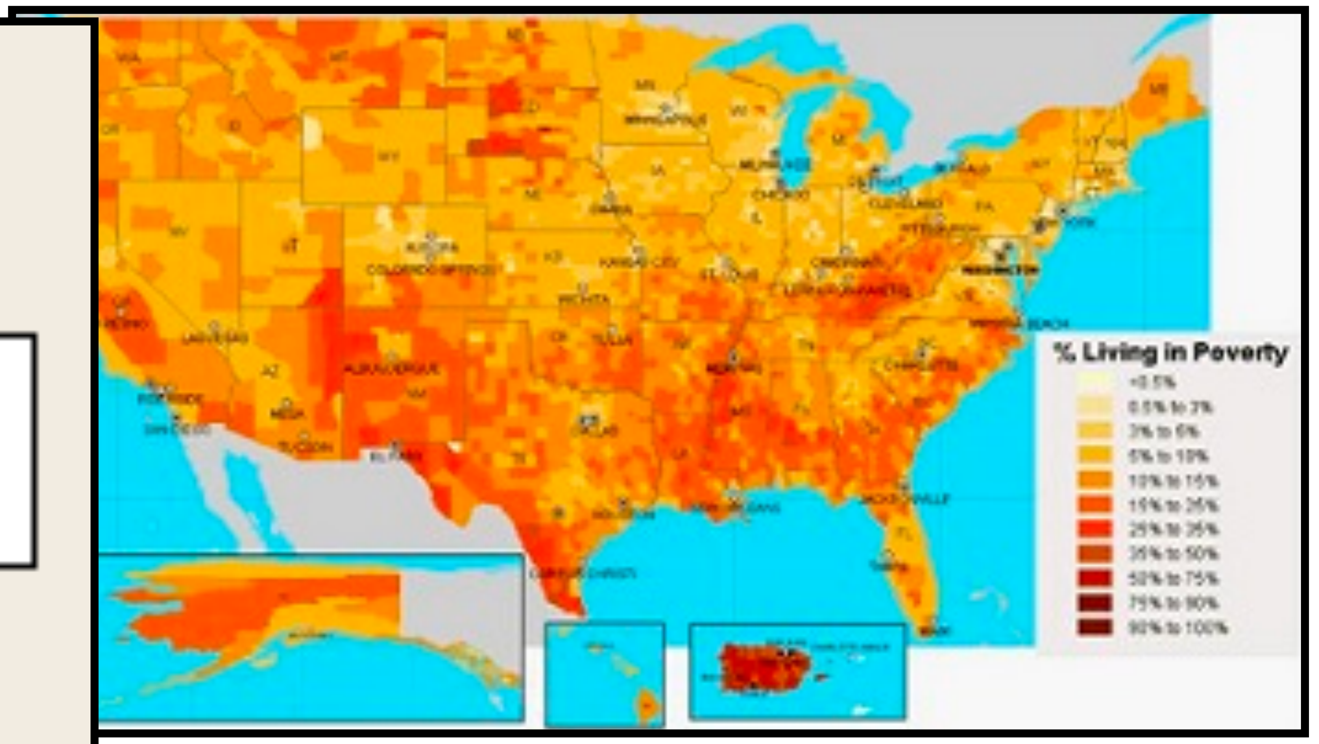




https://online.stat.psu.edu/stat462/node/101/

Linear Regression (example)





https://online.stat.psu.edu/stat462/node/101/

Linear Regression

- Input data: $X \in \mathbb{R}^{d \times n}$, $Y \in \mathbb{R}^{n}$, where $(\vec{x} \in \mathbb{R}^{d}, y \in \mathbb{R}^{1})$ corresponds to a data point.
    - $n \rightarrow$ # of data points, $d \rightarrow$ # of features / input dim.

Linear Regression

- Input data: $X \in \mathbb{R}^{d \times n}$, $Y \in \mathbb{R}^n$, where $(\vec{x} \in \mathbb{R}^d, y \in \mathbb{R}^1)$ corresponds to a data point.
  - $n \rightarrow$ # of data points, $d \rightarrow$ # of features / input dim.
- Goal: to find $\vec{w} \in \mathbb{R}^d$ such that $< \vec{w}, \vec{x} > = y$
  - Minimize $\|X^\top \vec{w} - Y\|^2$

Linear Regression

- Input data: $X \in \mathbb{R}^{d \times n}$, $Y \in \mathbb{R}^{n}$, where $(\vec{x} \in \mathbb{R}^{d}, y \in \mathbb{R}^{1})$ corresponds to a data point.
  - $n \rightarrow$ # of data points, $d \rightarrow$ # of features / input dim.
- Goal: to find $\vec{w} \in \mathbb{R}^{d}$ such that $< \vec{w}, \vec{x} > = y$
  - Minimize $\|X^{\top}\vec{w} - Y\|^{2}$
- Solution: $\vec{w} = (XX^{\top})^{-1}XY$
  - Easy way to remember $\vec{w} = (X^{\top})^{+}Y$

Linear Regression (non-linear?)

- Vanilla linear regression (OLS) assumes that: $y = \vec{w}^\top x$

Linear Regression (non-linear?)

- Vanilla linear regression (OLS) assumes that: $y = \overrightarrow{w}^{\top} x$

- What if we want to model non linear relationship:

$$y = \overrightarrow{w_1}^{\top} x + \overrightarrow{w_2}^{\top} x^2 + \overrightarrow{w_3}^{\top} x^3 + \cdots + \overrightarrow{w_p}^{\top} x^p$$

- Is there an easy way to do this?

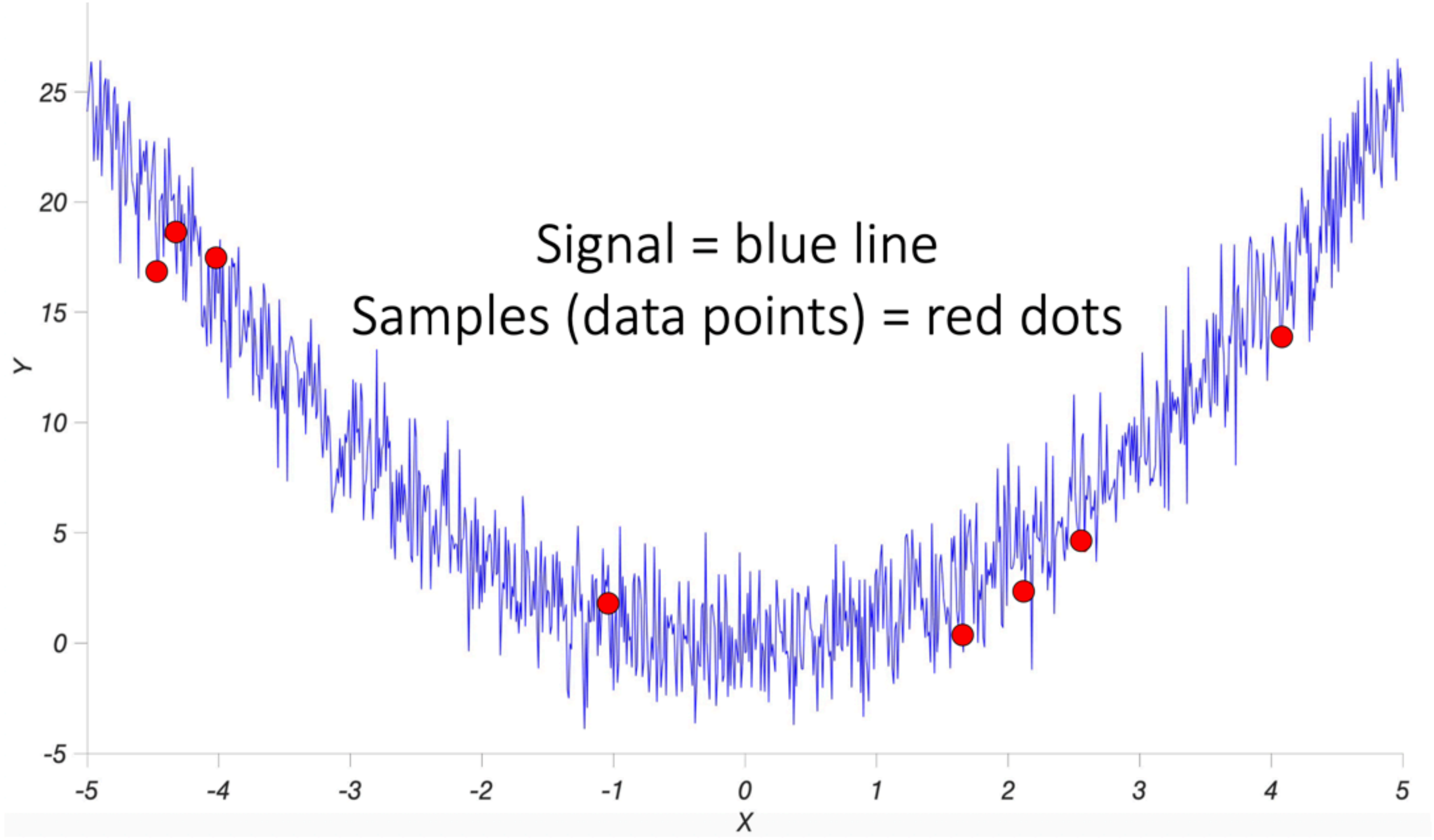
Linear Regression (non-linear?)

- Vanilla linear regression (OLS) assumes that: $y = \vec{w}^\top x$
- What if we want to model non linear relationship:

$$y = \vec{w_1}^\top x + \vec{w_2}^\top x^2 + \vec{w_3}^\top x^3 + \cdots + \vec{w_p}^\top x^p$$

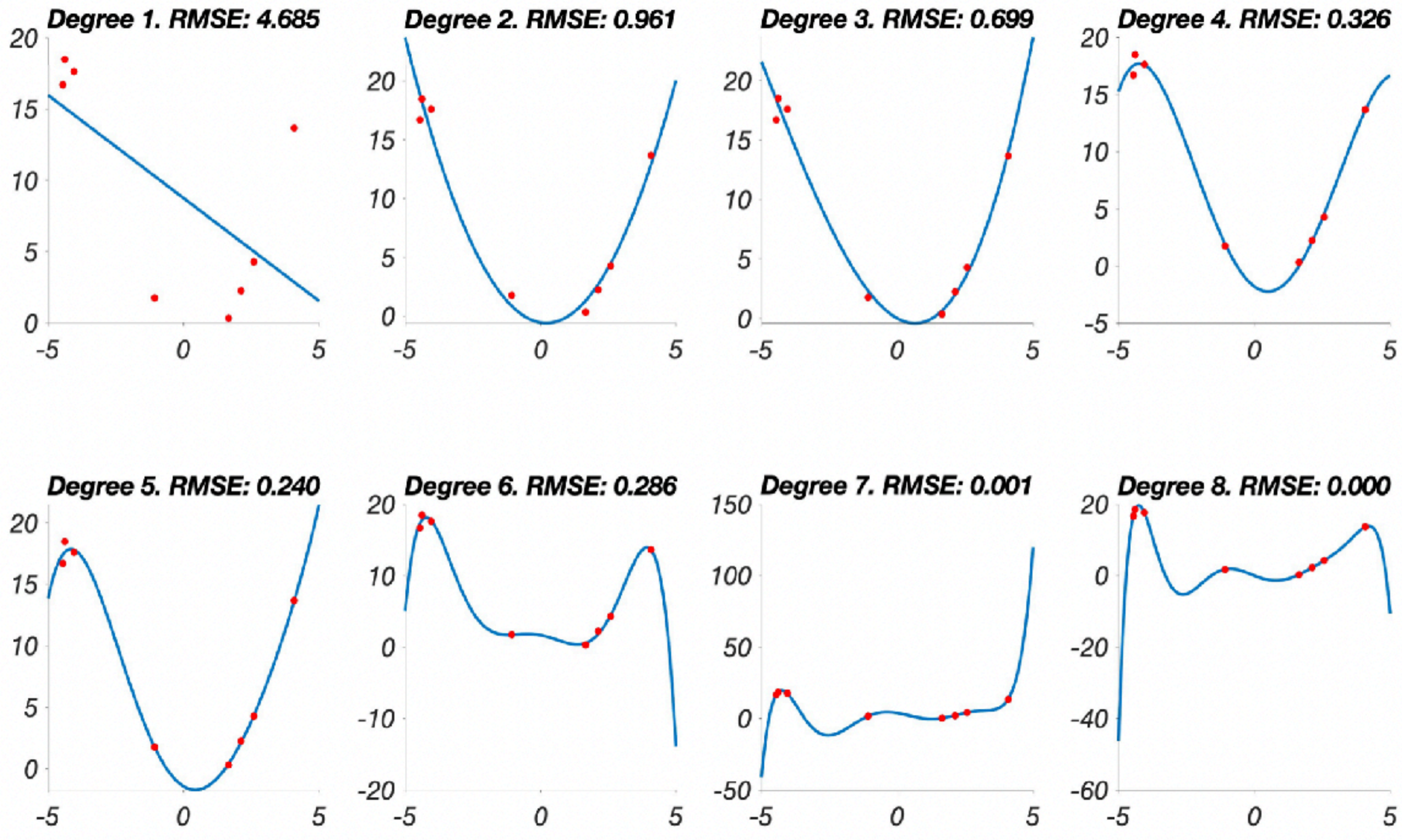
- Is there an easy way to do this?
    - Define new variable: $z = [x, x^2, x^3, \cdots, x^p]$
    - Do vanilla linear regression: $y = \vec{w}^\top z$

Linear Regression (bias vs variance)



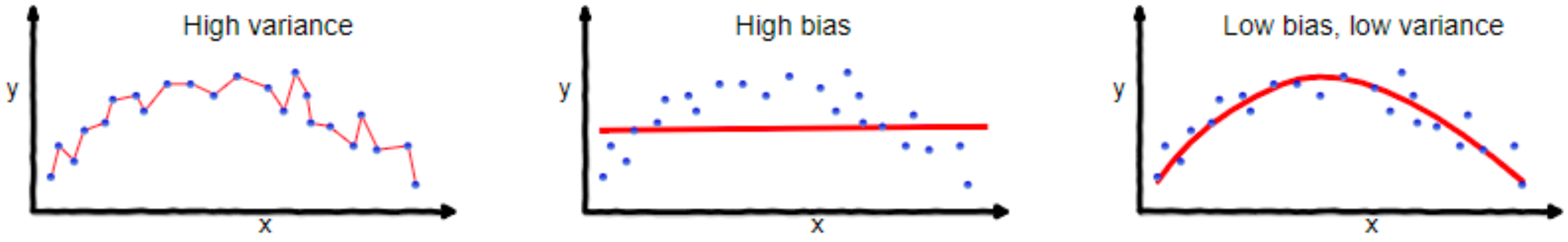
Slide credit: Pascal Wallisch

Linear Regression (bias vs variance)



Slide credit: Pascal Wallisch

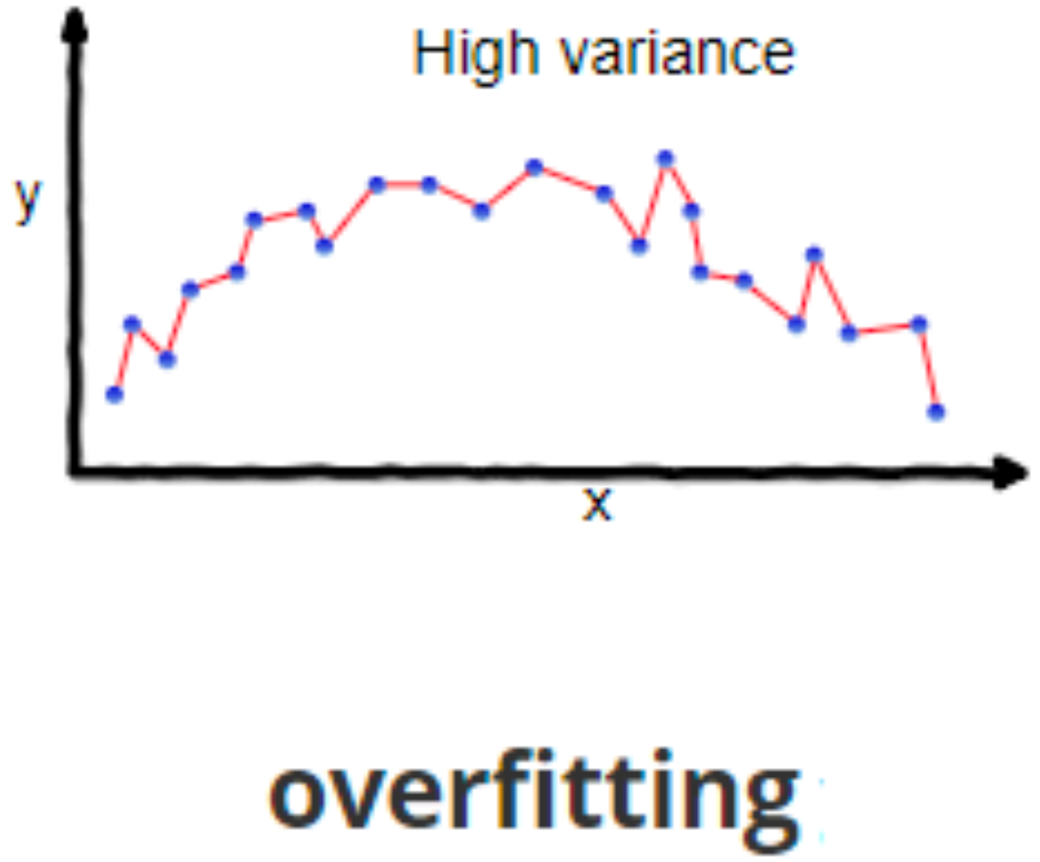
Bias vs variance tradeoff

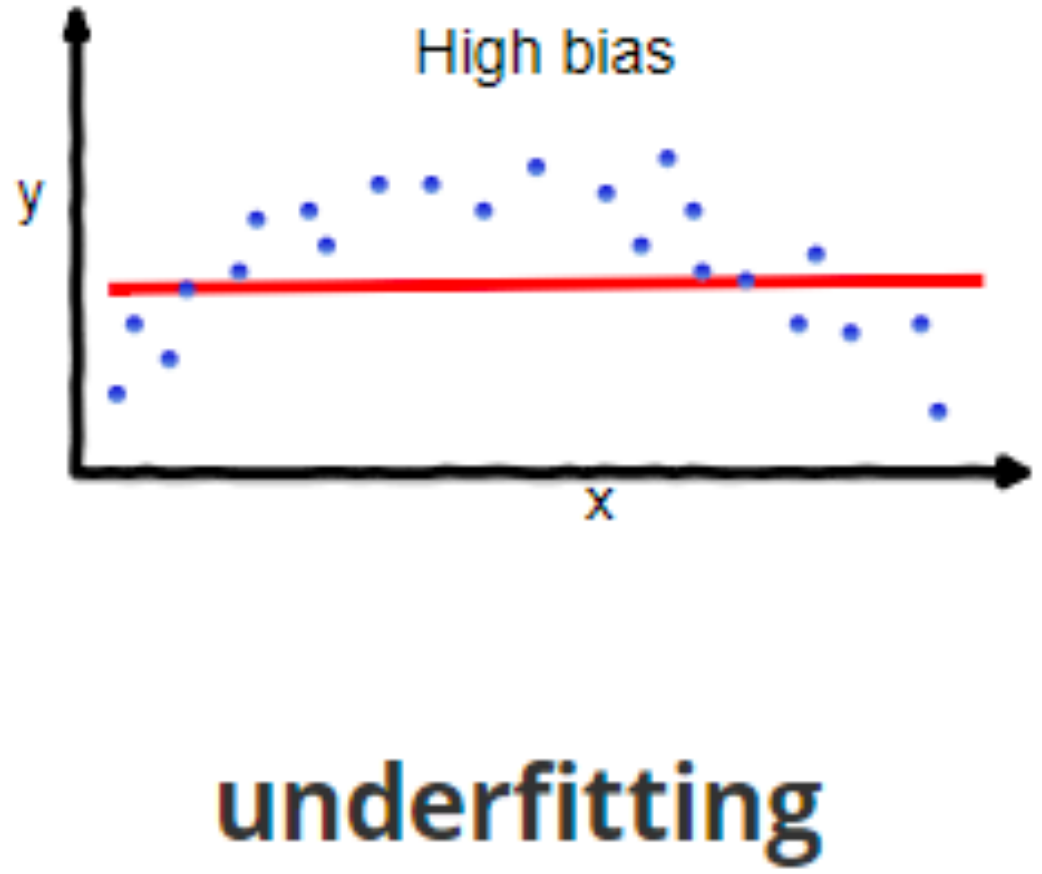


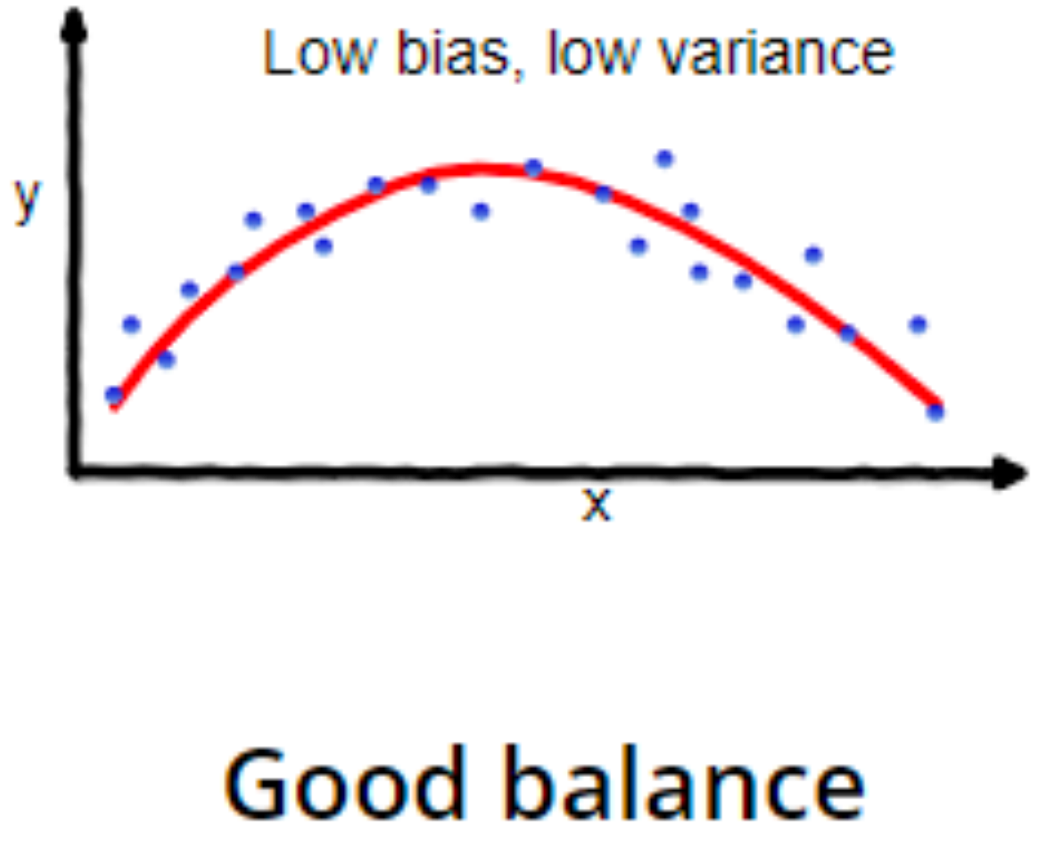
overfitting | **underfitting** | **Good balance**

https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

Bias vs variance tradeoff







- How do we know if a model has high bias or variance?
- How do we find the balance?
    - Algorithms, Validation, prior knowledge

https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

Scratchpad

Ridge Regression

- OLS: Minimize $||X^\top \vec{w} - Y||^2$
    - Solution: $\vec{w} = (XX^\top)^{-1}XY$

https://towardsdatascience.com/ridge-regression-for-better-usage-2f19b3a202db

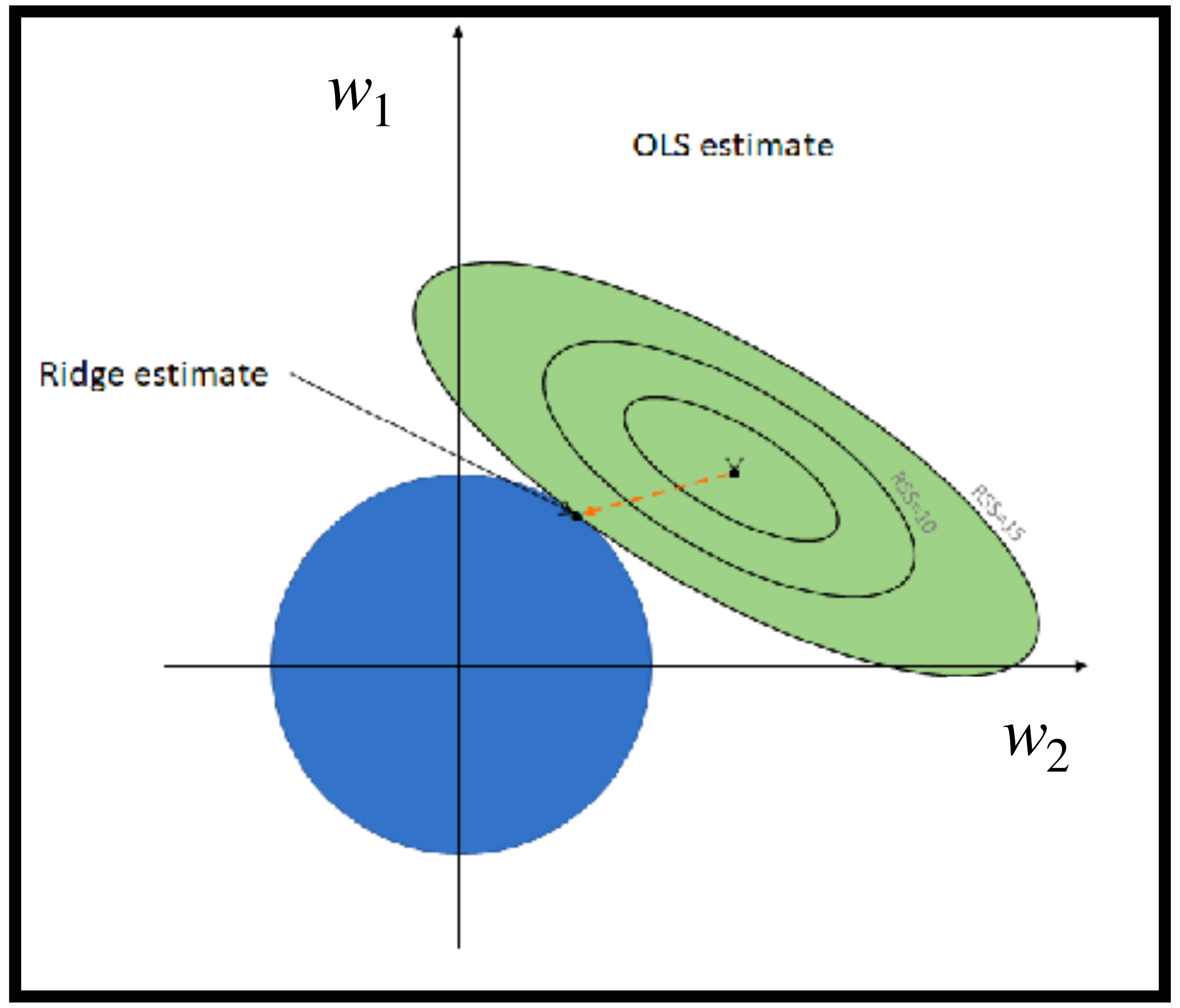
Ridge Regression

- OLS: Minimize $||X^\top \vec{w} - Y||^2$
  - Solution: $\vec{w} = (XX^\top)^{-1}XY$
- Ridge Regression: Minimize $||X^\top \vec{w} - Y||^2$ such that $||\vec{w}||^2 \leq c^2$

https://towardsdatascience.com/ridge-regression-for-better-usage-2f19b3a202db

Ridge Regression

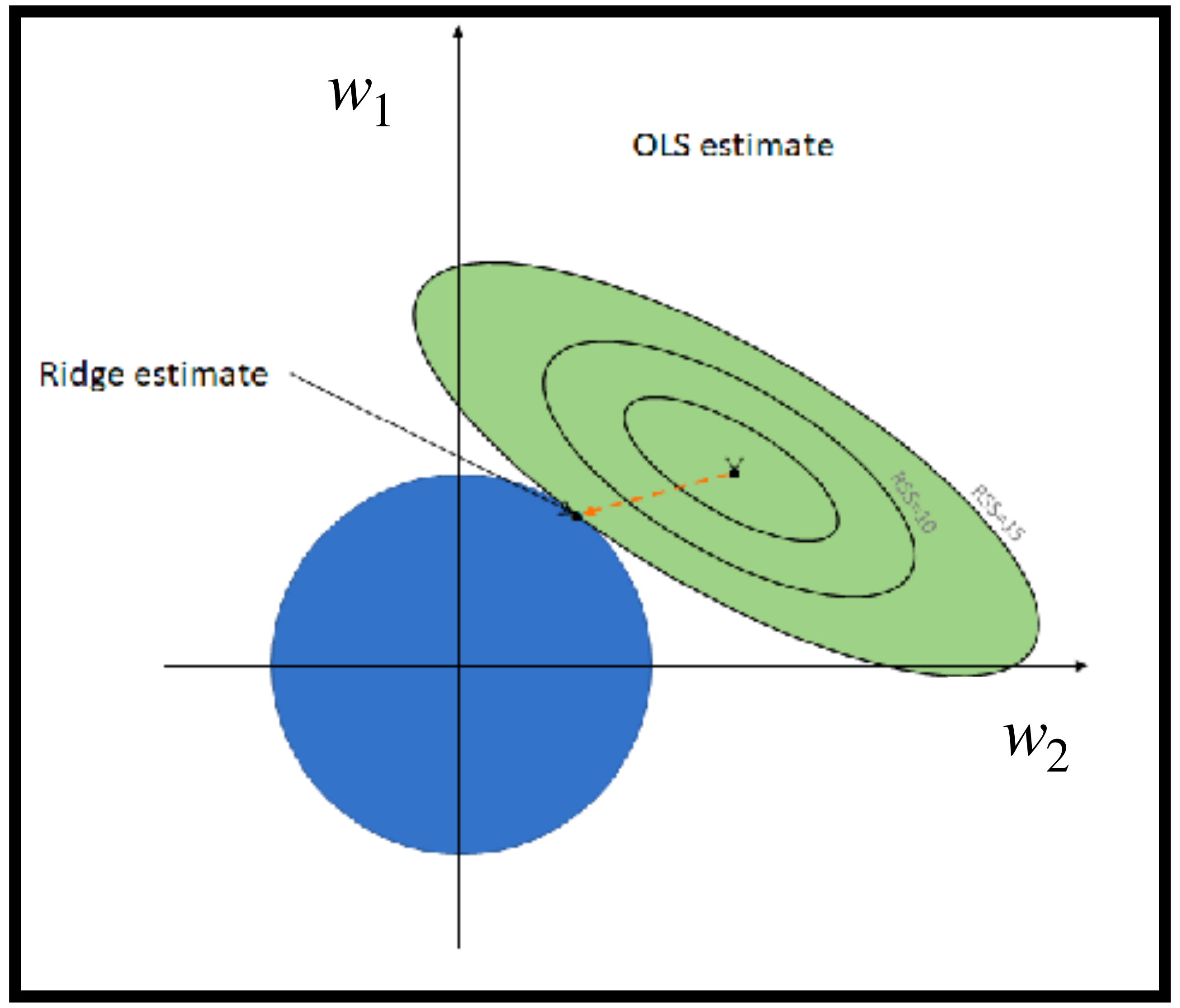
- OLS: Minimize $||X^{\top}\vec{w} - Y||^2$
  - Solution: $\vec{w} = (XX^{\top})^{-1}XY$
- Ridge Regression: Minimize $||X^{\top}\vec{w} - Y||^2$ such that $||\vec{w}||^2 \leq c^2$



https://towardsdatascience.com/ridge-regression-for-better-usage-2f19b3a202db

Ridge Regression

- OLS: Minimize $\|X^\top \vec{w} - Y\|^2$
  - Solution: $\vec{w} = (XX^\top)^{-1}XY$
- Ridge Regression: Minimize $\|X^\top \vec{w} - Y\|^2$ such that $\|\vec{w}\|^2 \leq c^2$
  - Loss function: $(\|X^\top \vec{w} - Y\|^2 + \lambda\|\vec{w}\|^2)$
  - Solution: $\vec{w} = (XX^\top + \lambda \mathbf{I})^{-1}XY$



https://towardsdatascience.com/ridge-regression-for-better-usage-2f19b3a202db

Lasso Regression

- OLS: Minimize $||X^\top \vec{w} - Y||^2$
    - Solution: $\vec{w} = (XX^\top)^{-1}XY$

https://towardsdatascience.com/ridge-regression-for-better-usage-2f19b3a202db

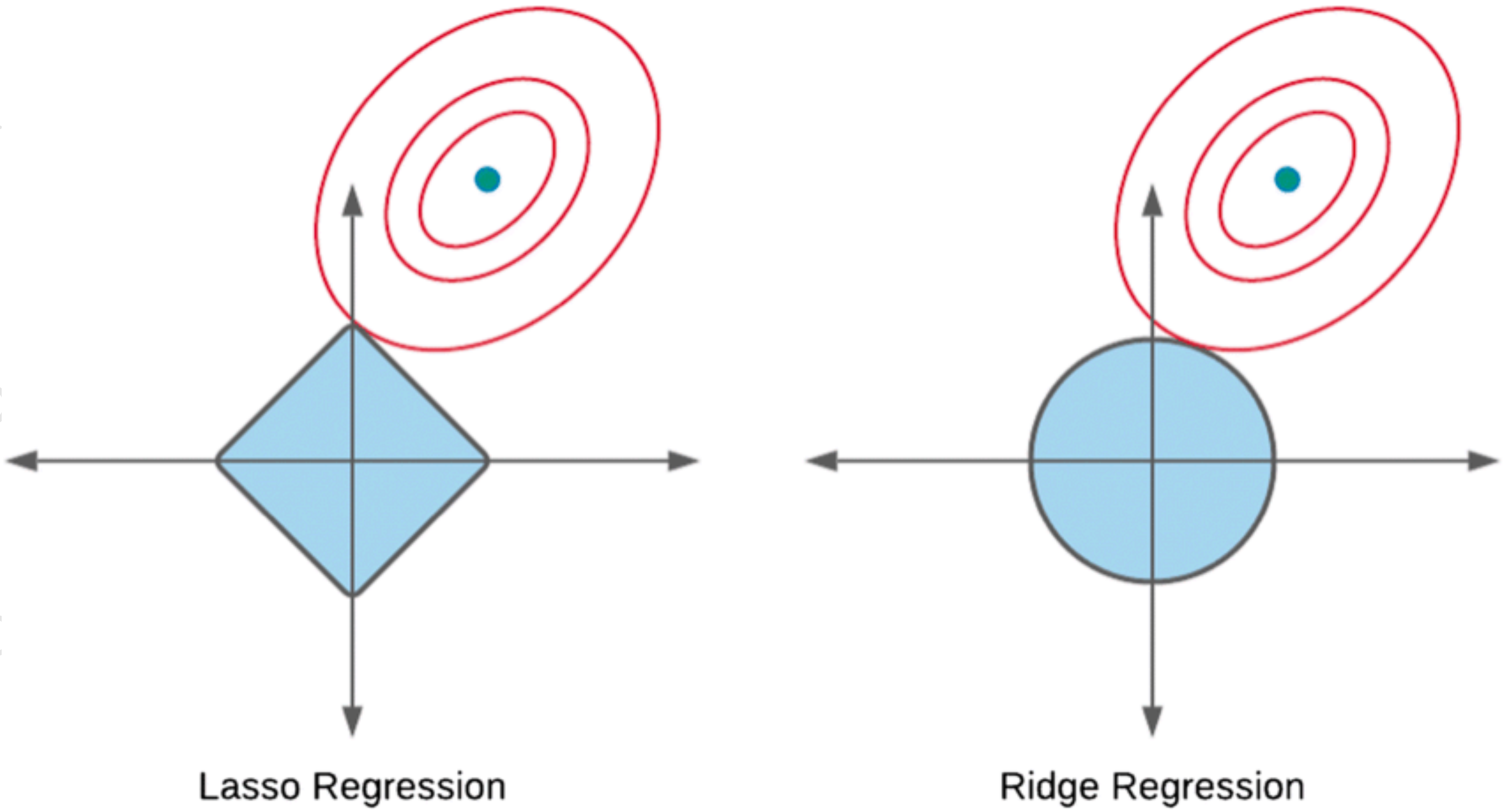
Lasso Regression

- OLS: Minimize $||X^\top \vec{w} - Y||^2$
  - Solution: $\vec{w} = (XX^\top)^{-1}XY$
- Ridge Regression: Minimize $||X^\top \vec{w} - Y||^2$ such that $||\vec{w}||_1 \leq c$
  - Loss function: $(||X^\top \vec{w} - Y||^2 + \lambda ||\vec{w}||_1)$
  - Solution: *From numerical methods.*

https://towardsdatascience.com/ridge-regression-for-better-usage-2f19b3a202db

Lasso Regression

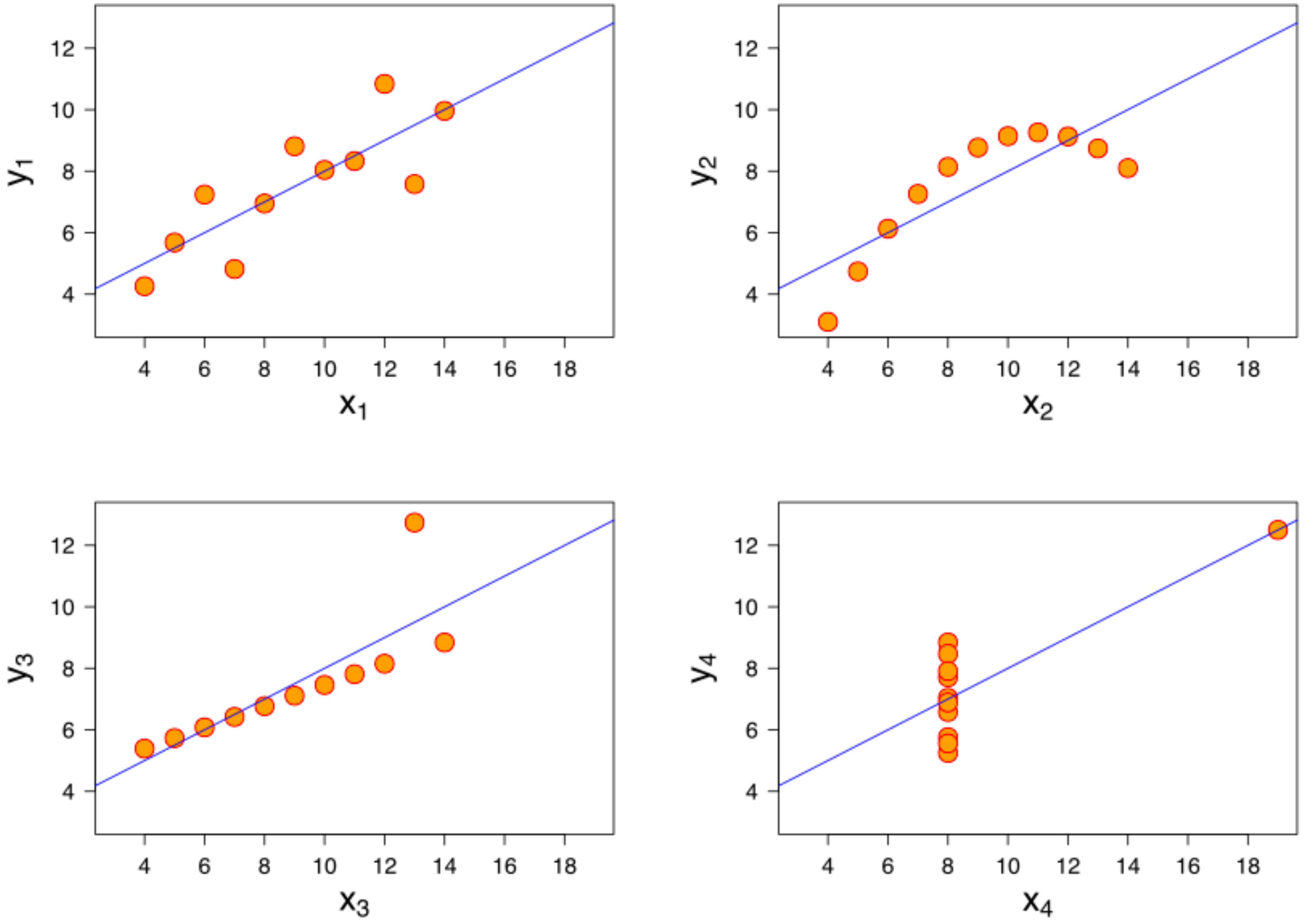
- OLS: Minimize $||X^{\top}\vec{w} - Y||^2$
    - Solution: $\vec{w}$
- Ridge Regressic $_1 \leq c$
    - Loss functic
    - Solution: *From numerical methods.*



https://towardsdatascience.com/ridge-regression-for-better-usage-2f19b3a202db

Pathological cases

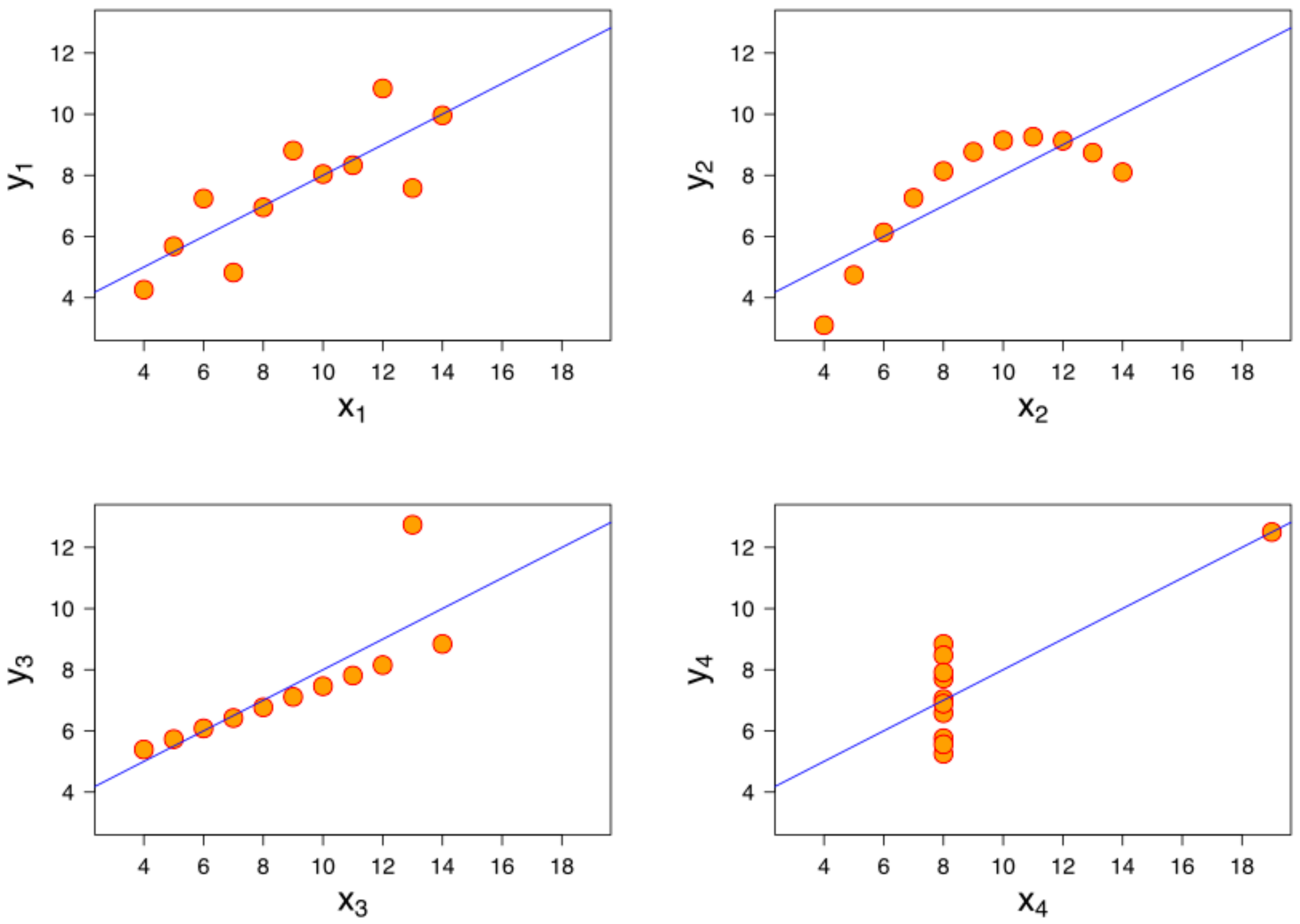
- Anscombe's quartet



https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Pathological cases

- Anscombe's quartet

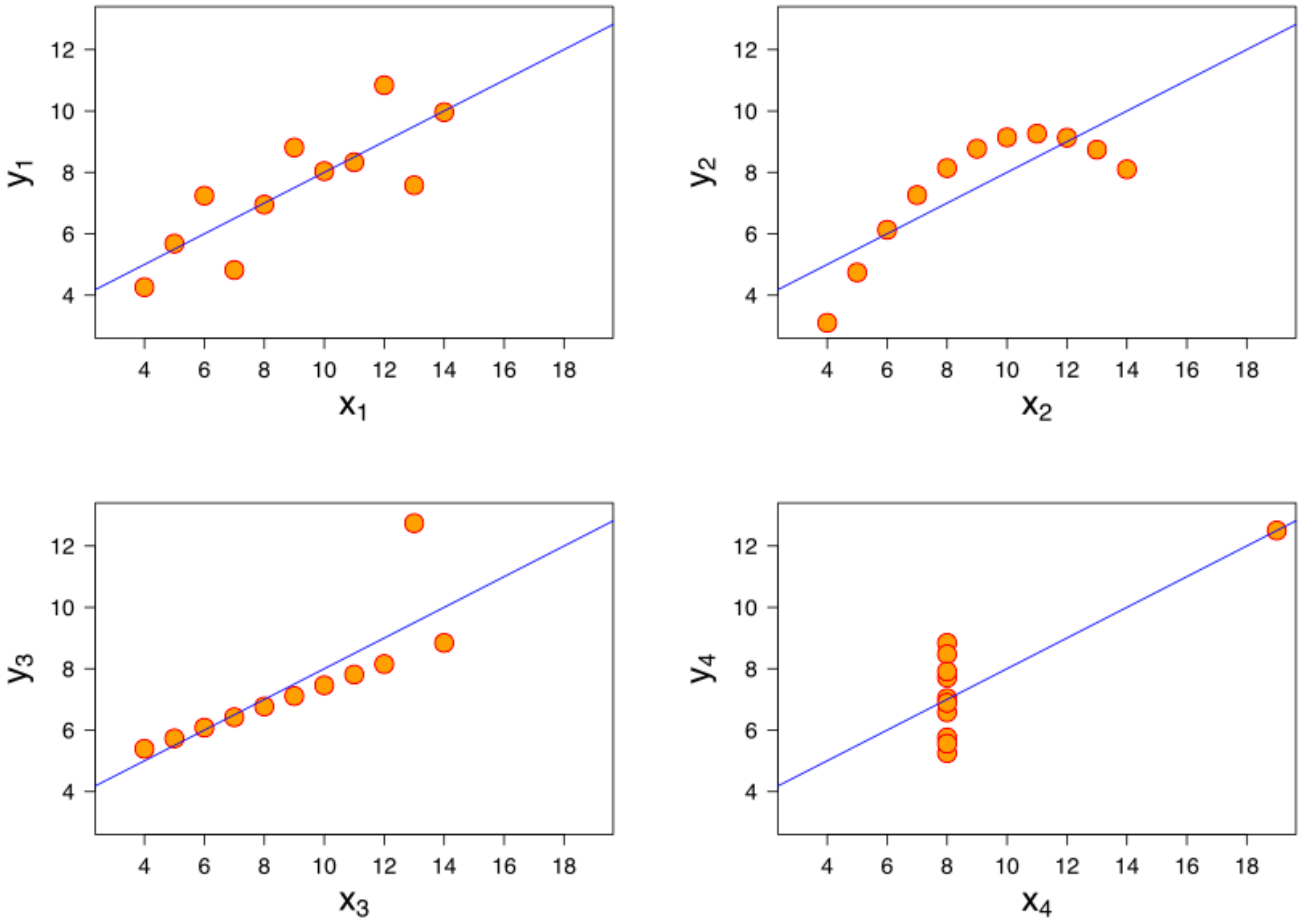


Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Pathological cases

- Anscombe's quartet



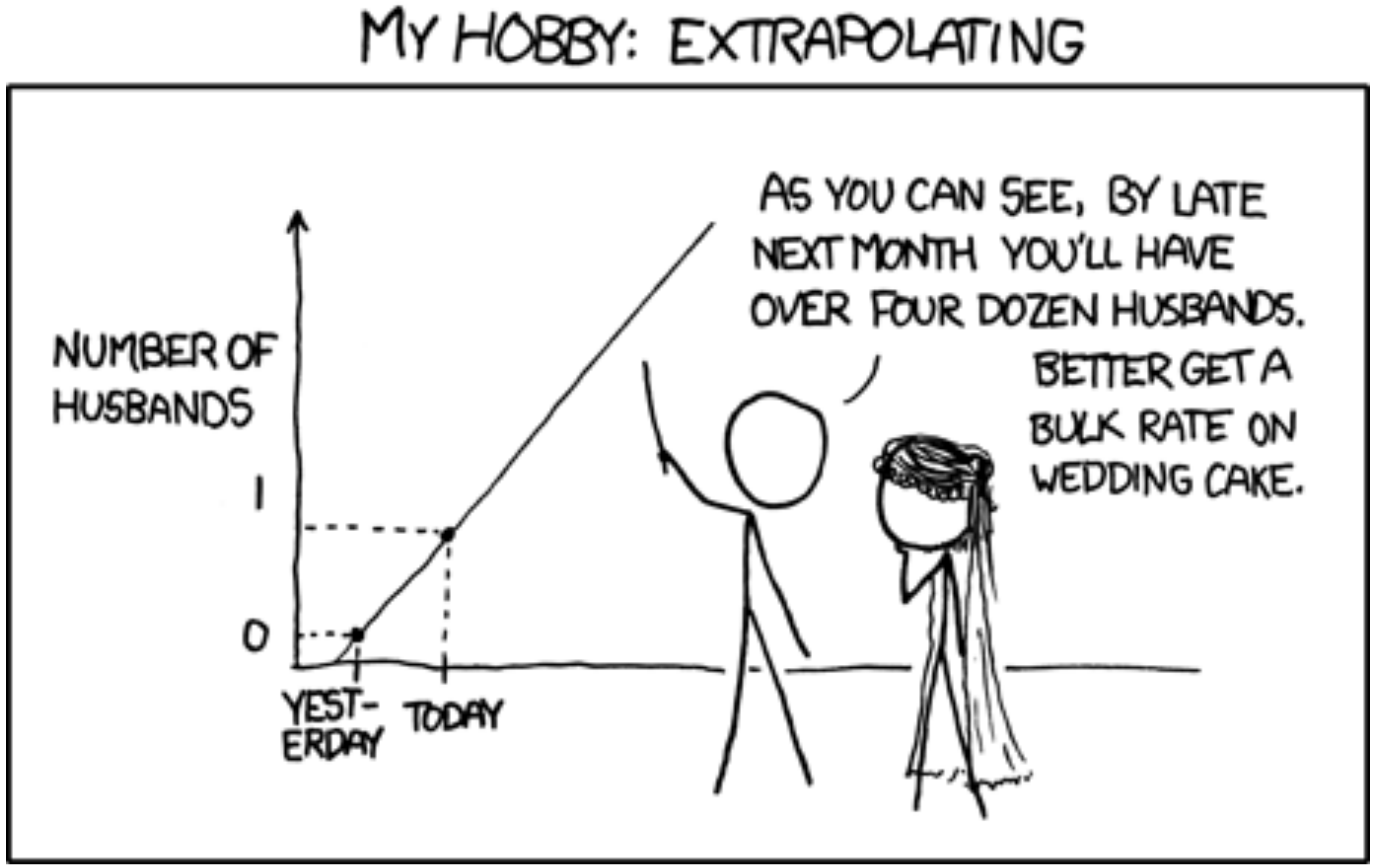
Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

Takeaway: Visualize your data!!

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Pathological cases

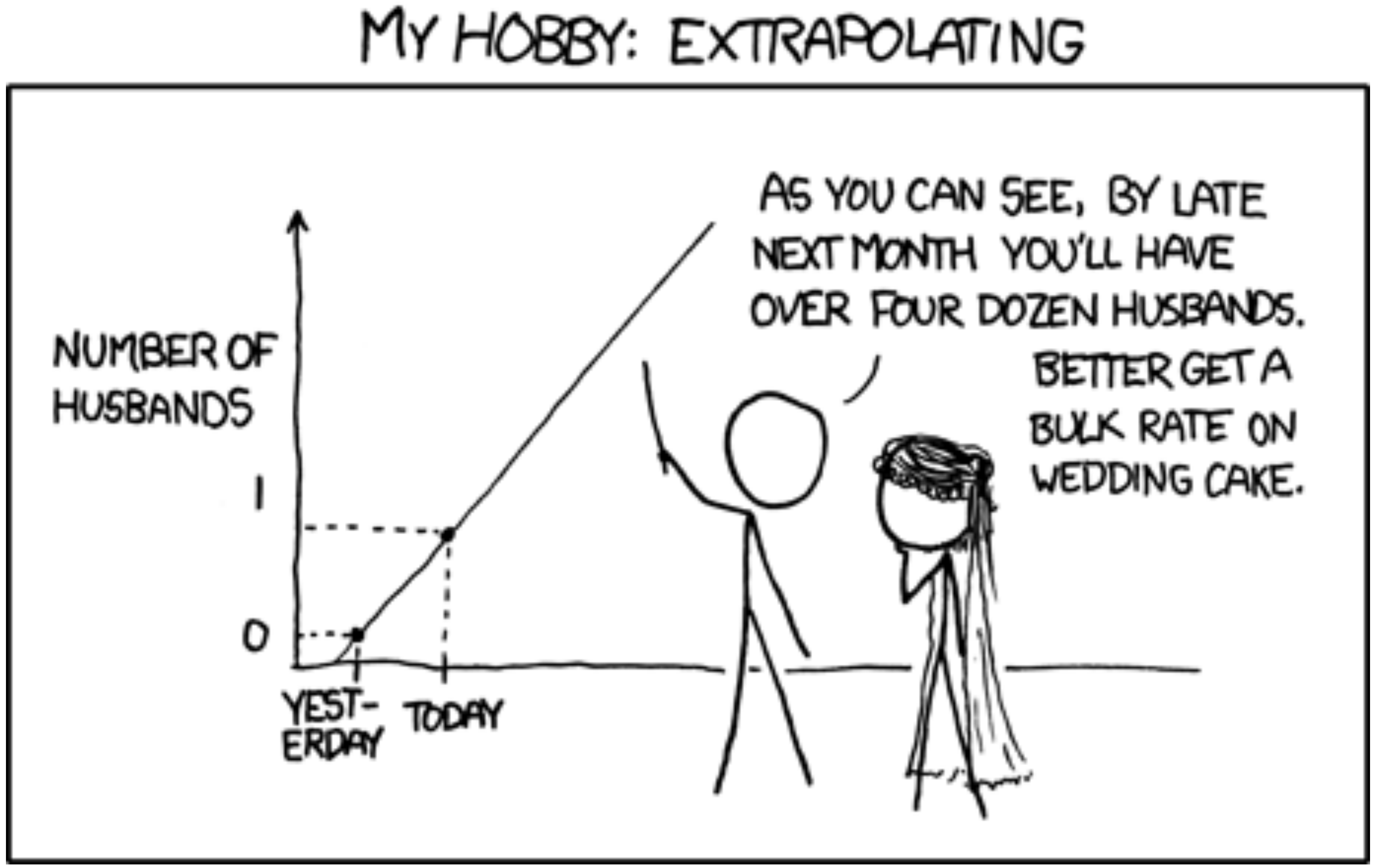
- Extrapolation



https://xkcd.com/605/

Pathological cases

- Extrapolation

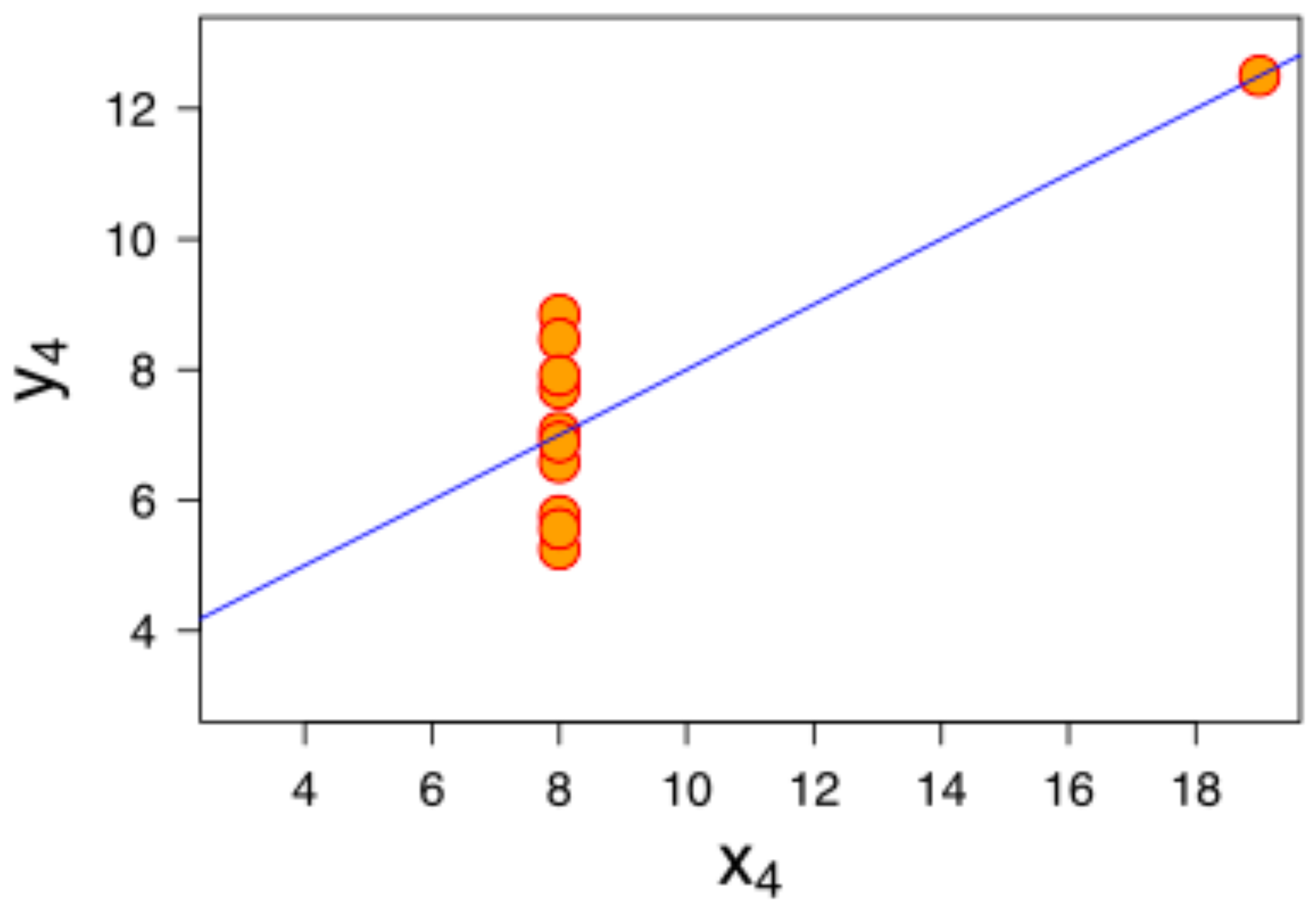


Takeaway: Don't trust your model outside the training data distribution

https://xkcd.com/605/

Pathological cases

- Noise in data



Pathological cases

- Large feature size

Additional Reading

- https://www.mit.edu/~6.s085/notes/lecture3.pdf
- https://www.coconino.edu/resources/files/pdfs/academics/sabbatical-reports/kate-kozak/chapter_10.pdf

Questions?