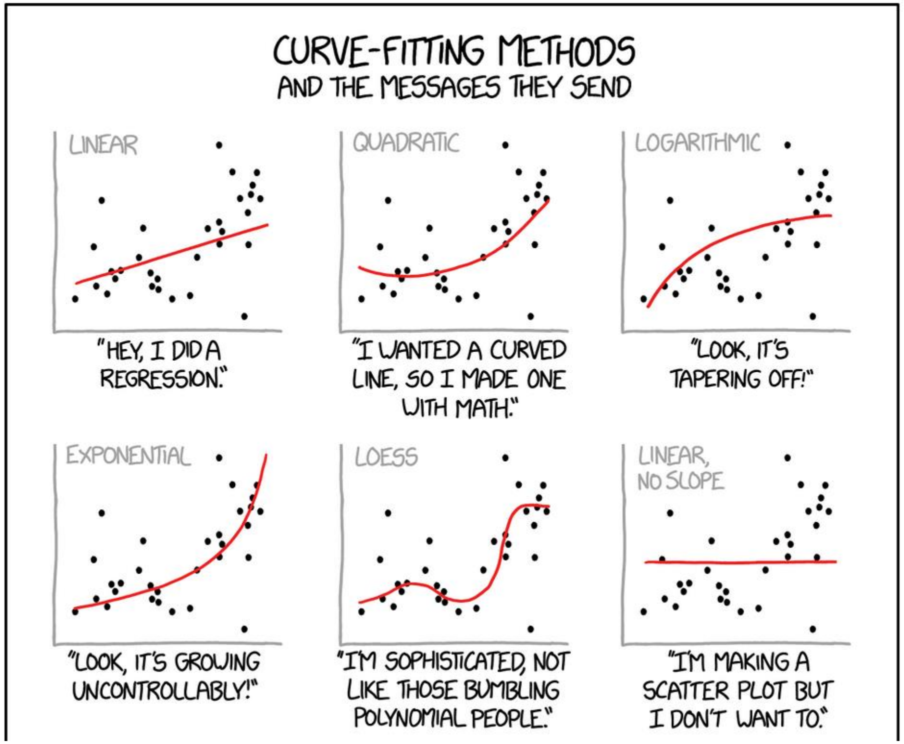
Inference in Linear Regression



https://xkcd.com/2048/

SML201: Introduction to Data Science, Spring 2020

Michael Guerzhoy

Refresher: Linear Regression

Inputs	Outputs
$x_1^{(1)}, x_2^{(1)}, \dots, x_n^{(1)}$	$y^{(1)}$
$x_1^{(2)}, x_2^{(2)}, \dots, x_n^{(2)}$	$y^{(2)}$
$x_1^{(3)}, x_2^{(3)}, \dots, x_n^{(3)}$	$y^{(3)}$

New prediction:

$$\hat{y}^{(i)} = a_0 + a_1 x_1^{(i)} + a_2 x_2^{(i)} + \cdots + a_n x_n^{(i)}$$

Error/residual:

$$e^{(i)} = y^{(i)} - \hat{y}^{(i)}$$

Sum of Squared Errors/Cost:

$$\sum_{i=1}^{m} \left(y^{(i)} - \hat{y}^{(i)}\right)^2$$

minimize

Linear Regression: Null Hypothesis

- Usually of the form $a_j = 0$
  - The j-th feature is not associated with the output

Linear Regression: Model Assumptions

- $y^{(i)} \approx a_0 + a_1 x_1^{(i)} + a_2 x_2^{(i)} + \cdots + a_n x_n^{(i)}$
  - Can check by plotting if there are few x's. Otherwise check with diagnostic plots
- $e^{(i)} \sim N(0, \sigma^2)$
  - Check with diagnostic plots
- The residuals $e^{(i)}$ are independent of each other, and independent of x
  - Check with diagnostic plots

Q-Q plots

- Sort all the observations from both distribution 1 and the normal distribution
- Plot the observations from distribution 1 (in order) vs. the observations from the normal (in order)
- Approx straight line if distribution 1 is normal

Q-Q plots

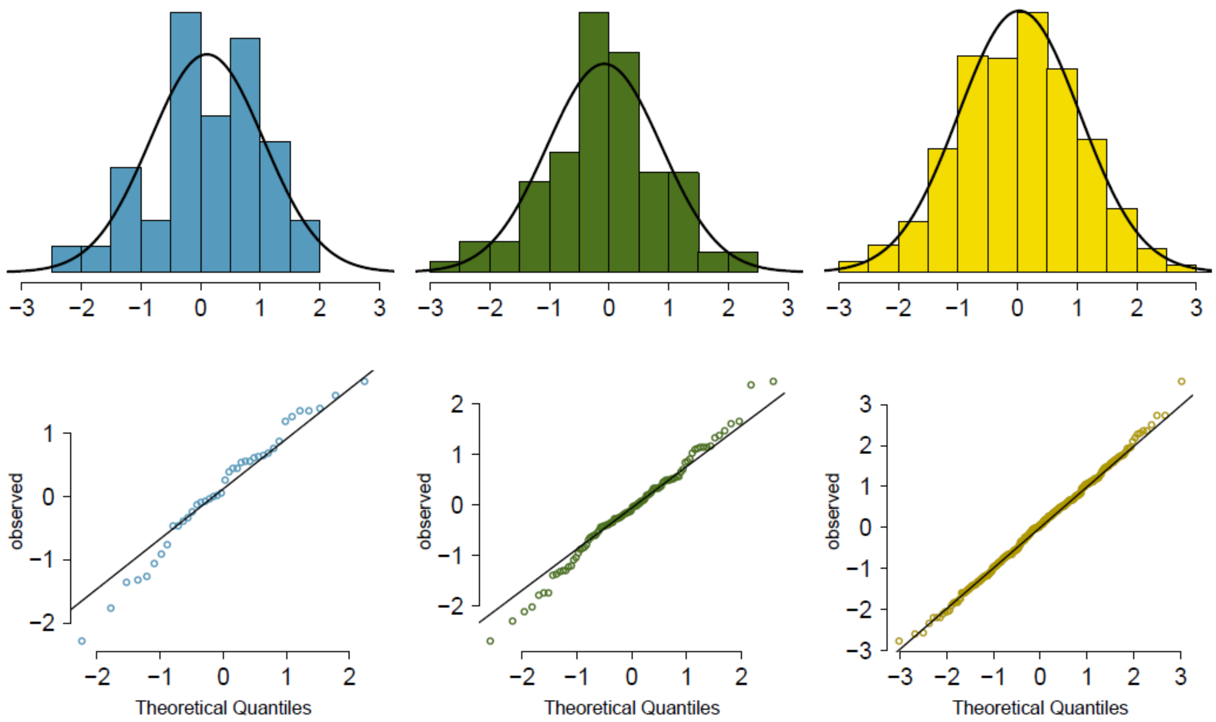


Figure 3.11: Histograms and normal probability plots for three simulated normal data sets; $n = 40$ (left), $n = 100$ (middle), $n = 400$ (right).

OpenIntro Statistics

Linear Regression: test

- For the null hypothesis $a_j = 0$, and assuming the model assumptions are satisfied, we can compute a p-value using a t-test
- (Switch to R)

Linear Regression: Multiple Comparisons warning + F-test

- We can only run *one* pre-registered t-test
  - If there are multiple features, cannot test the hypotheses that each of them is non-zero
- Can run an F-test, where the null hypothesis is that all the $a_j's$ are 0
  - (Switch to R)

Linear Regression: correlation is not causation

- Rejecting the hypothesis that $a_j = 0$ doesn't mean x_j influences the value of y
  - Reverse causation
  - Common cause
  - Indirect causation
  - Coincidence
  - …
  - (Type I error)

R^2

- If we are trying to predict $y^{(i)}$, the simplest thing is to predict $\bar{y}$ every time.
- Can compute

$$R^2 = 1 - \frac{\sum_{i=1}^{m}\left(y^{(i)} - \hat{y}^{(i)}\right)^2}{\sum_{i=1}^{m}\left(y^{(i)} - \bar{y}\right)^2}$$

Low ratio: our predictions are much better than the baseline

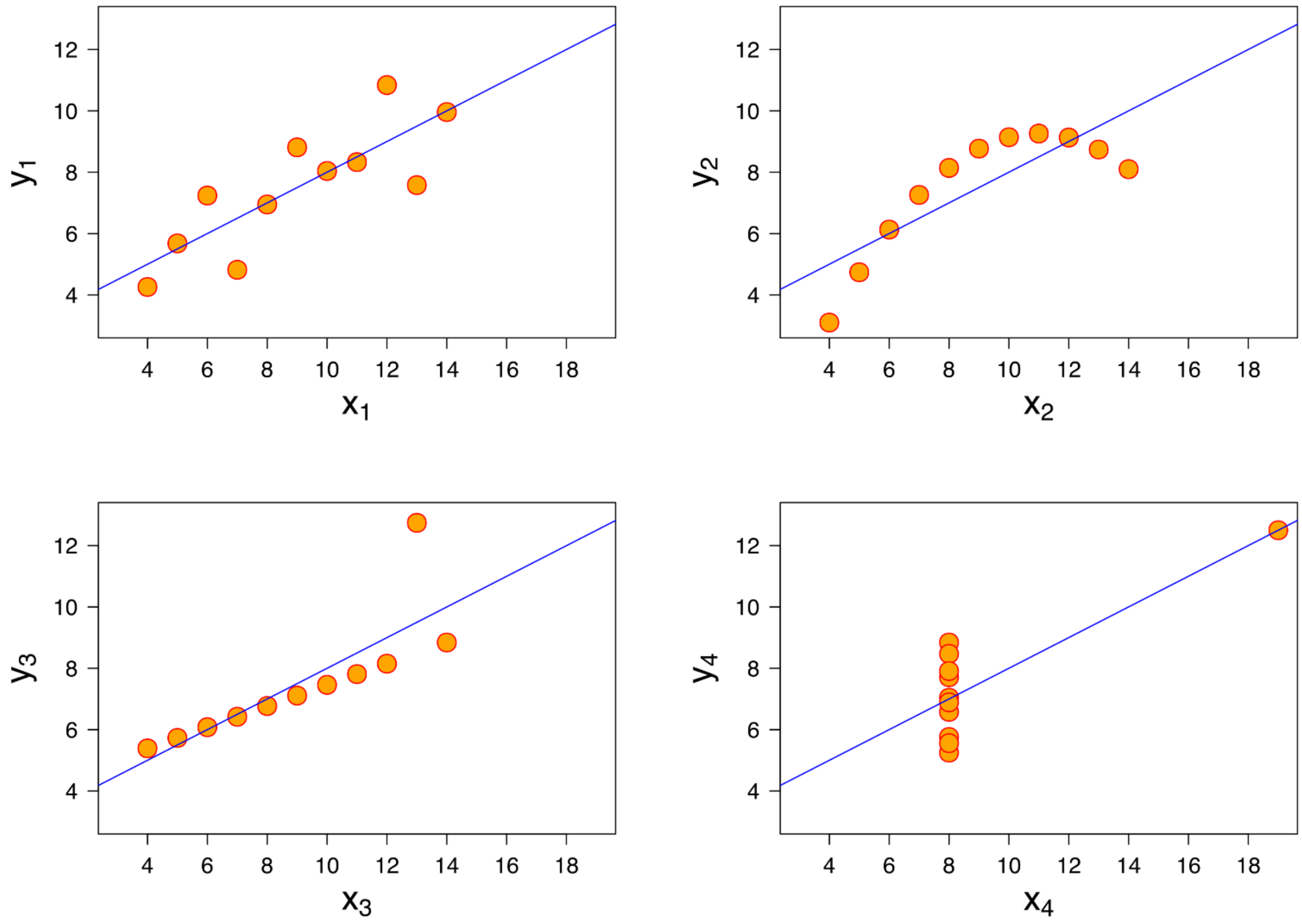
Ratio close to 1: our predictions are the same as the baseline

- R^2 close to 1 is usually interpreted as a strong linear relationship between the inputs and the outputs
- Low R^2 is usually interpreted as a weak (linear) relationship

Correlation

- Trying to predict $y \approx a_0 + a_1 x$
- The correlation is $\mathrm{r} = \sqrt{R^2}$ is y generally increases when x increases, and $\mathrm{r} = -\sqrt{R^2}$ otherwise

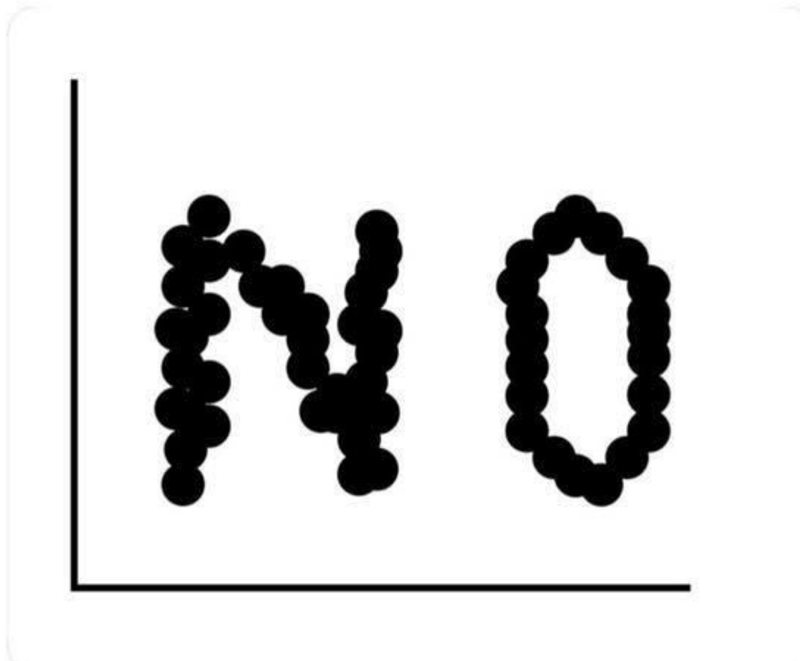
Anscombe's quartet



r=0.816 for all four datasets

Dan Quintana
@dsquintana

$r = .23, p = .042$



8:19 AM · 2019-04-17 · Twitter Web App

Linear Regression summary

- Formulate null hypothesis
- Collect data
- Visualize data to check model assumptions
- If model assumptions seem approximately satisfied, can run regression