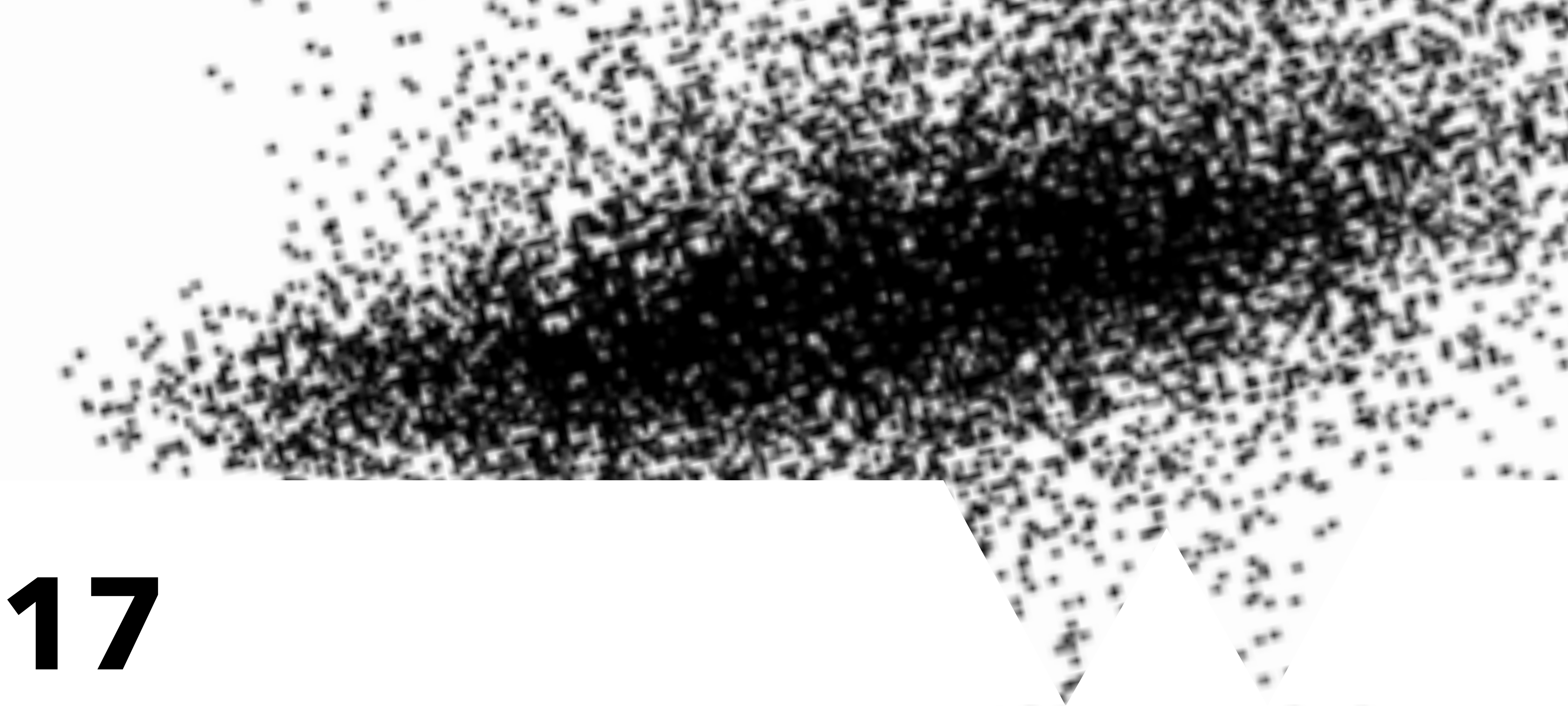


ST 117

5. Regression

WARWICK

Lecture 25
(Week 9)

Examples
Model fit
Data transformations

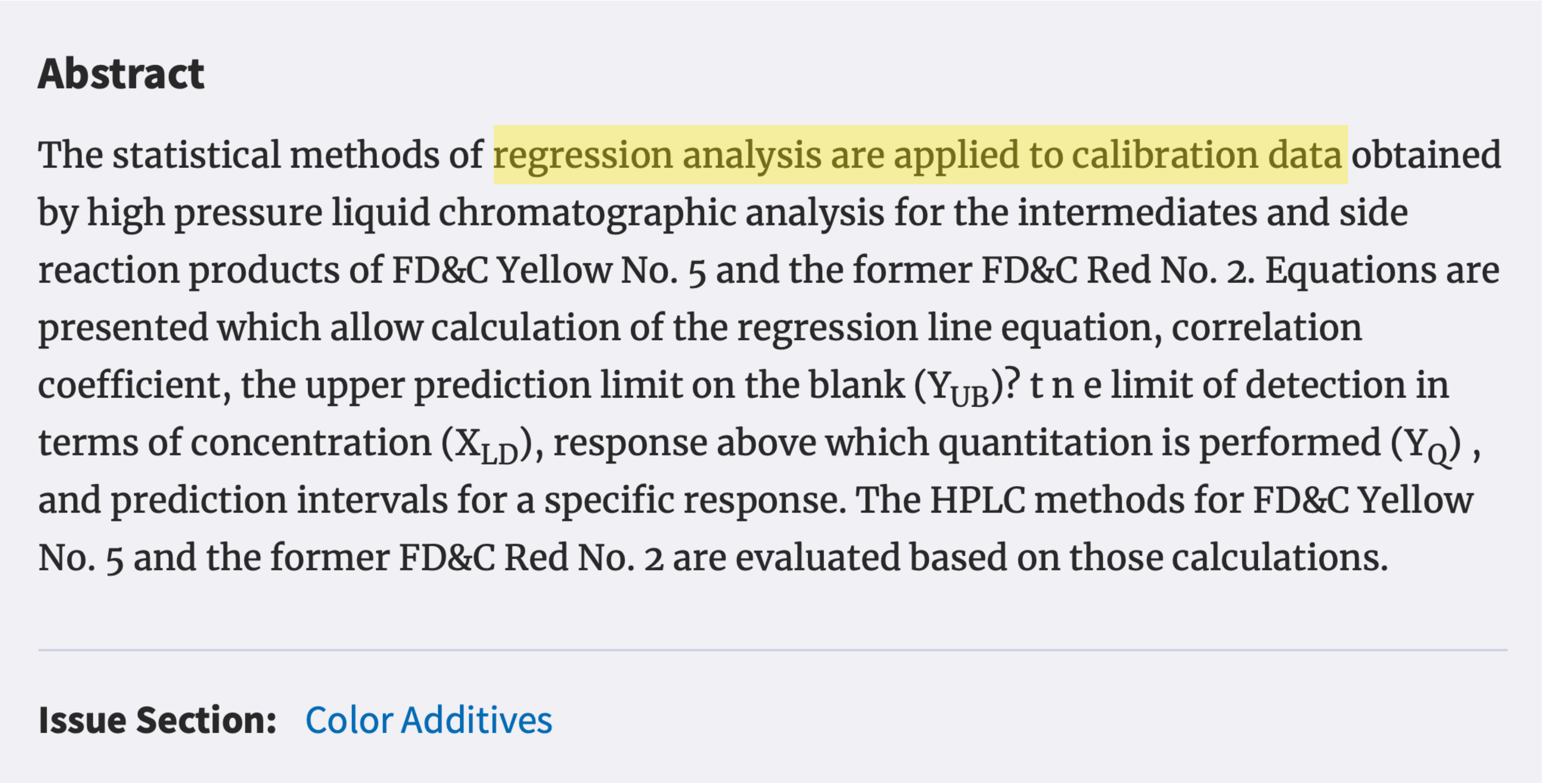
Examples for Linear Regression Fit and Diagnostics

Real-world examples from a range of domains (*)

1. **Analytical chemistry:** Yellow dye quantification by chromatography
2. **Environmental Sciences:** Stream depth and flow
3. **Medical geography/epidemiology:** Breast cancer mortality in 301 countries

(*) Source: John Rice, Mathematical Statistics and Data Analysis, Duxbury Press

Regression in practice: 1. analytical chemistry

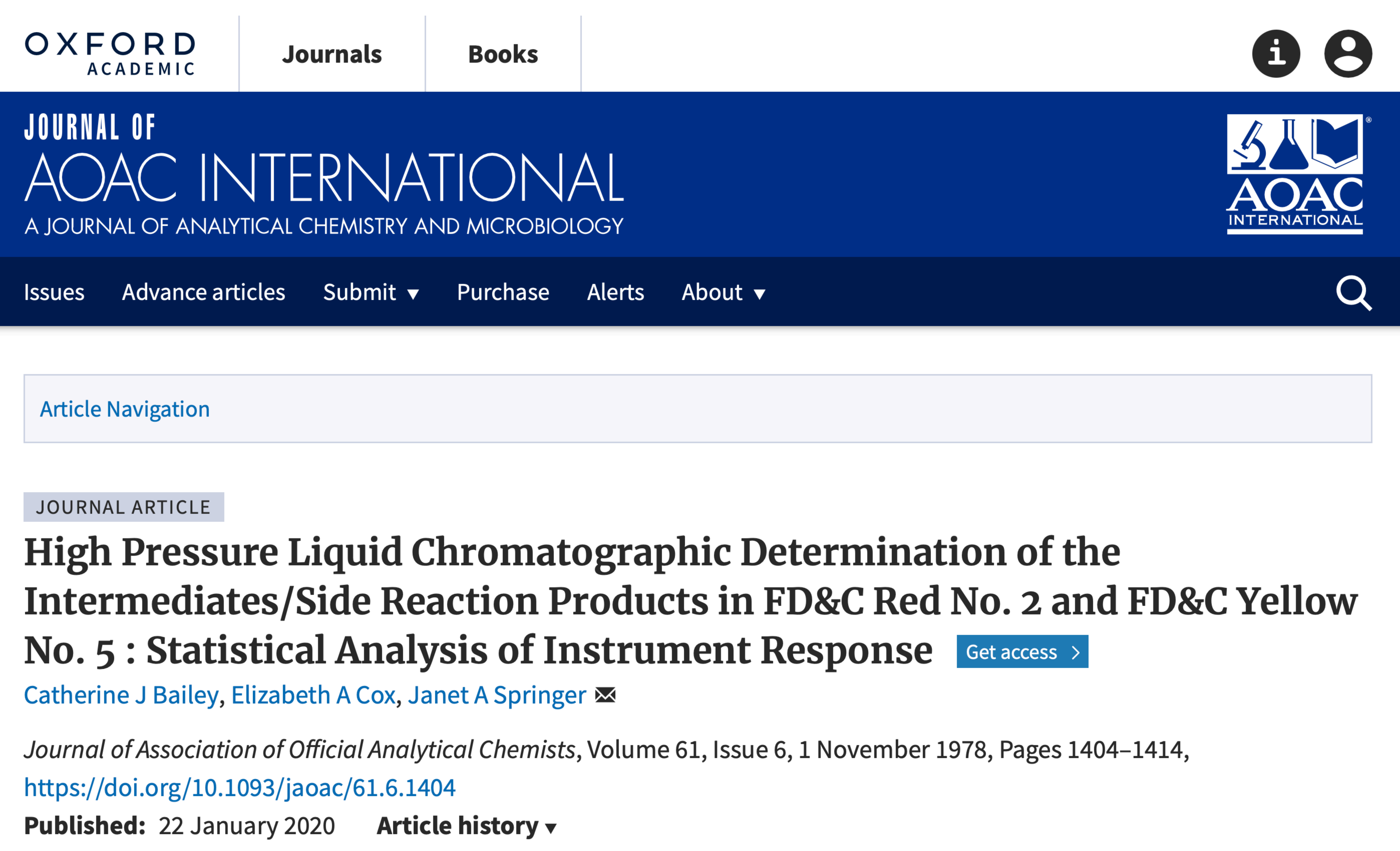


Abstract

The statistical methods of regression analysis are applied to calibration data obtained by high pressure liquid chromatographic analysis for the intermediates and side reaction products of FD&C Yellow No. 5 and the former FD&C Red No. 2. Equations are presented which allow calculation of the regression line equation, correlation coefficient, the upper prediction limit on the blank (Y_{UB})? t n e limit of detection in terms of concentration (X_{LD}), response above which quantitation is performed (Y_Q) , and prediction intervals for a specific response. The HPLC methods for FD&C Yellow No. 5 and the former FD&C Red No. 2 are evaluated based on those calculations.

Issue Section: Color Additives

Catherine J Bailey, Elizabeth A Cox, Janet A Springer, High Pressure Liquid Chromatographic Determination of the Intermediates/Side Reaction Products in FD&C Red No. 2 and FD&C Yellow No. 5 : Statistical Analysis of Instrument Response, *Journal of Association of Official Analytical Chemists*, Volume 61, Issue 6, 1 November 1978, Pages 1404–1414, https://doi.org/10.1093/jaoac/61.6.1404



High Pressure Liquid Chromatographic Determination of the Intermediates/Side Reaction Products in FD&C Red No. 2 and FD&C Yellow No. 5 : Statistical Analysis of Instrument Response

Catherine J Bailey, Elizabeth A Cox, Janet A Springer ✉

Journal of Association of Official Analytical Chemists, Volume 61, Issue 6, 1 November 1978, Pages 1404–1414, https://doi.org/10.1093/jaoac/61.6.1404

Published: 22 January 2020

Catherine J Bailey, Elizabeth A Cox, Janet A Springer, High Pressure Liquid Chromatographic Determination of the Intermediates/Side Reaction Products in FD&C Red No. 2 and FD&C Yellow No. 5 : Statistical Analysis of Instrument Response, *Journal of Association of Official Analytical Chemists*, Volume 61, Issue 6, 1 November 1978, Pages 1404–1414, https://doi.org/10.1093/jaoac/61.6.1404

1. Curves are often fit to data as part of the process of calibrating instruments. For example, Bailey, Cox, and Springer (1978) discuss a method for measuring the concentrations of food dyes and other substances by high-pressure chromatography. Measurements of the chromatographic peak areas corresponding to sulfanilic acid were taken for several known concentrations of FD&C Yellow No. 5.

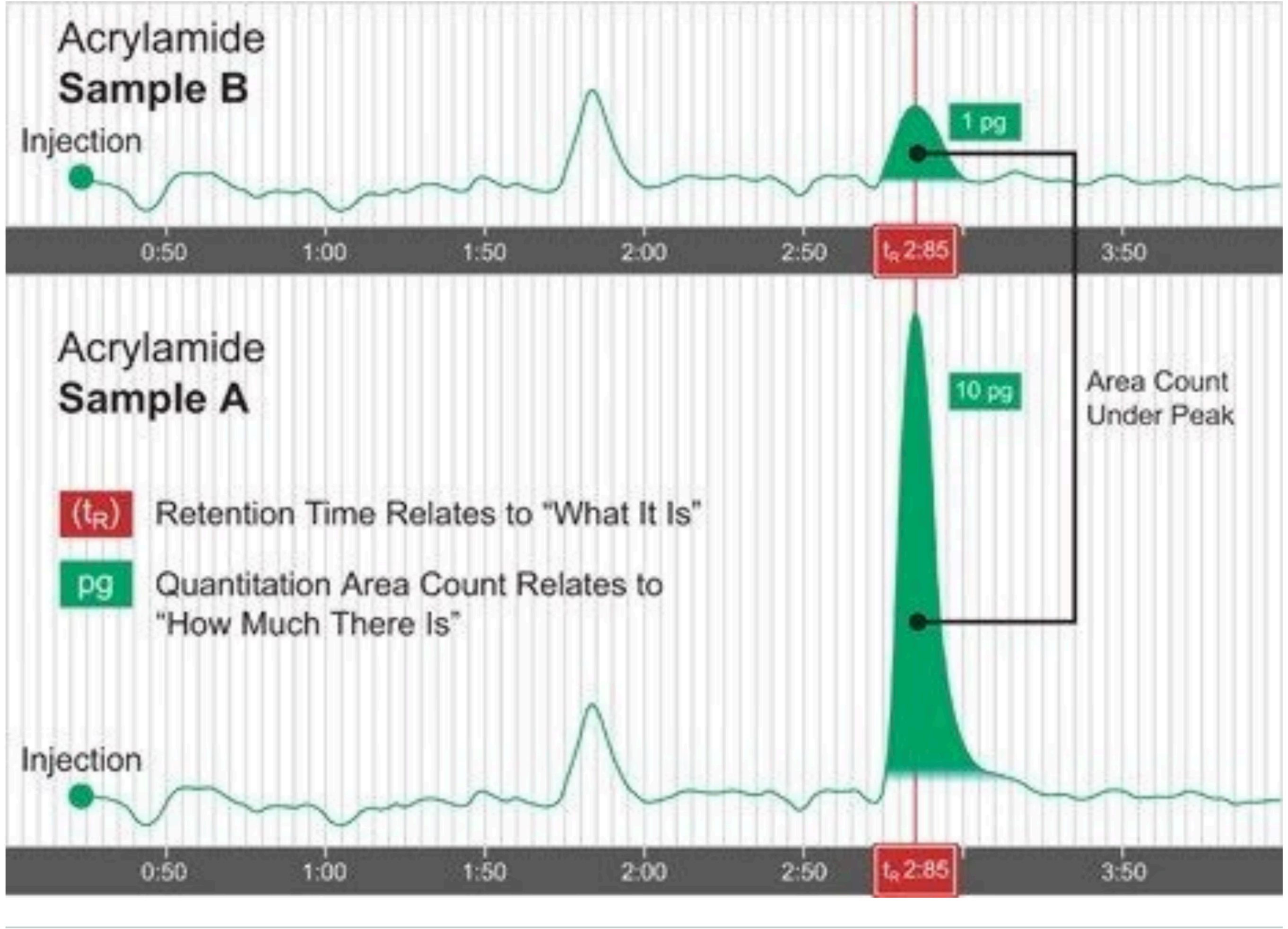


Figure I-2. Identification and Quantitation.

- Three **dye compounds are represented by** three **peaks** separated in time in the chromatogram.
- Each elutes at a specific location.
- Is the **area under the peak** linked to relative **amount of the dye**?

John Rice: Mathematical Statistics and Data Analysis, Duxbury Press
Chromatography figure source: https://www.waters.com/nextgen/us/en/education/primers/beginner-s-guide-to-liquid-chromatography/identifying-and-quantitating-compounds.html

1. Curves are often fit to data as part of the process of calibrating instruments. For example, Bailey, Cox, and Springer (1978) discuss a method for measuring the concentrations of food dyes and other substances by high-pressure chromatography. Measurements of the chromatographic peak areas corresponding to sulfanilic acid were taken for several known concentrations of FD&C Yellow No. 5.

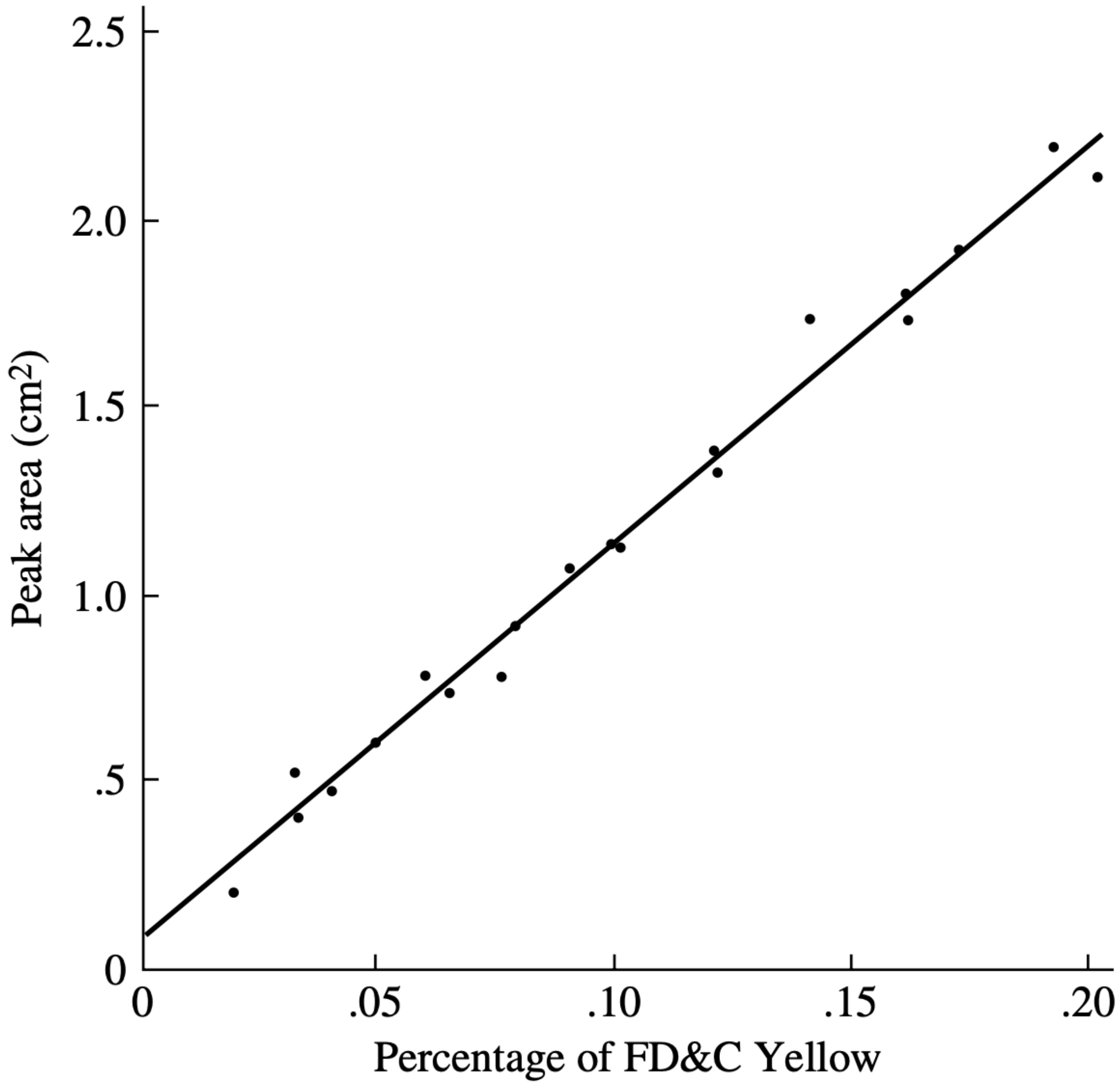


FIGURE **14.2** Data points and the least squares line for the relation of sulfanilic acid peak area to percentage of FD&C Yellow.

Independent variable:

percentage of DF&C yellow

Dependent variable (observed response):

peak area

Regression like looks like a very good fit!

John Rice: Mathematical Statistics and Data Analysis, Duxbury Press

1. Curves are often fit to data as part of the process of calibrating instruments. For example, Bailey, Cox, and Springer (1978) discuss a method for measuring the concentrations of food dyes and other substances by high-pressure chromatography. Measurements of the chromatographic peak areas corresponding to sulfanilic acid were taken for several known concentrations of FD&C Yellow No. 5.

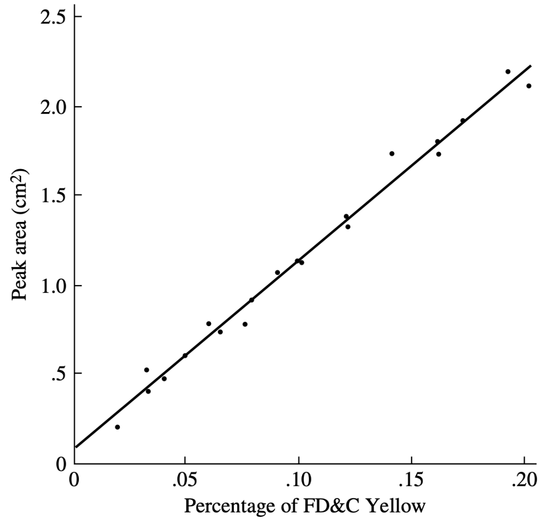
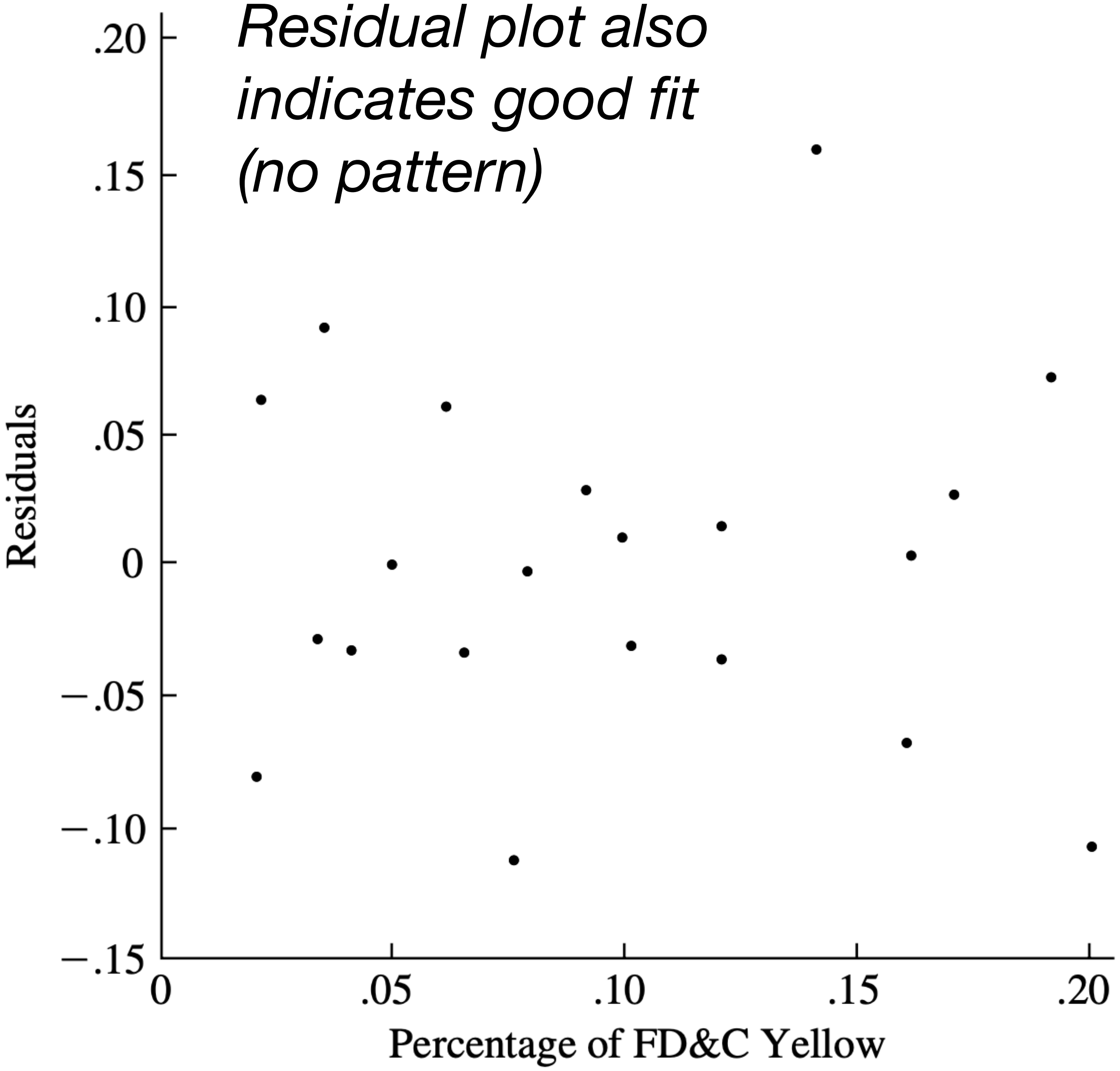


FIGURE **14.2** Data points and the least squares line for the relation of sulfanilic acid peak area to percentage of FD&C Yellow.

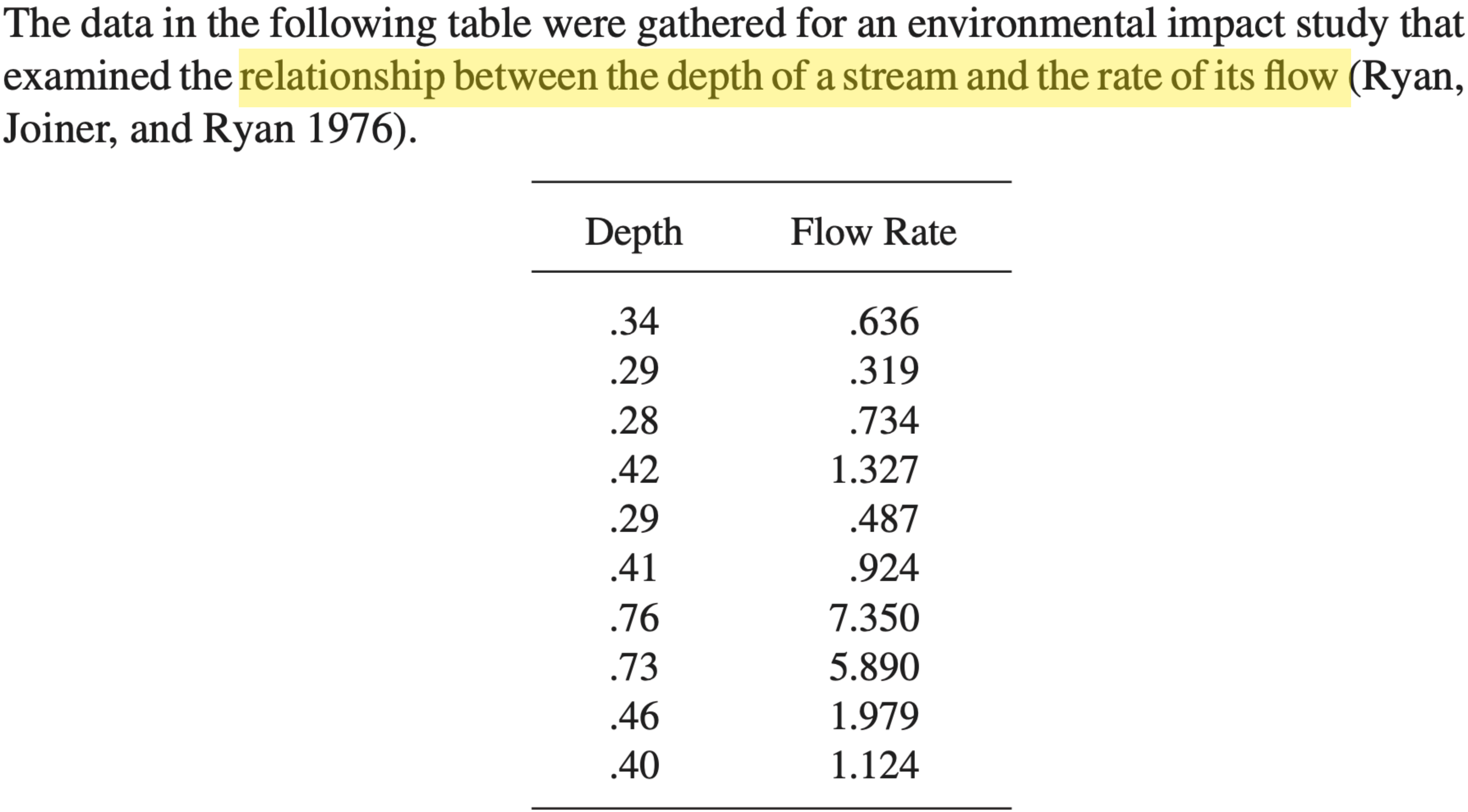


Residual plot also indicates good fit (no pattern)

FIGURE **14.5** A plot of residuals for the data on chromatographic peak area.

John Rice: Mathematical Statistics and Data Analysis, Duxbury Press

Regression in practice: 2. Environmental Sciences



The data in the following table were gathered for an environmental impact study that examined the relationship between the depth of a stream and the rate of its flow (Ryan, Joiner, and Ryan 1976).

Depth	Flow Rate
.34	.636
.29	.319
.28	.734
.42	1.327
.29	.487
.41	.924
.76	7.350
.73	5.890
.46	1.979
.40	1.124

John Rice: Mathematical Statistics and Data Analysis, Duxbury Press

The data in the following table were gathered for an environmental impact study that examined the relationship between the depth of a stream and the rate of its flow (Ryan, Joiner, and Ryan 1976).

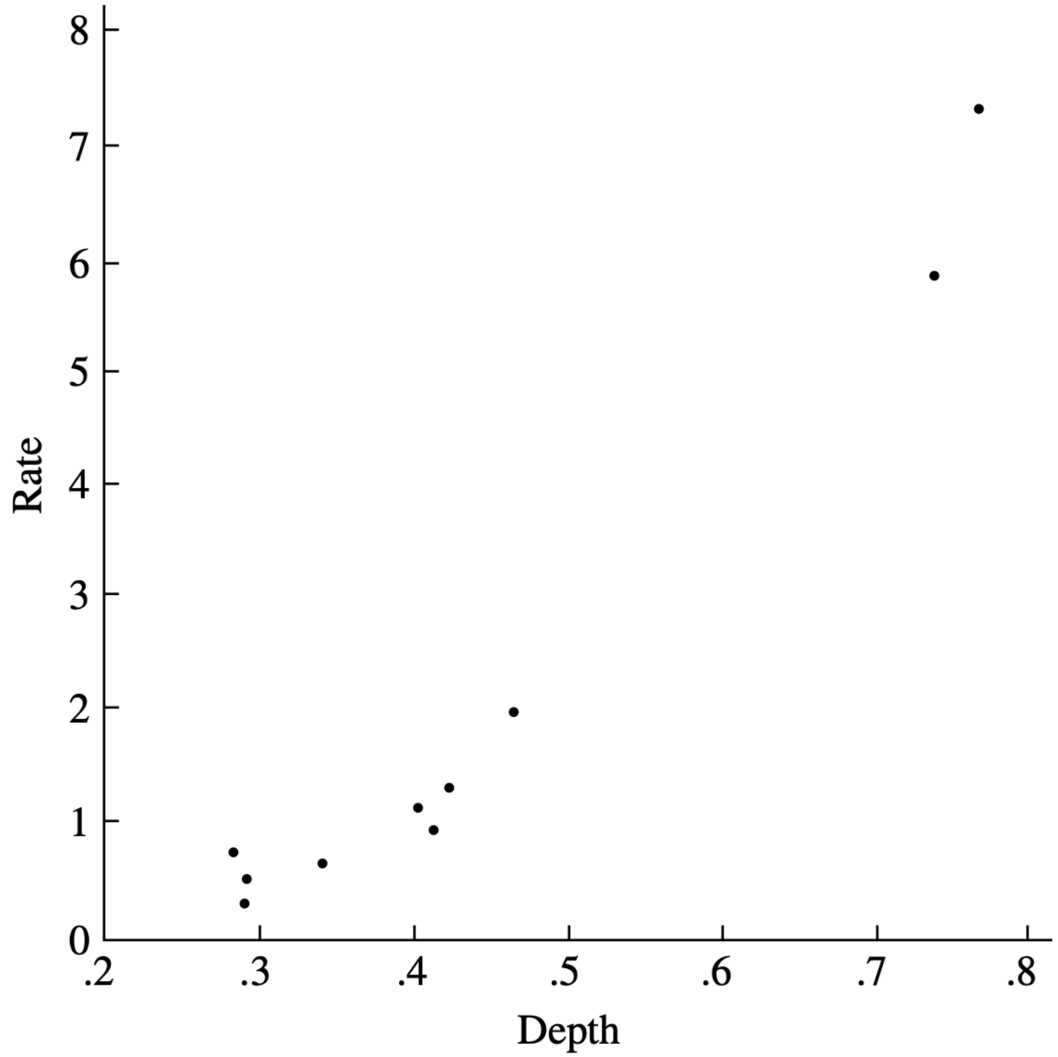


FIGURE 14.6 A plot of flow rate versus stream depth.

Independent variable:

Depth of stream

Dependent variable (observed response):

Flow rate

John Rice: Mathematical Statistics and Data Analysis, Duxbury Press

The data in the following table were gathered for an environmental impact study that examined the relationship between the depth of a stream and the rate of its flow (Ryan, Joiner, and Ryan 1976).

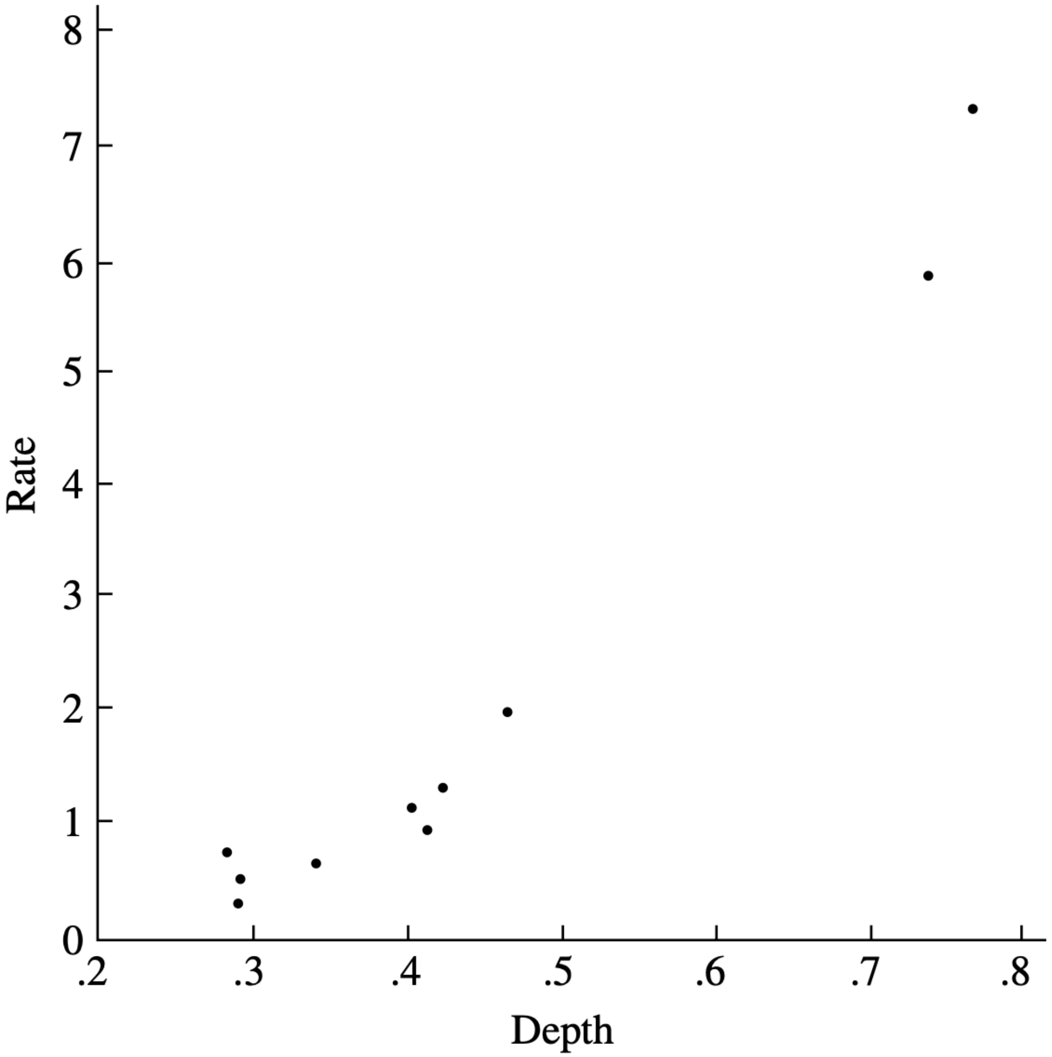


FIGURE 14.6 A plot of flow rate versus stream depth.

Independent variable:

Depth of stream

Dependent variable (observed response):

Flow rate

Does not look like a linear relationship...

John Rice: Mathematical Statistics and Data Analysis, Duxbury Press

The data in the following table were gathered for an environmental impact study that examined the relationship between the depth of a stream and the rate of its flow (Ryan, Joiner, and Ryan 1976).

A plot of flow rate versus depth suggests that the relation is not linear (Figure 14.6). This is even more immediately apparent from the bowed shape of the plot of the residuals versus depth (Figure 14.7).

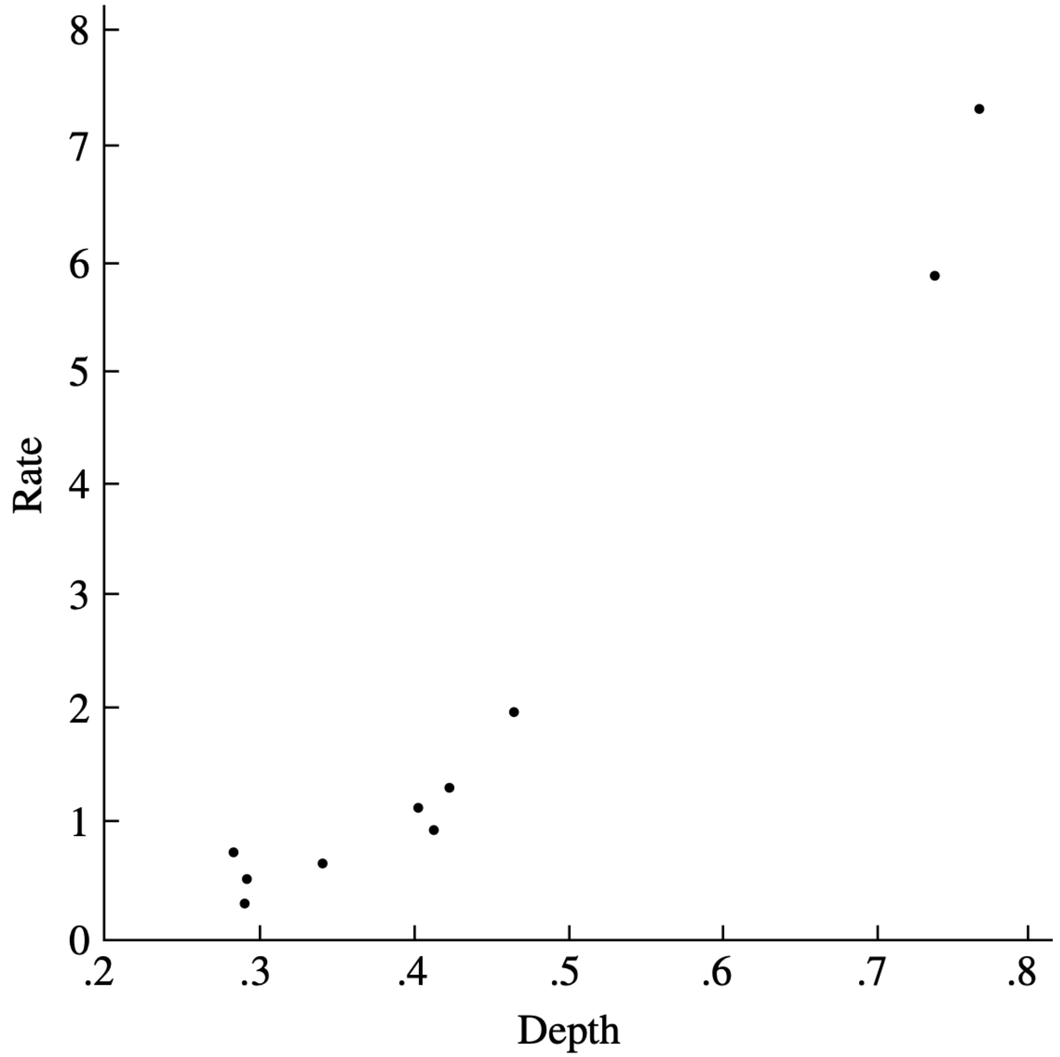
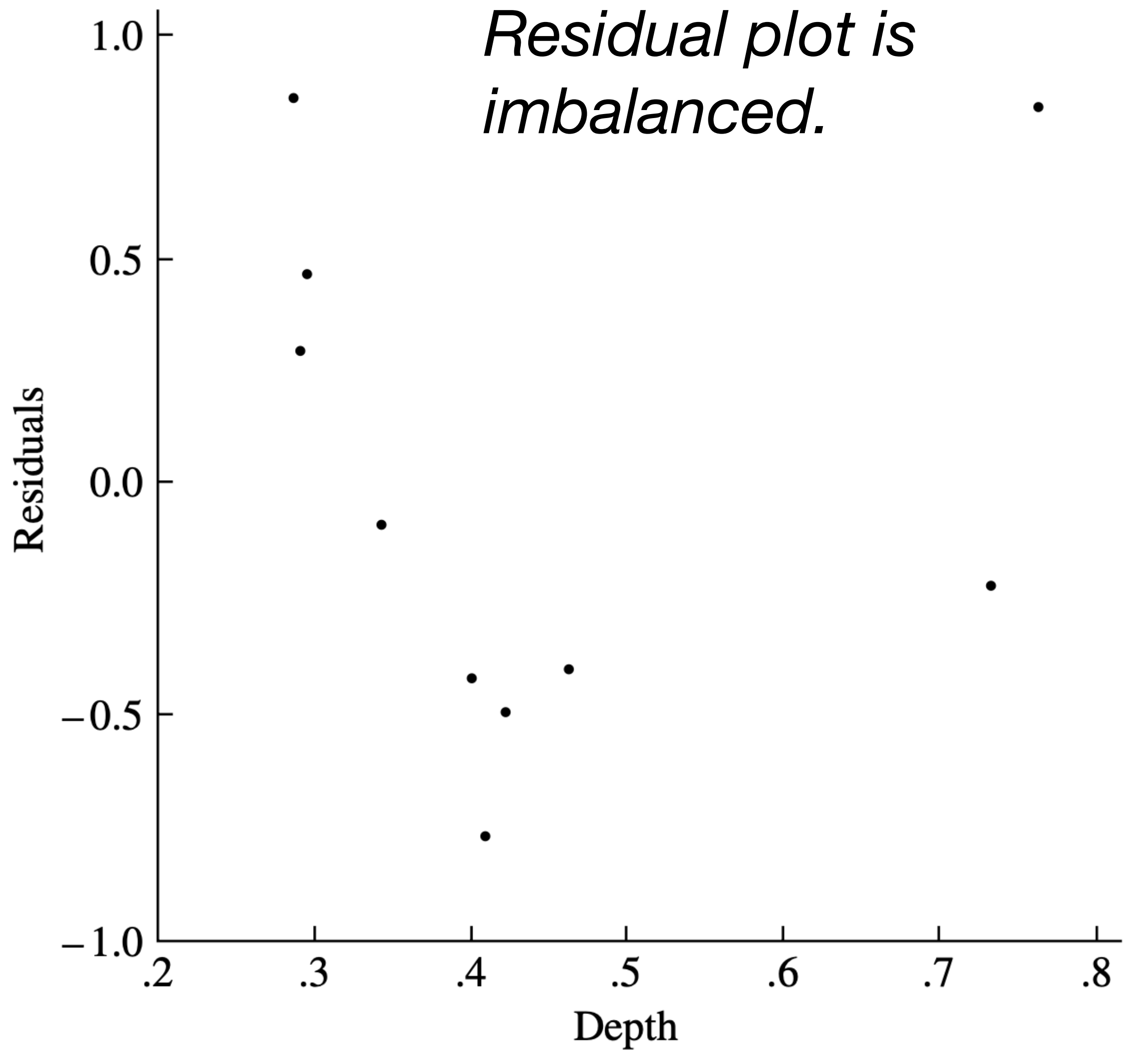


FIGURE **14.6** A plot of flow rate versus stream depth.



Residual plot is imbalanced.

FIGURE **14.7** Residuals from the regression of flow rate on depth.

John Rice: Mathematical Statistics and Data Analysis, Duxbury Press

What to do if we identify an issue?

These diagnostic plots are not a strict “go” or “stop” sign. It can tell you several things about the data.
You may want to rethink your model and hypotheses. You may want to:

- Transform variables
- Add new variables in the model
- Remove a few influential points
- Need better or different data collection methods, because of systematic bias in the data
- Possibly other things.

Use of log transforms: In order to empirically linearize relationships, transformations are frequently employed. Figure 14.8 is a plot of log rate versus log depth, and Figure 14.9 shows the residuals for the corresponding fit. There is no sign of obvious misfit.

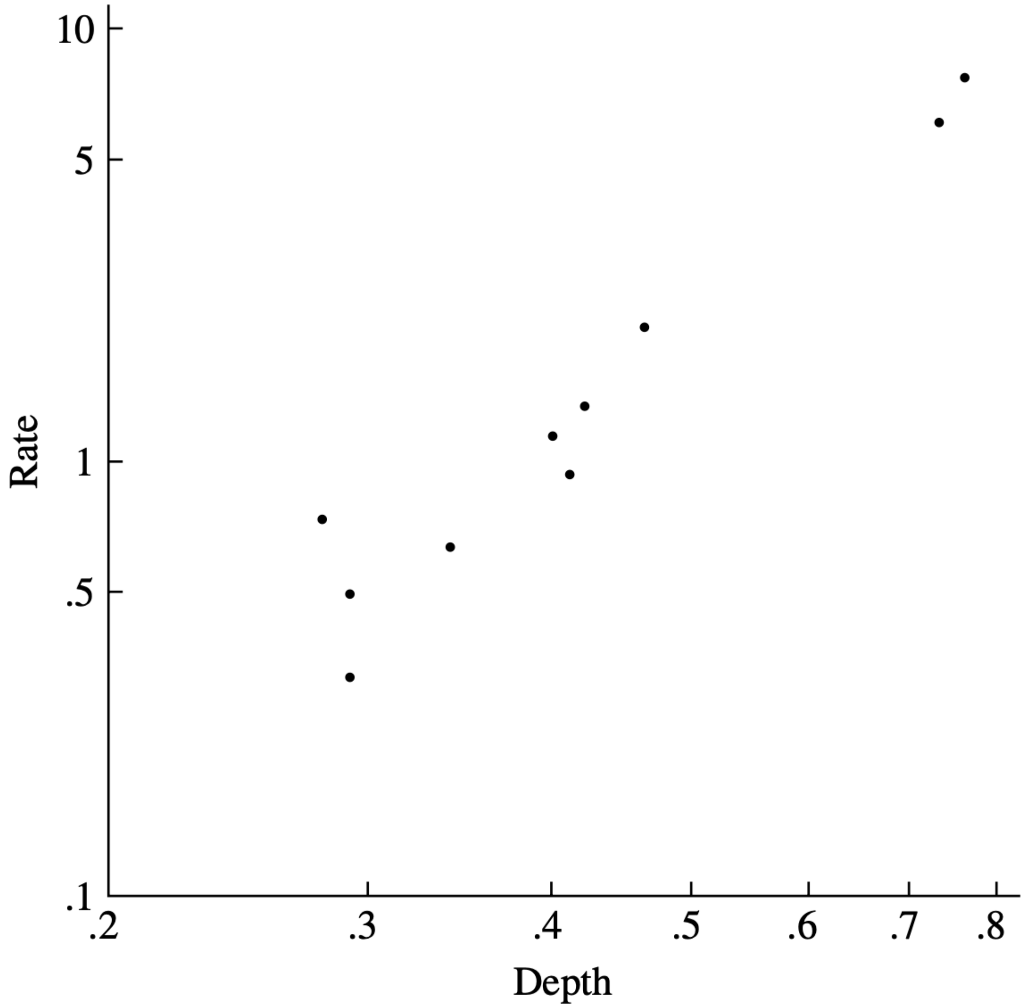


FIGURE **14.8** Plot of log flow rate versus log depth.

Residual plot indicates good fit (no pattern)

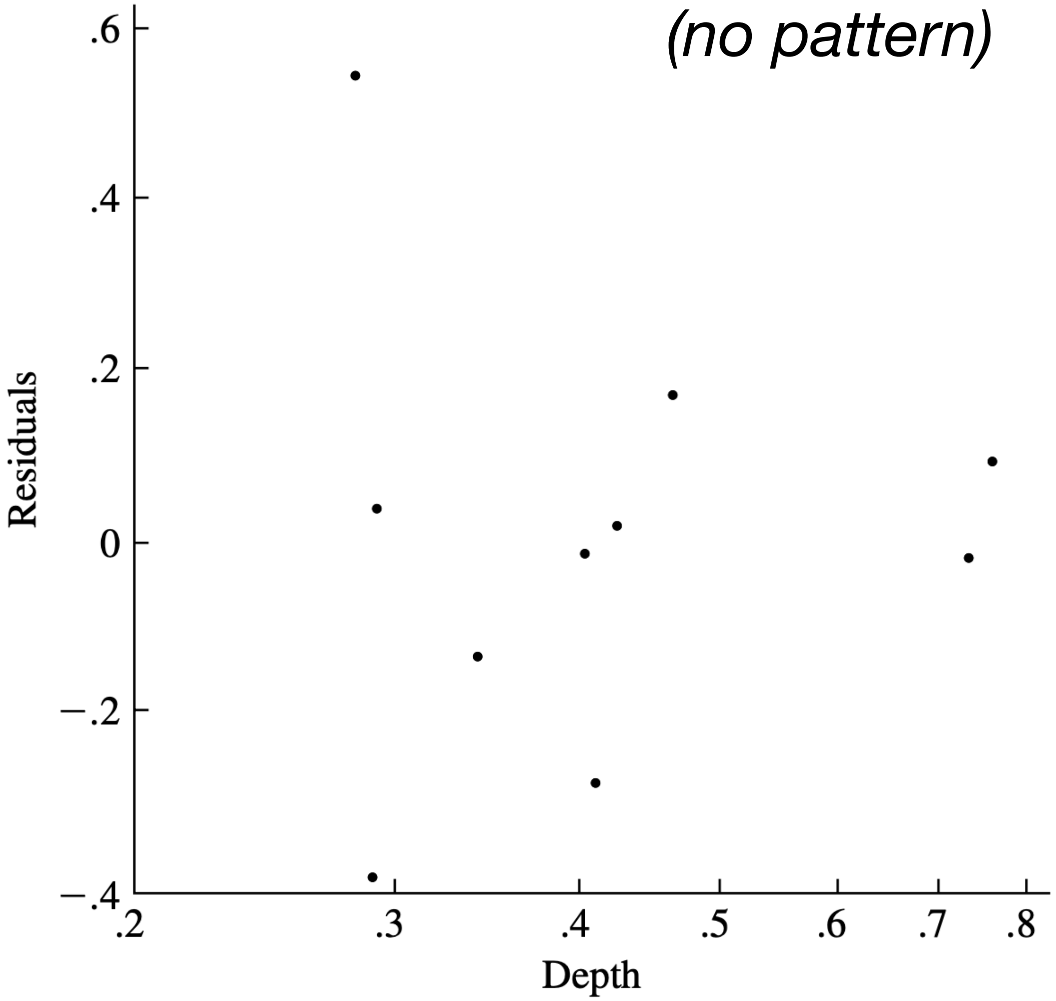


FIGURE **14.9** Residuals from the regression of log flow rate on log depth.

Regression in practice: 3. Cancer mortality

Dataset of population size and breast cancer mortality in 301 countries

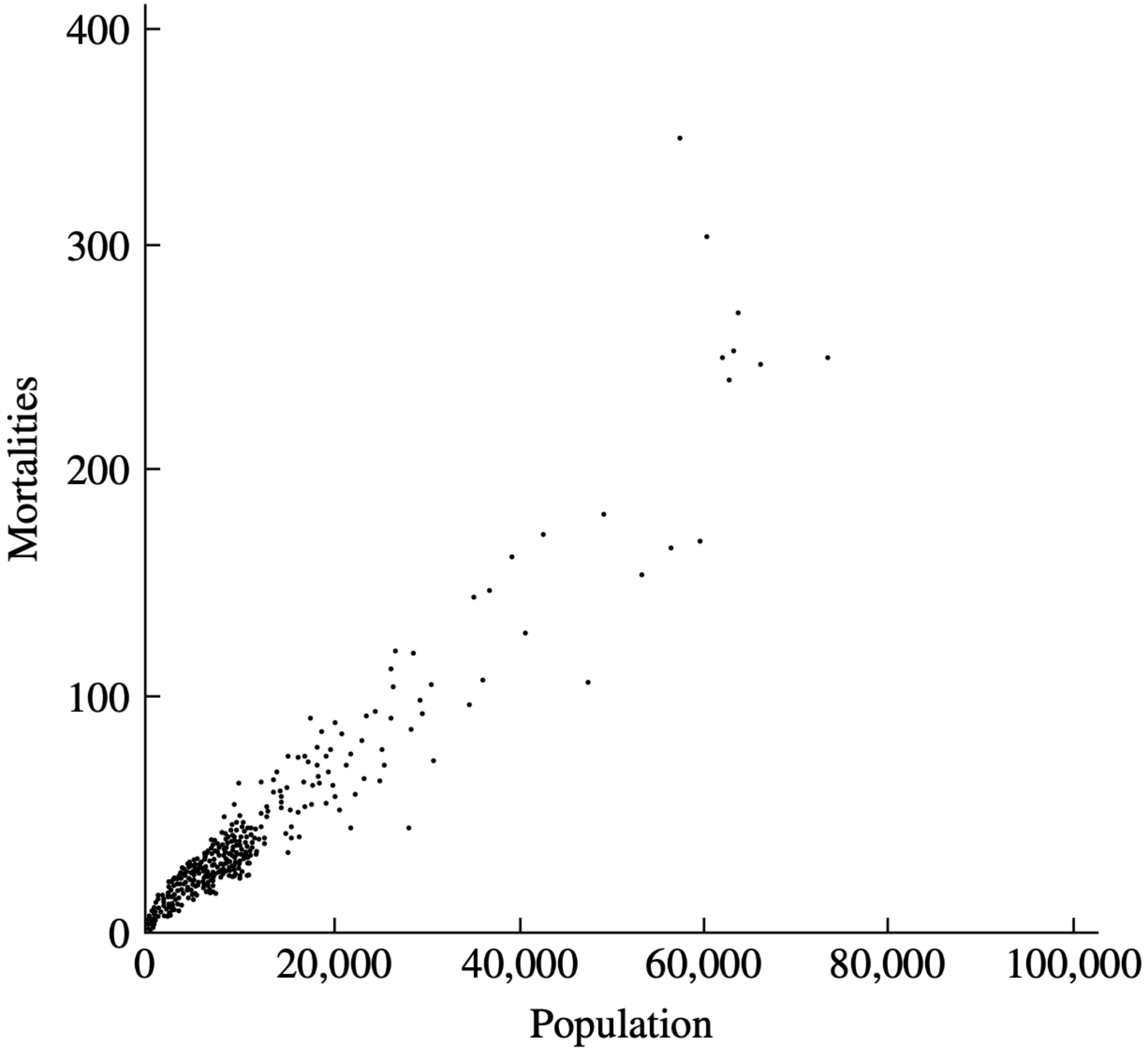


FIGURE **14.10** Scatterplot showing breast cancer mortality versus population.

Independent variable:

Population size

Dependent variable (observed response):

Mortalities

Can fit simple linear regression model:

- *intercept 0*
- *slope 3.56/10000*

John Rice: Mathematical Statistics and Data Analysis, Duxbury Press

Check model fit with residual plot….

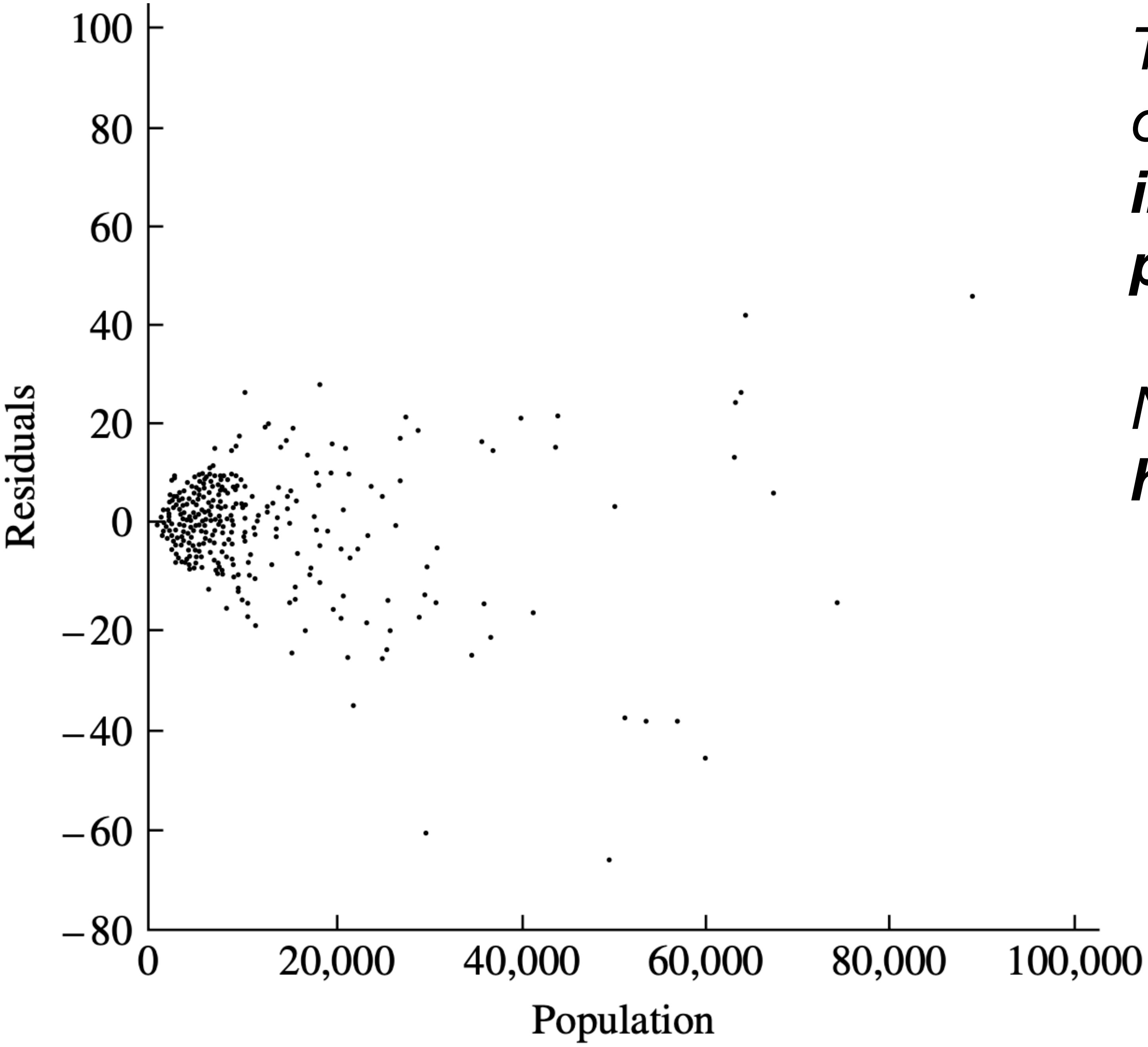


FIGURE **14.11** Residuals from the regression of mortality on population.

*The residual plot has no slope or other trends, but shows **increased variance as population grows**.*

*Non-constant variance is called **heteroscedasticity***

John Rice: Mathematical Statistics and Data Analysis, Duxbury Press

Check model fit with residual plot….

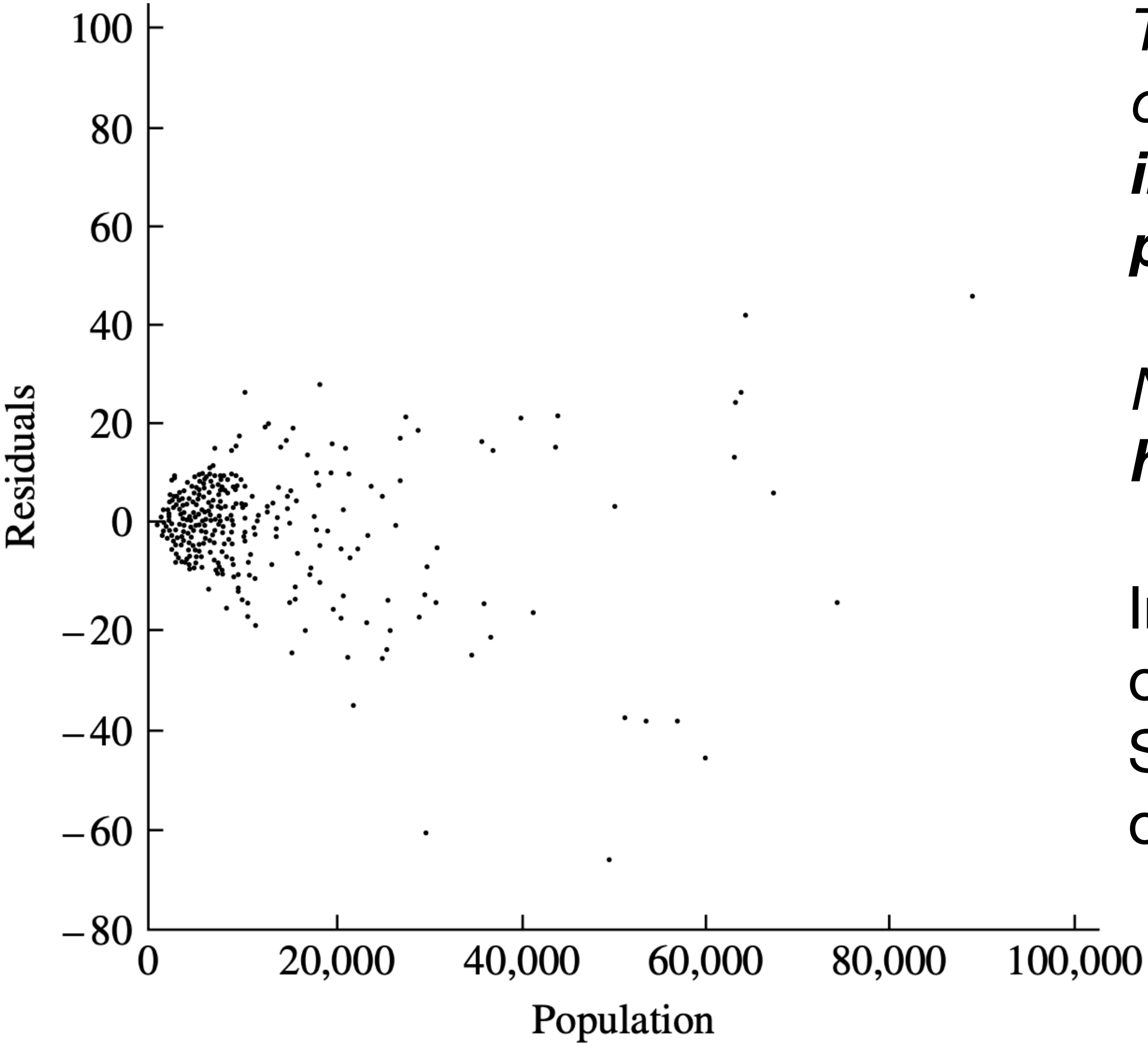


FIGURE **14.11** Residuals from the regression of mortality on population.

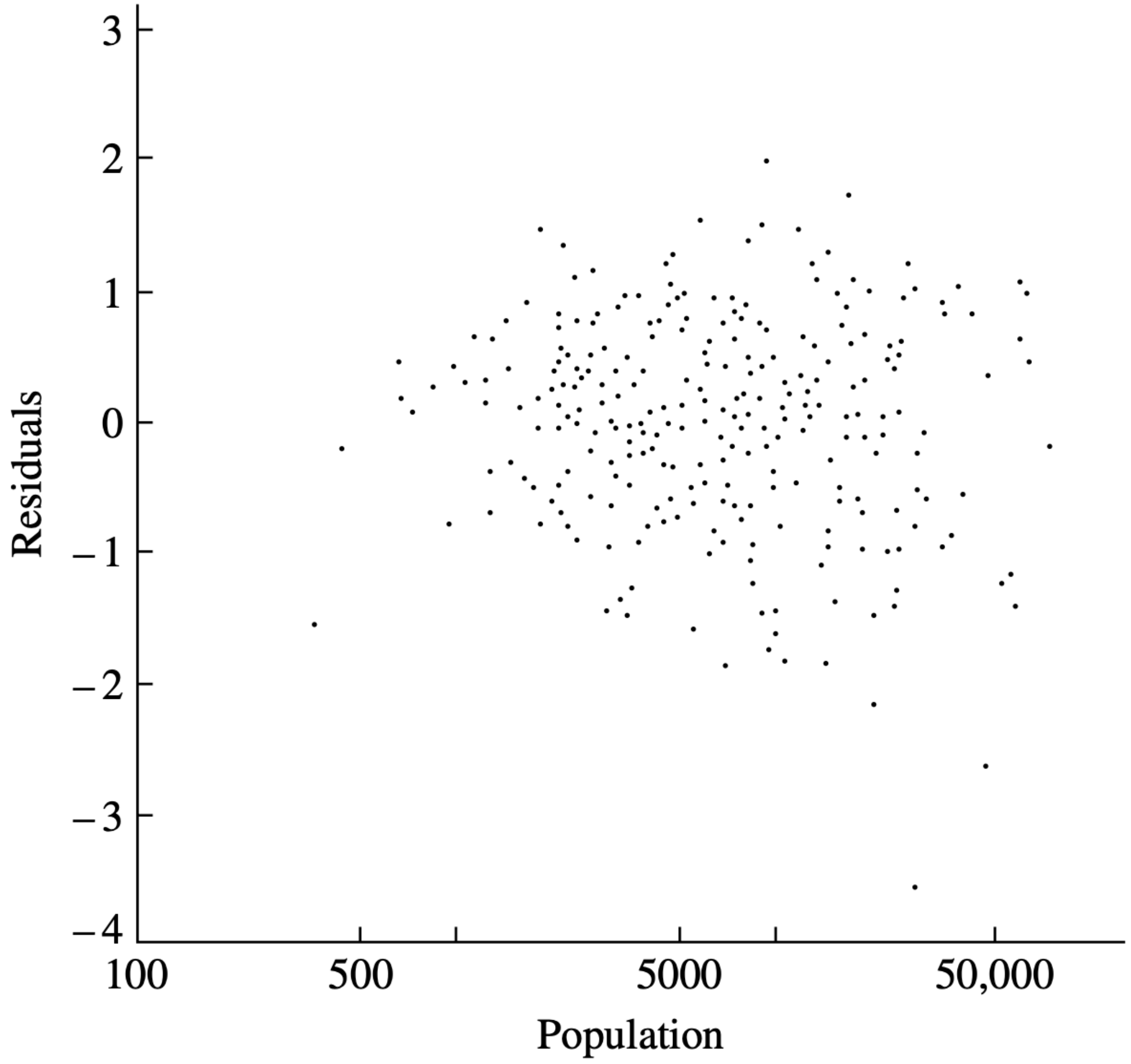
The residual plot has no slope or other trends, but shows ***increased variance as population grows****.*

Non-constant variance is called ***heteroscedasticity***

In count data the variance often growth with the mean. Square root transformations often stabilise the variance.

John Rice: Mathematical Statistics and Data Analysis, Duxbury Press

The residual plot in Figure 14.12 shows no curvature but indicates that the variance is not constant. For counted data, the variability often grows with the mean, and frequently a square root transformation is used in an attempt to stabilize the variance. We therefore fit a model of the form $\sqrt{y} \approx \gamma\sqrt{x}$. Figure 14.13 shows the plot of residuals for this fit. The residual variability is more nearly constant here; β is estimated by the square of the slope, $\hat{\gamma}$, which for this example gives $\tilde{\beta} = \hat{\gamma}^2 = 3.471 \times 10^{-3}$.



Residual plot has not slope or other trends, but shows increased variance as population grows.

FIGURE **14.13** Residuals from the regression of the square root of mortality on the square root of population.

John Rice: Mathematical Statistics and Data Analysis, Duxbury Press

ST 117

5. Context

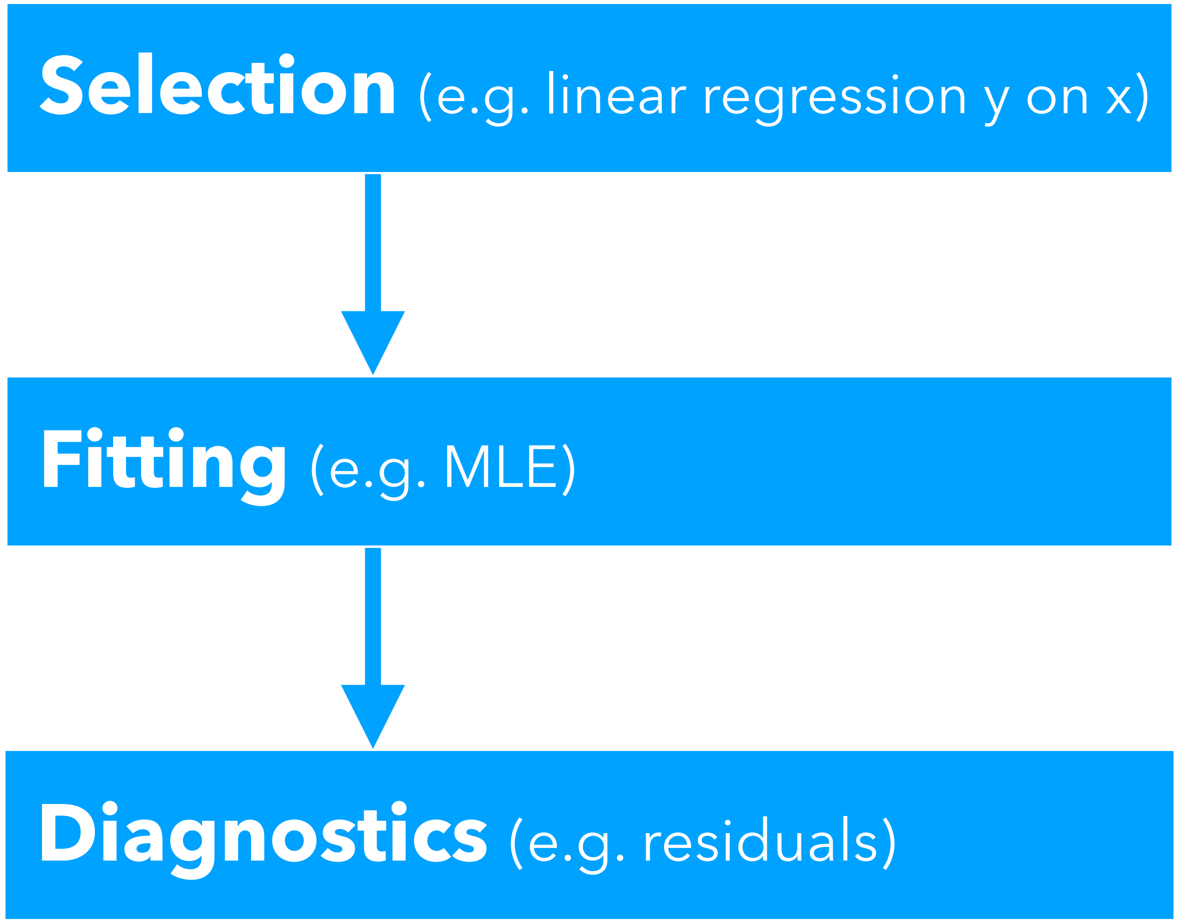
WARWICK

Lecture 26
(Week 9)

Model fit
Model improvement
Model choice
Beginnings of data science
Anscombe's quartet

Does my model fit?

A first simple model about statistical modelling



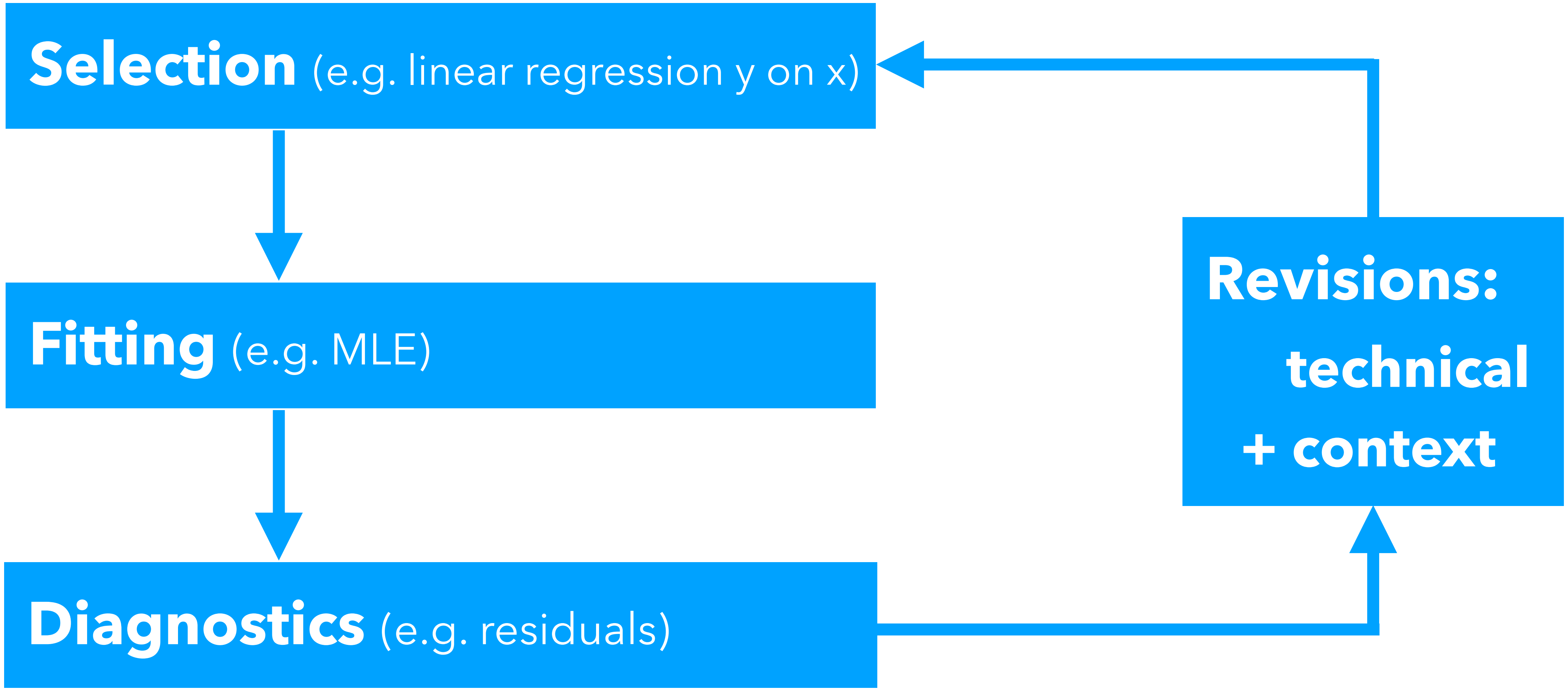
Selection (e.g. linear regression y on x)

Fitting (e.g. MLE)

Diagnostics (e.g. residuals)

Does my model fit and how can I improve it?

A first simple model about statistical modelling


Selection (e.g. linear regression y on x)
Fitting (e.g. MLE)
Diagnostics (e.g. residuals)
Revisions:
technical
+ context

How do we know if a model fits?

JOURNAL ARTICLE

The Examination and Analysis of Residuals

F. J. Anscombe and John W. Tukey

Technometrics

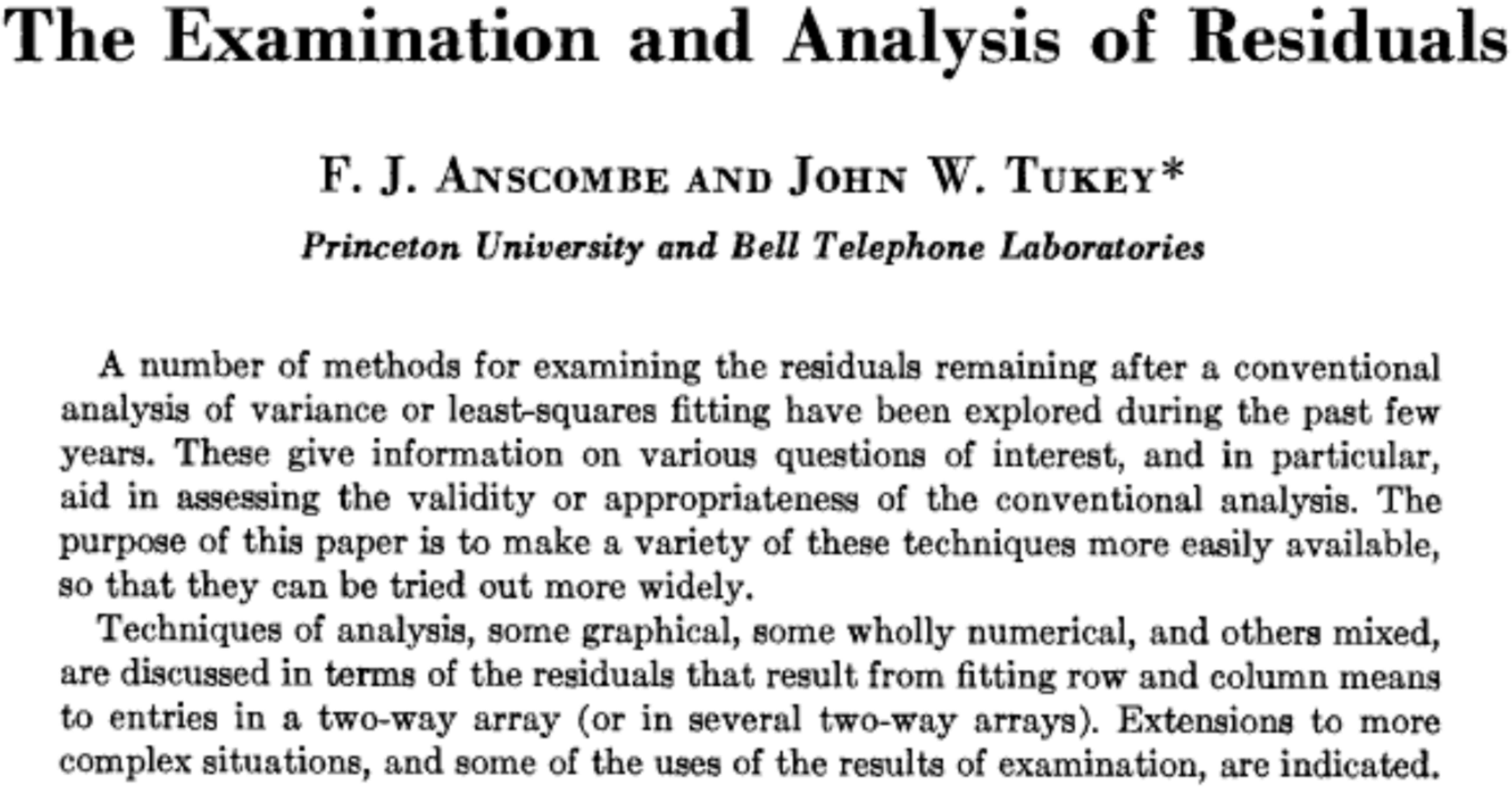
Vol. 5, No. 2 (May, 1963), pp. 141-160 (20 pages)

Published By: Taylor & Francis, Ltd.

https://doi.org/10.2307/1266059

https://www.jstor.org/stable/1266059

How do we know if a model fits?



The Examination and Analysis of Residuals

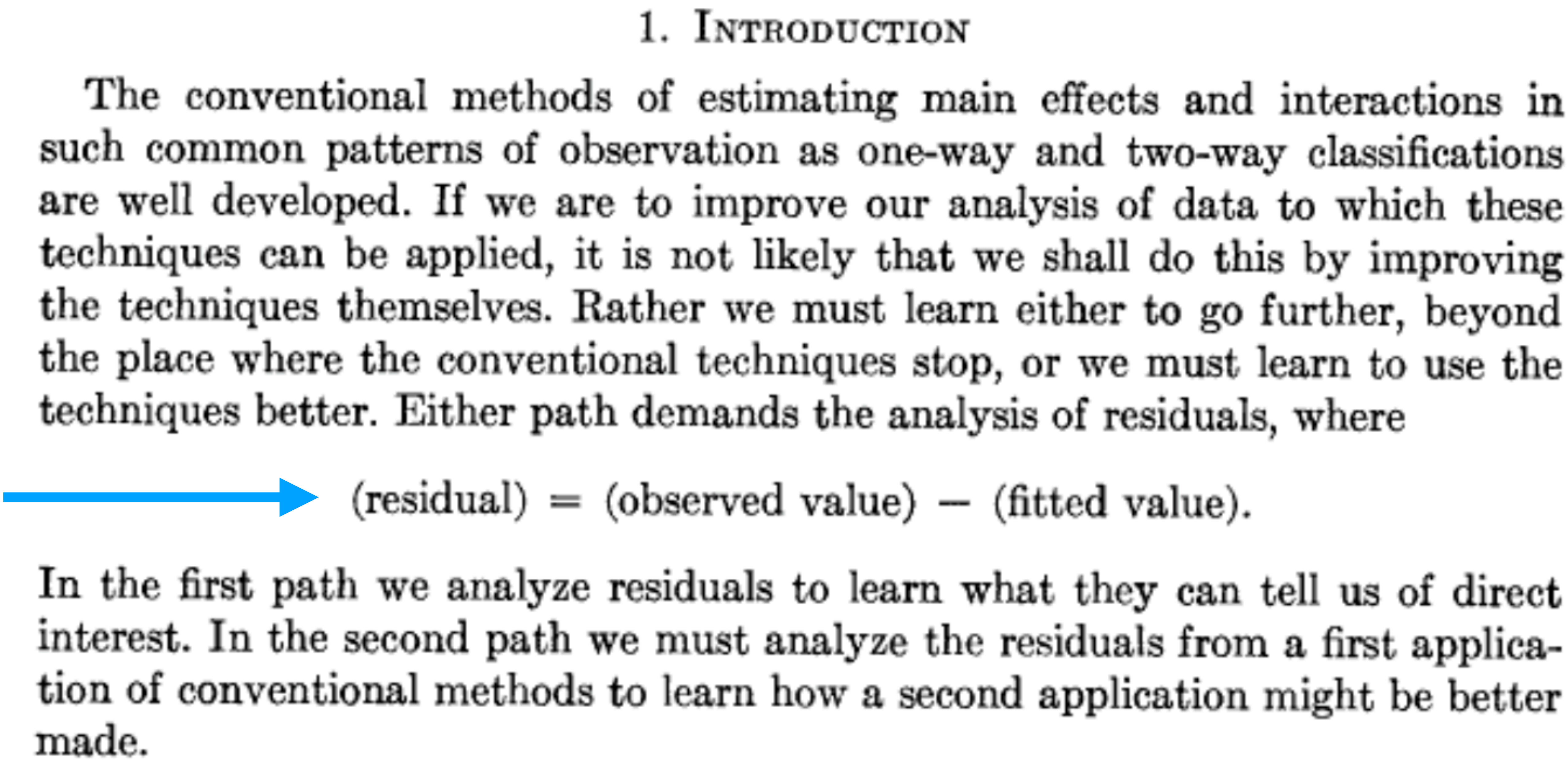
F. J. Anscombe and John W. Tukey*

Princeton University and Bell Telephone Laboratories

A number of methods for examining the residuals remaining after a conventional analysis of variance or least-squares fitting have been explored during the past few years. These give information on various questions of interest, and in particular, aid in assessing the validity or appropriateness of the conventional analysis. The purpose of this paper is to make a variety of these techniques more easily available, so that they can be tried out more widely.

Techniques of analysis, some graphical, some wholly numerical, and others mixed, are discussed in terms of the residuals that result from fitting row and column means to entries in a two-way array (or in several two-way arrays). Extensions to more complex situations, and some of the uses of the results of examination, are indicated.

How do we know if a model fits?



1. Introduction

The conventional methods of estimating main effects and interactions in such common patterns of observation as one-way and two-way classifications are well developed. If we are to improve our analysis of data to which these techniques can be applied, it is not likely that we shall do this by improving the techniques themselves. Rather we must learn either to go further, beyond the place where the conventional techniques stop, or we must learn to use the techniques better. Either path demands the analysis of residuals, where

$$(\text{residual}) = (\text{observed value}) - (\text{fitted value}).$$

In the first path we analyze residuals to learn what they can tell us of direct interest. In the second path we must analyze the residuals from a first application of conventional methods to learn how a second application might be better made.

Regression model and residuals

Bivariate data: $(x_i, y_i)\ (i = 1, \ldots, n)$

Model: $Y_i = \alpha + \beta x_i + \varepsilon_i$

Assumptions: ε_i i.i.d. $N(0, \sigma^2)$

Unknown parameters: α, β

Parameter estimates: $\hat{\alpha}, \hat{\beta}$

Fitted values: $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$

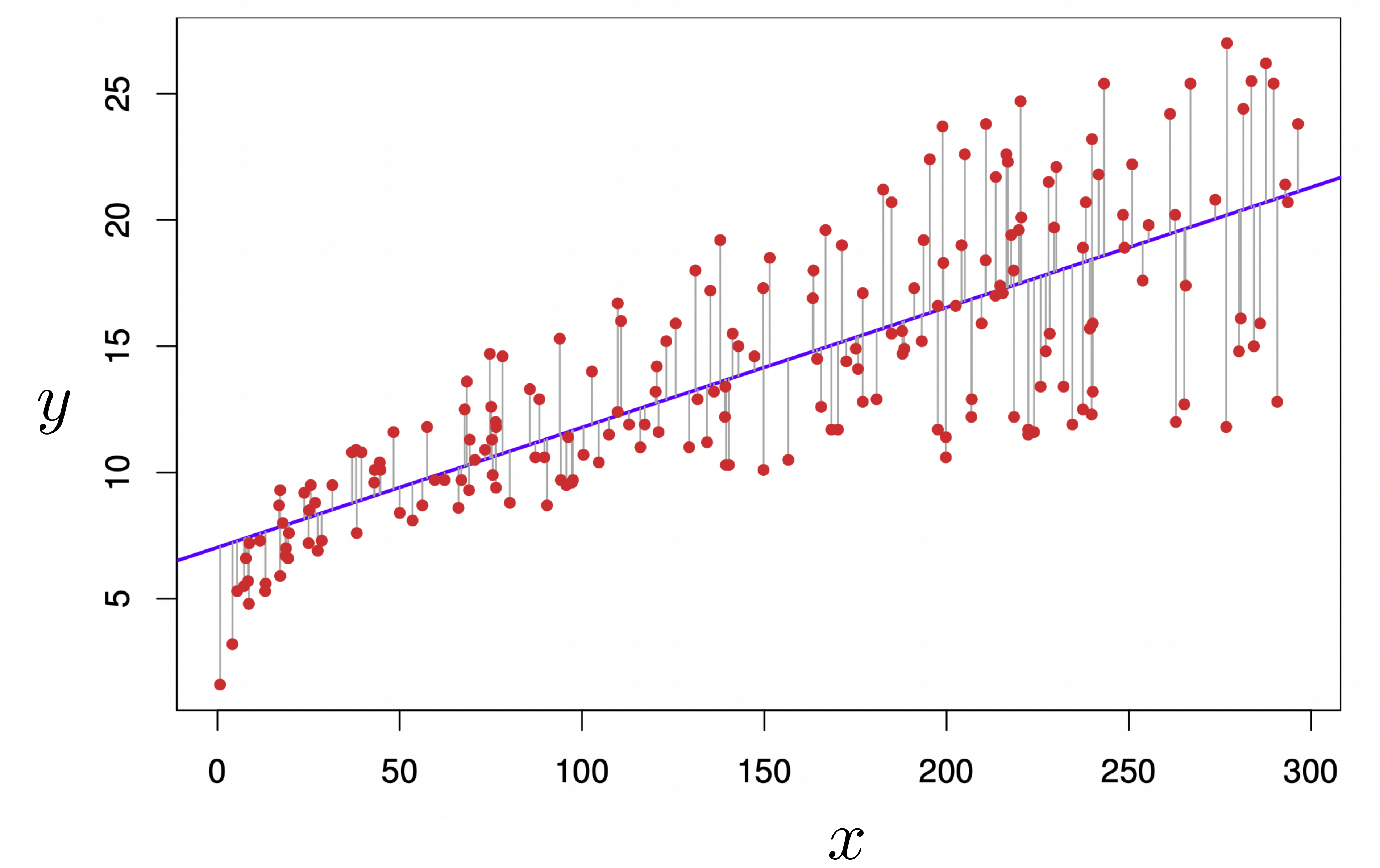
Residuals: $e_i = y_i - \hat{y}_i$

Residuals are the difference between observed values (data) and the model-based estimates.

Least squares estimator (same as MLE under given assumptions) minimises the residual sum of squares (RSS): $\sum_{i=1}^{n} e_i^2$

Source: https://hastie.su.domains/ISLR2/ISLRv2_website.pdf

Residuals



$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

$$e_i = y_i - \hat{y}_i$$

Least squares estimator (same as MLE under given assumptions) minimises the residual sum of squares (RSS): $\sum_{i=1}^{n} e_i^2$

Source: https://hastie.su.domains/ISLR2/ISLRv2_website.pdf

Residual plots

After defining a linear model `m`, if we type `plot(m)`, R gives us four plots (might have to press `ENTER` to make each new plot appear):

1. Residuals vs Fitted Values plot: This plot shows if residuals have non-linear patterns.
2. Normal Q-Q plot: This plot shows if residuals are normally distributed.
3. Scale-Location plot: This plot shows if residuals are spread equally along the ranges of predictors.
4. Residuals vs Leverage plot: This plot helps us to find influential cases.

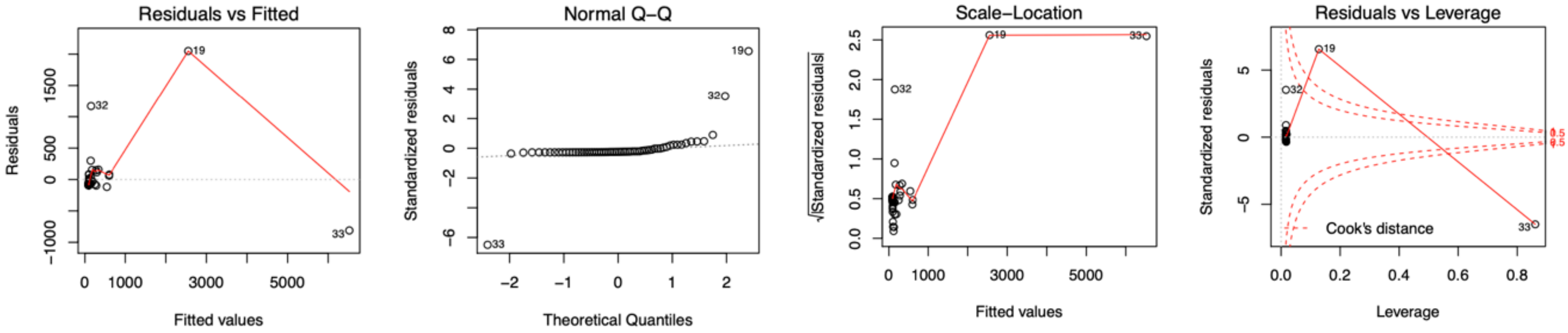
http://data.library.virginia.edu/diagnostic-plots/

Residual plots

After defining a linear model `m`, if we type `plot(m)`, `R` gives us four plots (might have to press `ENTER` to make each new plot appear):

1. Residuals vs Fitted Values plot: This plot shows if residuals have non-linear patterns.
2. Normal Q-Q plot: This plot shows if residuals are normally distributed.
3. Scale-Location plot: This plot shows if residuals are spread equally along the ranges of predictors.
4. Residuals vs Leverage plot: This plot helps us to find influential cases.

Material for following 4 slides taken from http://data.library.virginia.edu/diagnostic-plots/

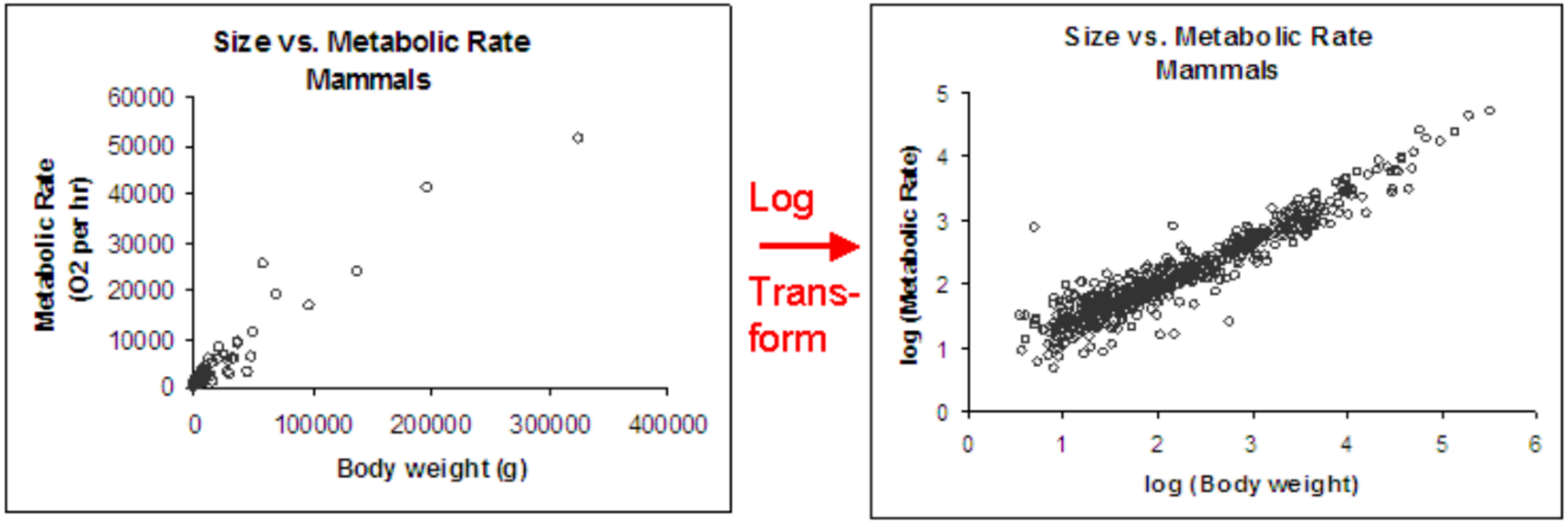


What can we do to improve a model fit?

- transform variables (with log, square root etc.)

Example: square root transformation to stability variance in international breast cancer mortality dataset (last week)

Example: metabolic rate vs body weight in 600 mammals



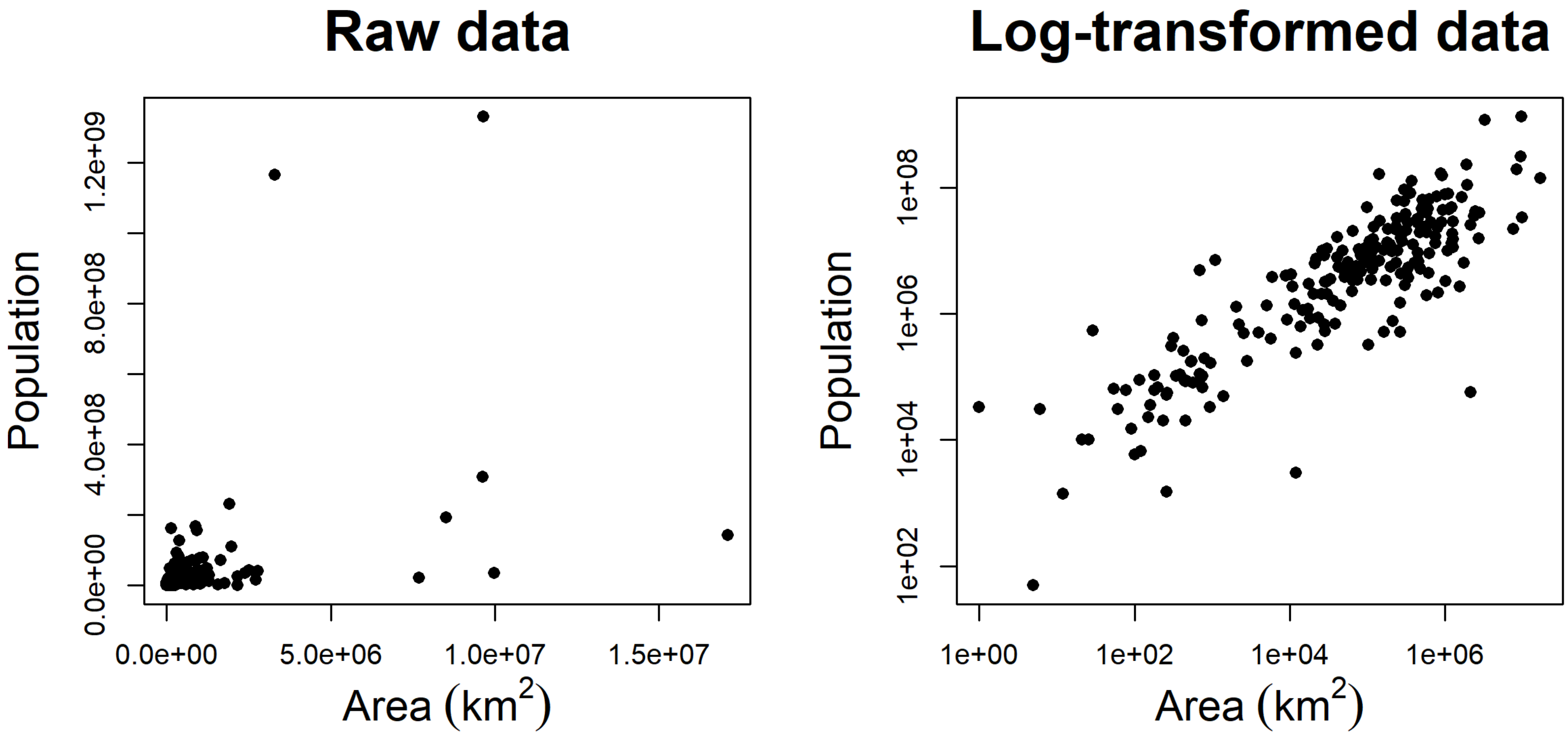
JM Bland, DG Altmann, Statistics Notes, MBJ
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2350481/pdf/bmj00534-0056.pdf

https://mathbench.umd.edu/modules/misc_scaling/page07.htm

What can we do to improve a model fit?

- transform variables (with log, square root etc.)

Example: area of a county and its populations



https://www.davidzeleny.net/anadat-r/doku.php/en:data_preparation

What can we do to improve a model fit?

- transform variables (with log, square root etc.)
  - e.g. by log square root, reciprocal…
  - more suitable scale to deal with large ranges
  - variance stabilisation (when closer to normal distribution, independence of mean and variance)
  - require care for interpretation of results

JM Bland, DG Altmann, Statistics Notes, MBJ
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2350481/pdf/bmj00534-0056.pdf

https://mathbench.umd.edu/modules/misc_scaling/page07.htm

What can we do to improve a model fit?

- transform variables (with log, square root etc.)
- identify and consider confounding variables
    - discrete data: fit separate regression lines to subsets of the dataset stratified by the confounding variable

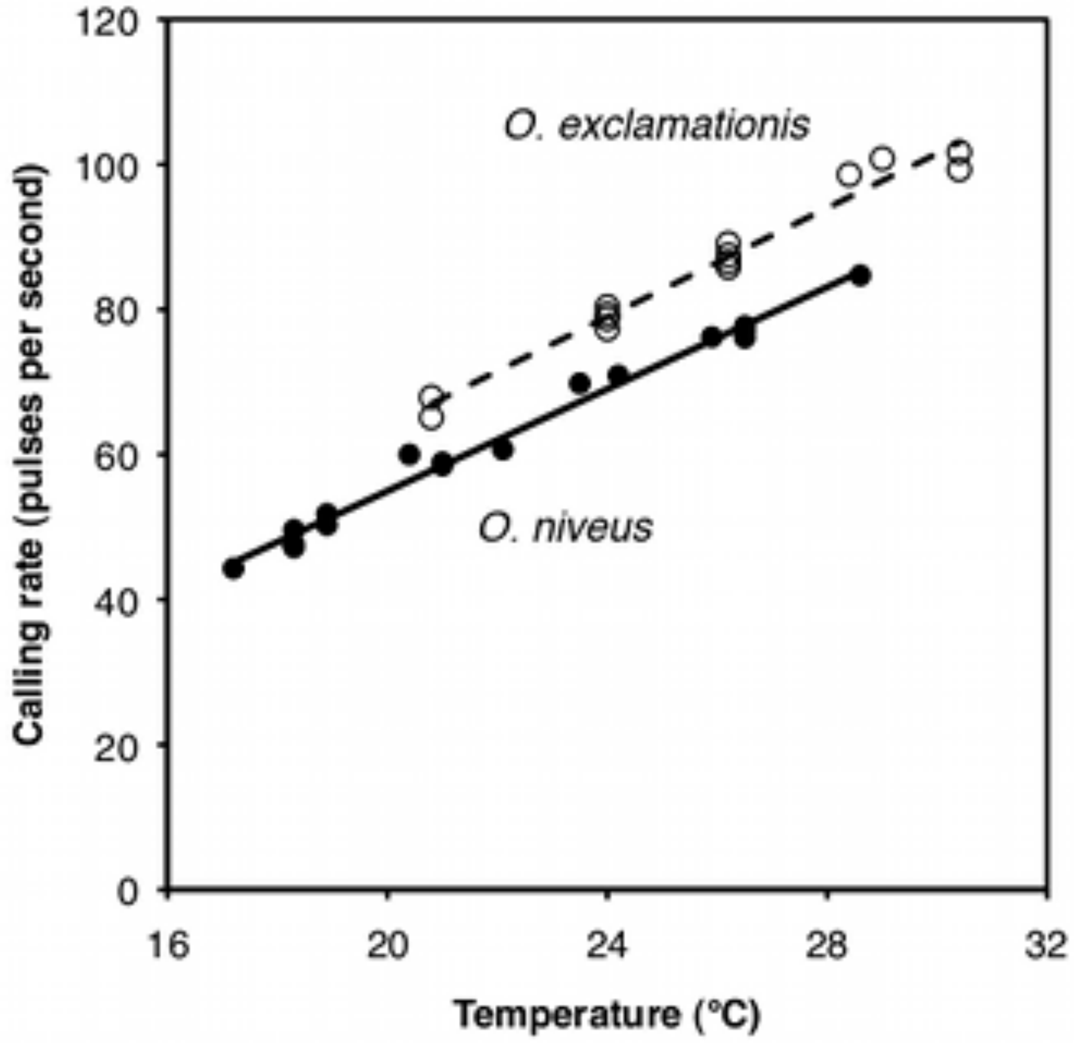
What can we do to improve a model fit?

- transform variables (with log, square root etc.)
- identify and consider confounding variables

Example: calling rate and temperature in two cricket species

Background:

Walker (1962) studied the mating songs of male tree crickets. Each wingstroke by a cricket produces a pulse of song, and females may use the number of pulses per second to identify males of the correct species. Walker (1962) wanted to know whether the chirps of the crickets *Oecanthus exclamationis* and *Oecanthus niveus* had different pulse rates. He measured the pulse rate of the crickets at a variety of temperatures.



JH McDonald, Handbook of Biological Statistics,
http://www.biostathandbook.com/ancova.html

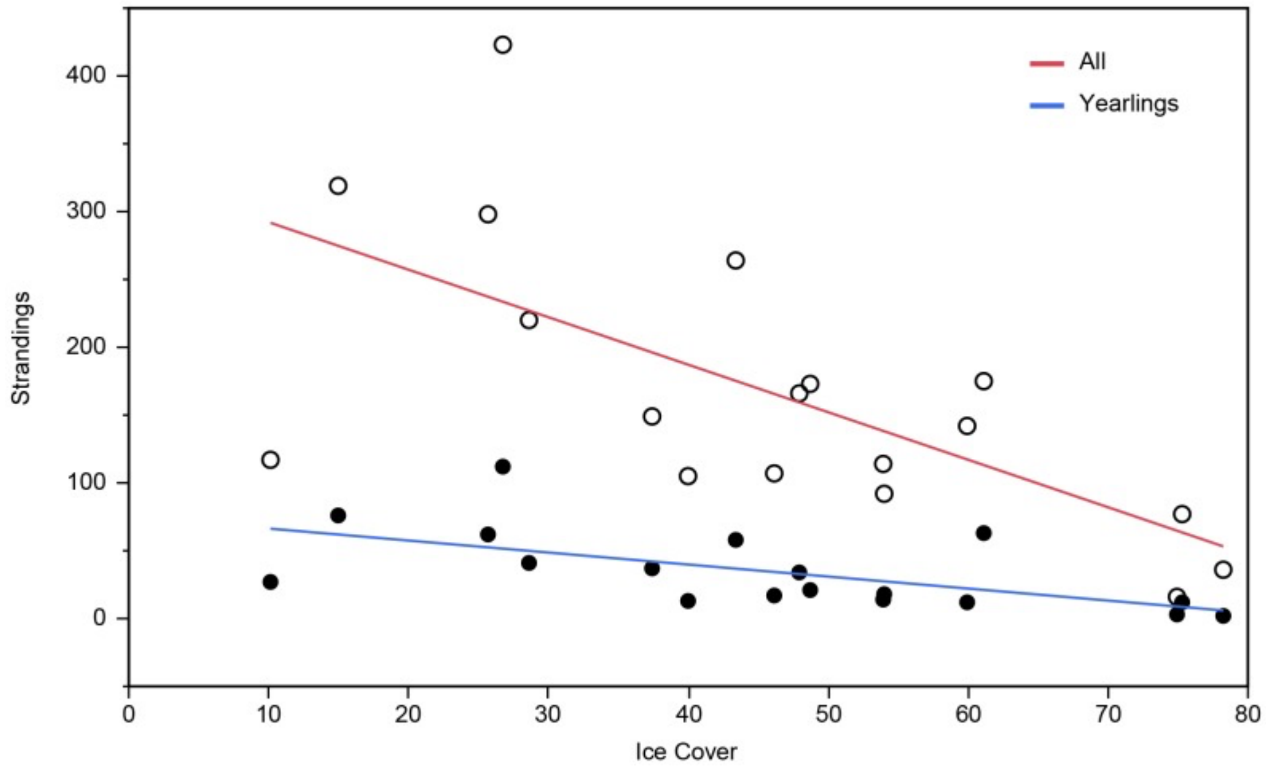
What can we do to improve a model fit?

- transform variables (with log, square root etc.)
- identify and consider confounding variables

Example: effect of ice cover on seal standings

Background:

Research about Factors Affecting Harp Seal (Pagophilus groenlandicus) Strandings in the Northwest Atlantic. Stranding: Seals and sea lions (pinnipeds) are considered stranded when they are found dead on land or in the water, or are in need of medical attention.



Linear regression of the present dataset (total strandings) and (dead yearling strandings). Percent sea ice cover and total strandings (open circles and red line) and percent sea ice cover and dead yearling strandings (solid circles and blue line).

BK Soulen et al, PLoS ONE 8(7):e68779
http://www.biostathandbook.com/ancova.html

What can we do to improve a model fit?

- transform variables (with log, square root etc.)
- identify and consider confounding variables
    - discrete data: fit separate regression lines to subsets of the dataset stratified by the confounding variable
    - continuous data (simple version): binning and use above
    - Continuous data (general approach): including them into the model

What can we do to improve a model fit?

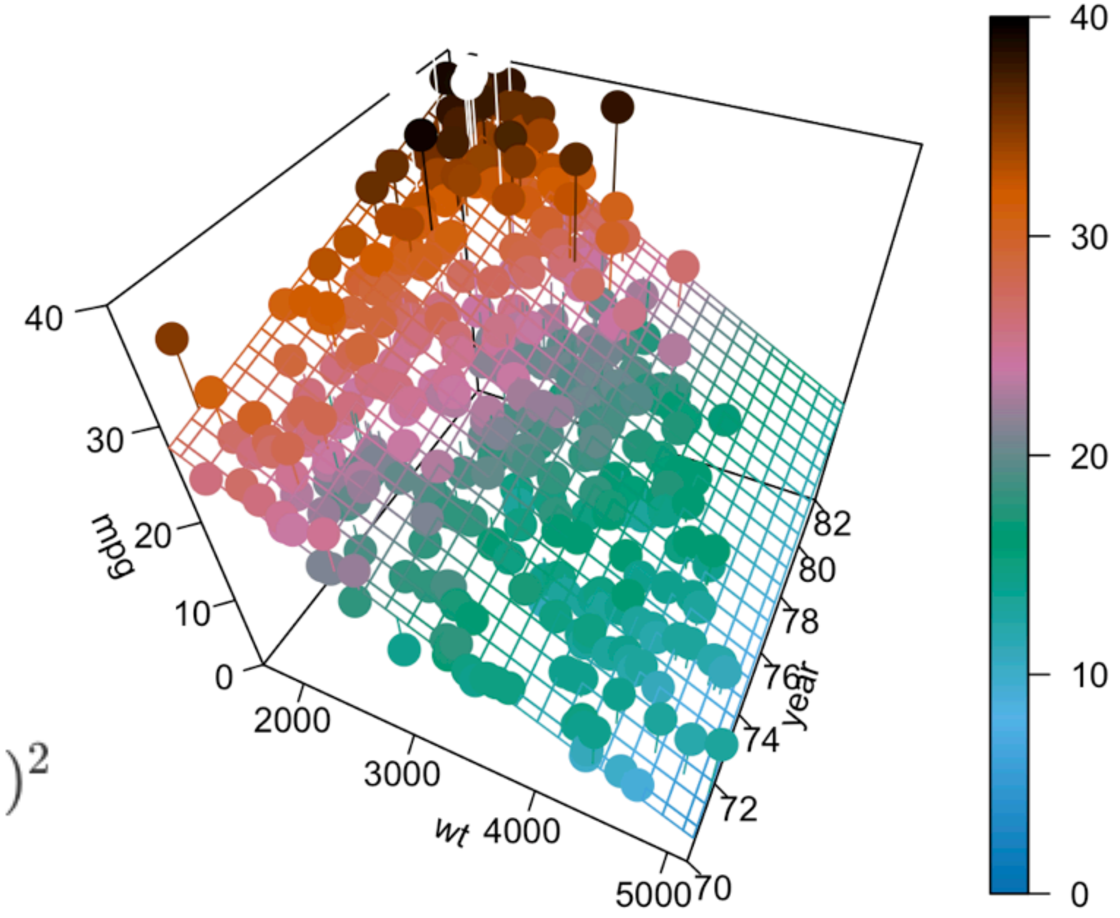
- transform variables (with log, square root etc.)
- identify and consider confounding variables
  - Continuous data (general approach): including them into the model (multivariate regression, see Second Year)

Task:

Fit a plane using two or more independent variables to predict the dependent variable y

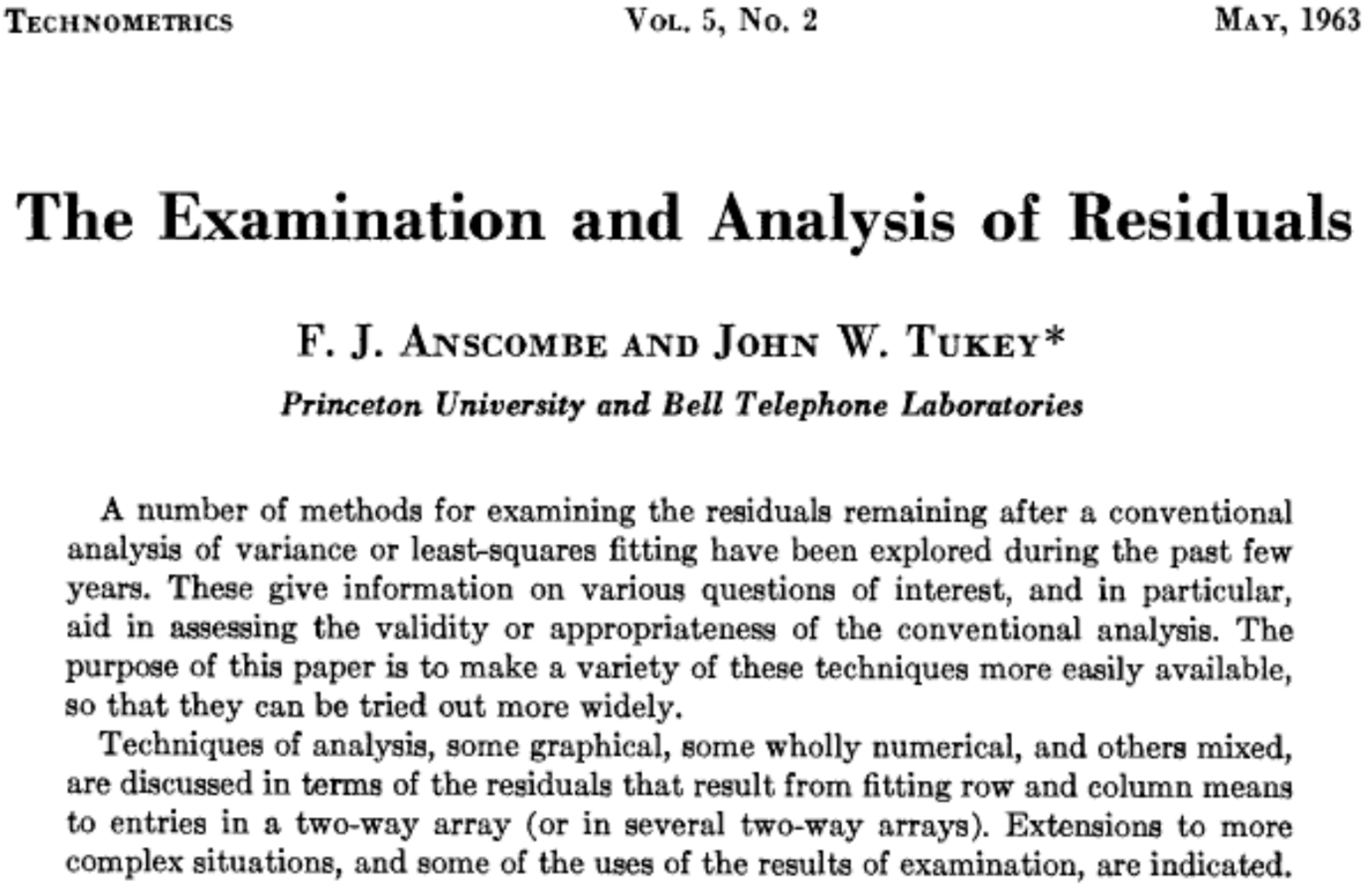
minimize

$$f(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}))^2$$



Applied Statistics with R
https://book.stat420.org/multiple-linear-regression.html

How do we know if a model fits?



TECHNOMETRICS VOL. 5, NO. 2 MAY, 1963

The Examination and Analysis of Residuals

F. J. ANSCOMBE AND JOHN W. TUKEY*

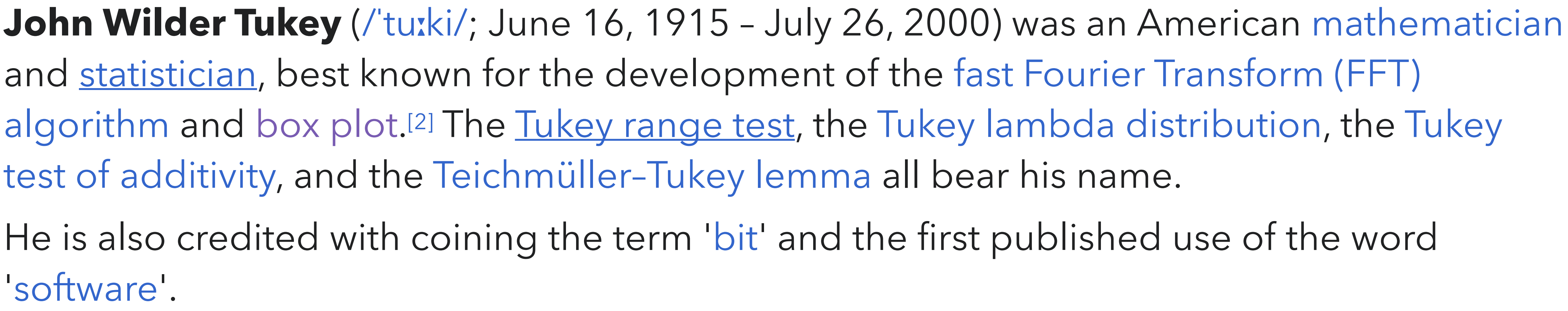
Princeton University and Bell Telephone Laboratories

A number of methods for examining the residuals remaining after a conventional analysis of variance or least-squares fitting have been explored during the past few years. These give information on various questions of interest, and in particular, aid in assessing the validity or appropriateness of the conventional analysis. The purpose of this paper is to make a variety of these techniques more easily available, so that they can be tried out more widely.

Techniques of analysis, some graphical, some wholly numerical, and others mixed, are discussed in terms of the residuals that result from fitting row and column means to entries in a two-way array (or in several two-way arrays). Extensions to more complex situations, and some of the uses of the results of examination, are indicated.

John Wilder Tukey

Seen as the father of data science



John Wilder Tukey (/ˈtuːki/; June 16, 1915 – July 26, 2000) was an American mathematician and statistician, best known for the development of the fast Fourier Transform (FFT) algorithm and box plot.[2] The Tukey range test, the Tukey lambda distribution, the Tukey test of additivity, and the Teichmüller–Tukey lemma all bear his name.

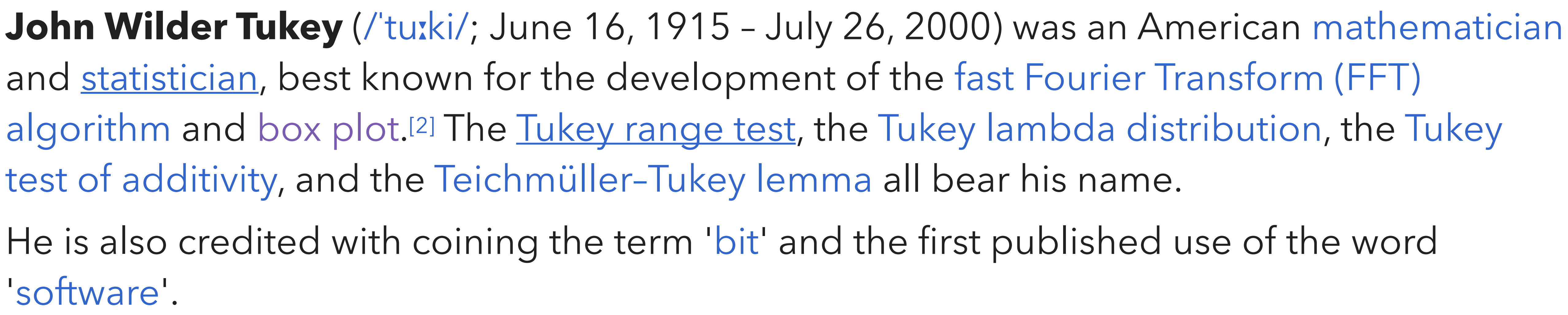
He is also credited with coining the term 'bit' and the first published use of the word 'software'.

- Mathematician, Statistician, who is known for many(!) things including
- UG in Chemistry, PhD in Topology (Princeton, 1939)
- Fire Control Research office during WWII
- invented fast Fourier transform, Teichmüller-Tukey lemma
- the terms “bit” and “software”
- one of the founders of data science (seminal paper in the 1960s)
- exploratory data analysis (EDA) which can be understood as precursor to data science, philosophy and tools (boxplot, median polish, MvA plots)
- Bell Labs, Full Professor and Founding Chairman at Princeton Statistics Department

https://www.stat.berkeley.edu/~brill/Papers/life.pdf
https://en.wikipedia.org/wiki/John_Tukey

John Wilder Tukey

Seen as the father of data science



John Wilder Tukey (/ˈtuːki/; June 16, 1915 – July 26, 2000) was an American mathematician and statistician, best known for the development of the fast Fourier Transform (FFT) algorithm and box plot.[2] The Tukey range test, the Tukey lambda distribution, the Tukey test of additivity, and the Teichmüller–Tukey lemma all bear his name.

He is also credited with coining the term 'bit' and the first published use of the word 'software'.

> *"The best thing about **being a statistician** is that you **get to play in everyone else's backyard.**"*

https://www.stat.berkeley.edu/~brill/Papers/life.pdf
https://en.wikipedia.org/wiki/John_Tukey

John Wilder Tukey

In **"The Future of Data Analysis" (1962)**:

"For a long time I thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and doubt…

I have come to feel that my central interest is in data analysis…

*Data analysis, and the parts of statistics which adhere to it, must…take on the characteristics of science rather than those of mathematics… **data analysis is intrinsically an empirical science…**"*

A very short history of data science
https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/

Applied mathematics as bridge

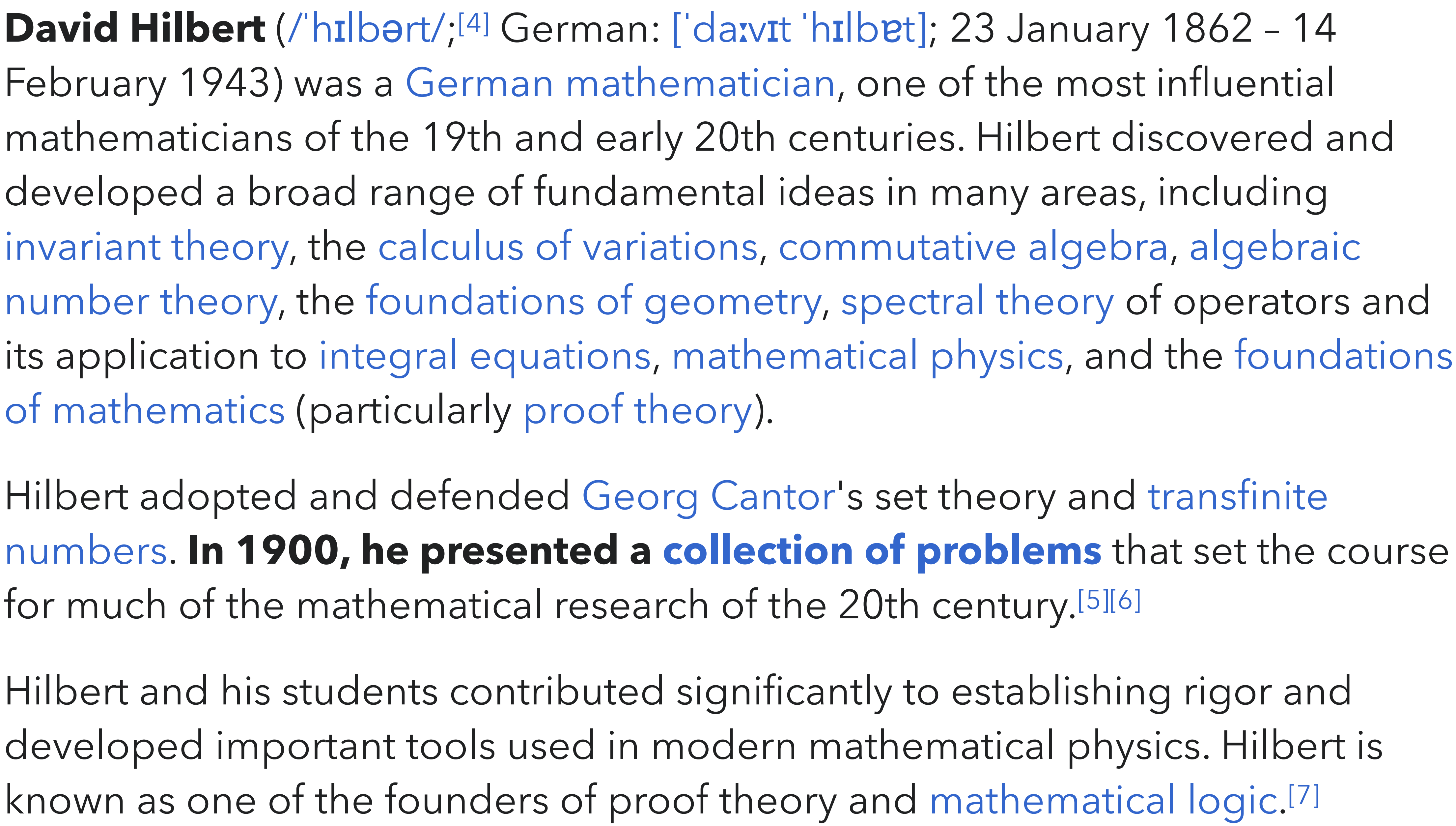
> *"The **instrument that mediates between theory and practice**, between thought and observation, is mathematics; it builds the **connecting bridge** and makes it stronger and stronger. Thus it happens that our entire present-day culture, insofar as it rests on intellectual insight into and harnessing of nature, is founded on mathematics."*

David Hilbert

In Königsberg on 8 September 1930, David Hilbert addressed the yearly meeting of the Society of German Natural Scientists and Physicians (Gesellschaft der Deutschen Naturforscher und Ärzte). Full text of the speech in English and German at url below, including audio file [*]

[*] http://math.sfsu.edu/smith/Documents/HilbertRadio/HilbertRadio.pdf

David Hilbert



David Hilbert (/ˈhɪlbərt/;[4] German: [ˈdaːvɪt ˈhɪlbɐt]; 23 January 1862 – 14 February 1943) was a German mathematician, one of the most influential mathematicians of the 19th and early 20th centuries. Hilbert discovered and developed a broad range of fundamental ideas in many areas, including invariant theory, the calculus of variations, commutative algebra, algebraic number theory, the foundations of geometry, spectral theory of operators and its application to integral equations, mathematical physics, and the foundations of mathematics (particularly proof theory).

Hilbert adopted and defended Georg Cantor's set theory and transfinite numbers. **In 1900, he presented a collection of problems** that set the course for much of the mathematical research of the 20th century.[5][6]

Hilbert and his students contributed significantly to establishing rigor and developed important tools used in modern mathematical physics. Hilbert is known as one of the founders of proof theory and mathematical logic.[7]

https://en.wikipedia.org/wiki/David_Hilbert

Hilbert’s problems

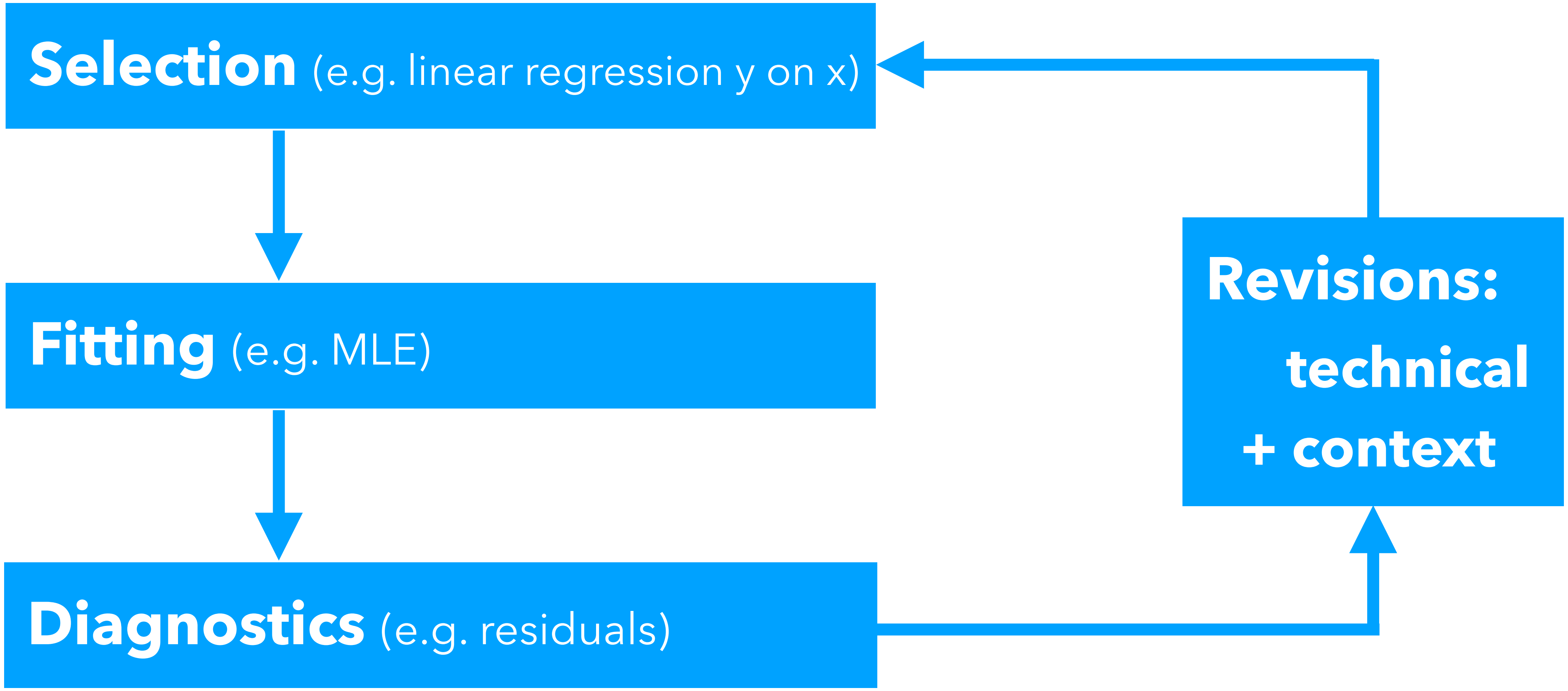
Hilbert's problems are **23 problems in mathematics** published by German mathematician David Hilbert in 1900. They were all unsolved at the time, and several proved to be very influential for 20th-century mathematics.

Hilbert presented ten of the problems (1, 2, 6, 7, 8, 13, 16, 19, 21, and 22) at the Paris conference of the International Congress of Mathematicians, speaking on August 8 at the Sorbonne.

The complete list of 23 problems was published later, in English translation in 1902 by Mary Frances Winston Newson in the Bulletin of the American Mathematical Society.[1]

https://en.wikipedia.org/wiki/Hilbert%27s_problems

Does my model fit and how can I improve it?

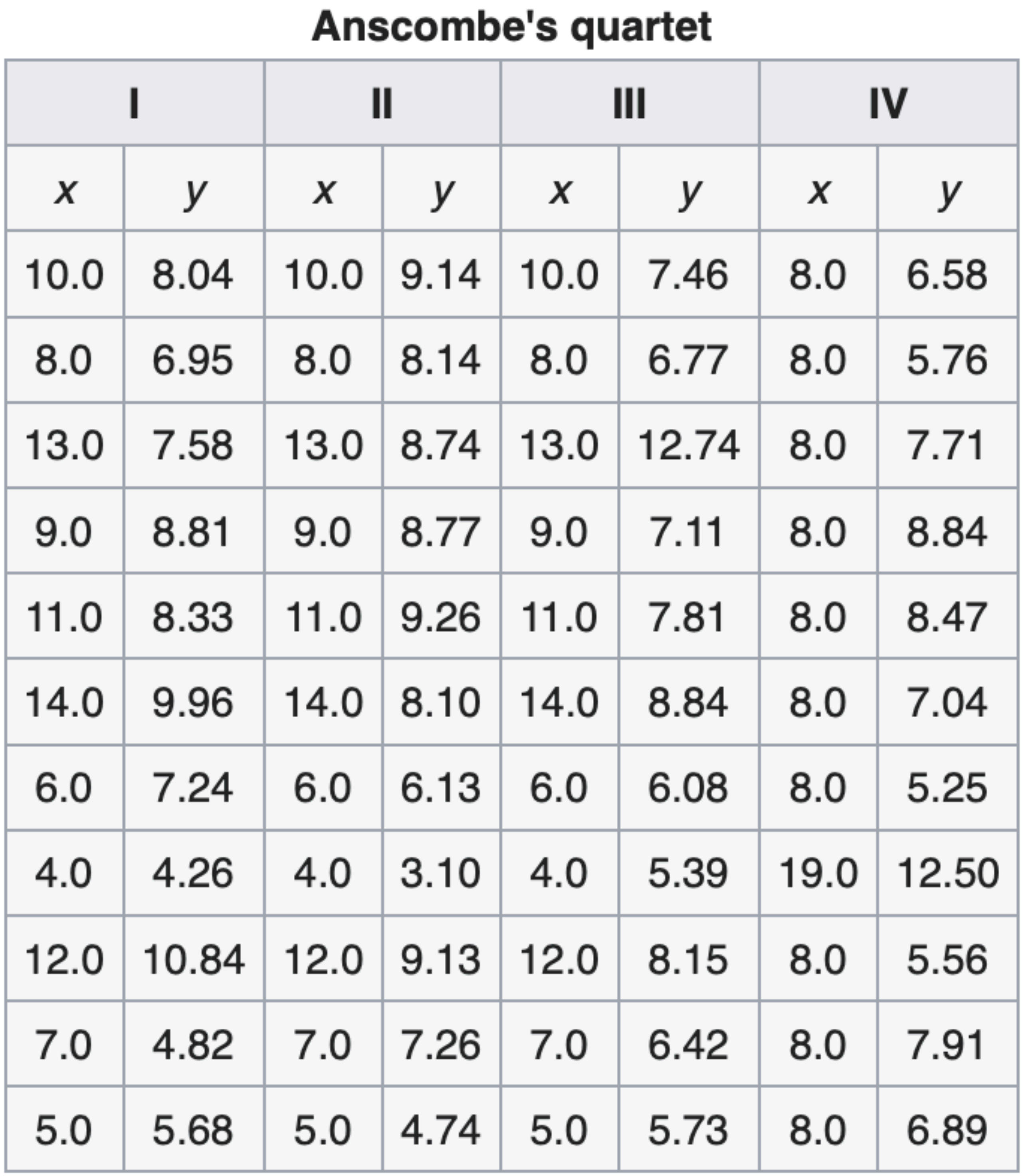

Selection (e.g. linear regression y on x)
Fitting (e.g. MLE)
Diagnostics (e.g. residuals)
Revisions:
technical
+ context

G Box

https://en.wikipedia.org/wiki/All_models_are_wrong

Anscombe’s quartet

Four datasets



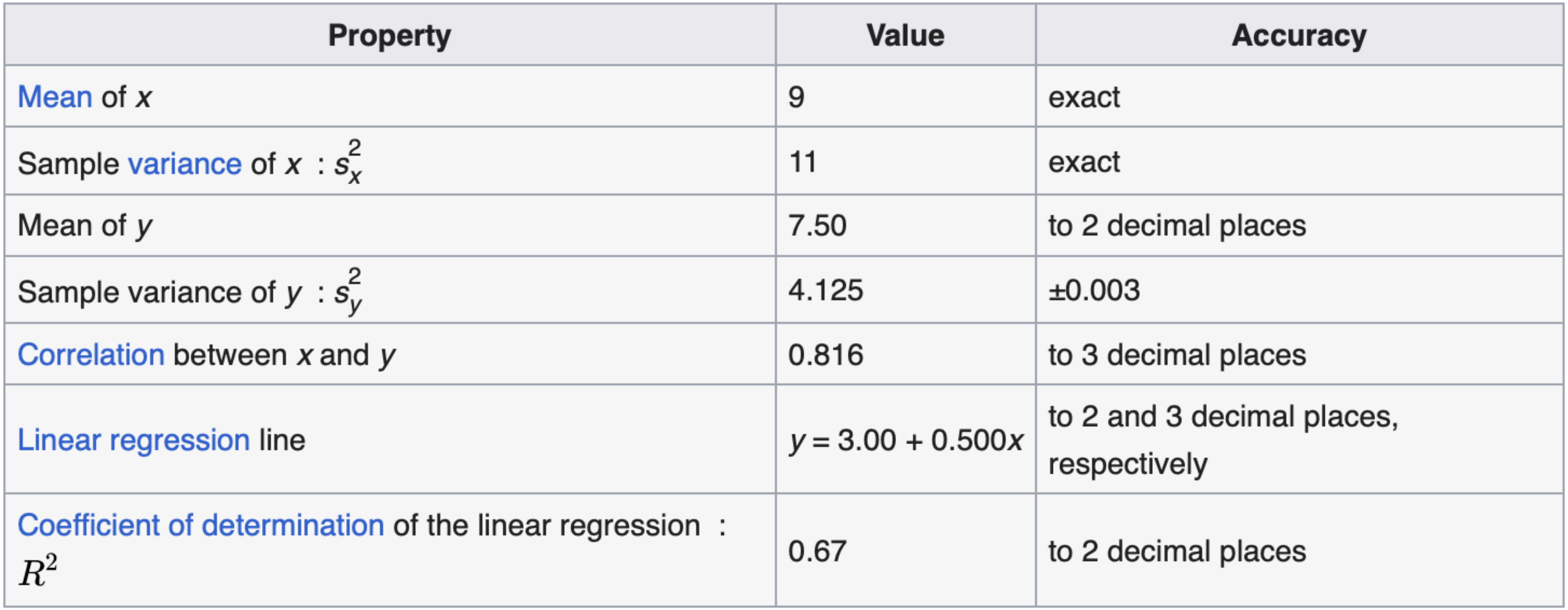
Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	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

https://en.wikipedia.org/wiki/Anscombe%27s_quartet and original references therein

Anscombe’s quartet

Four datasets with similar characteristics:



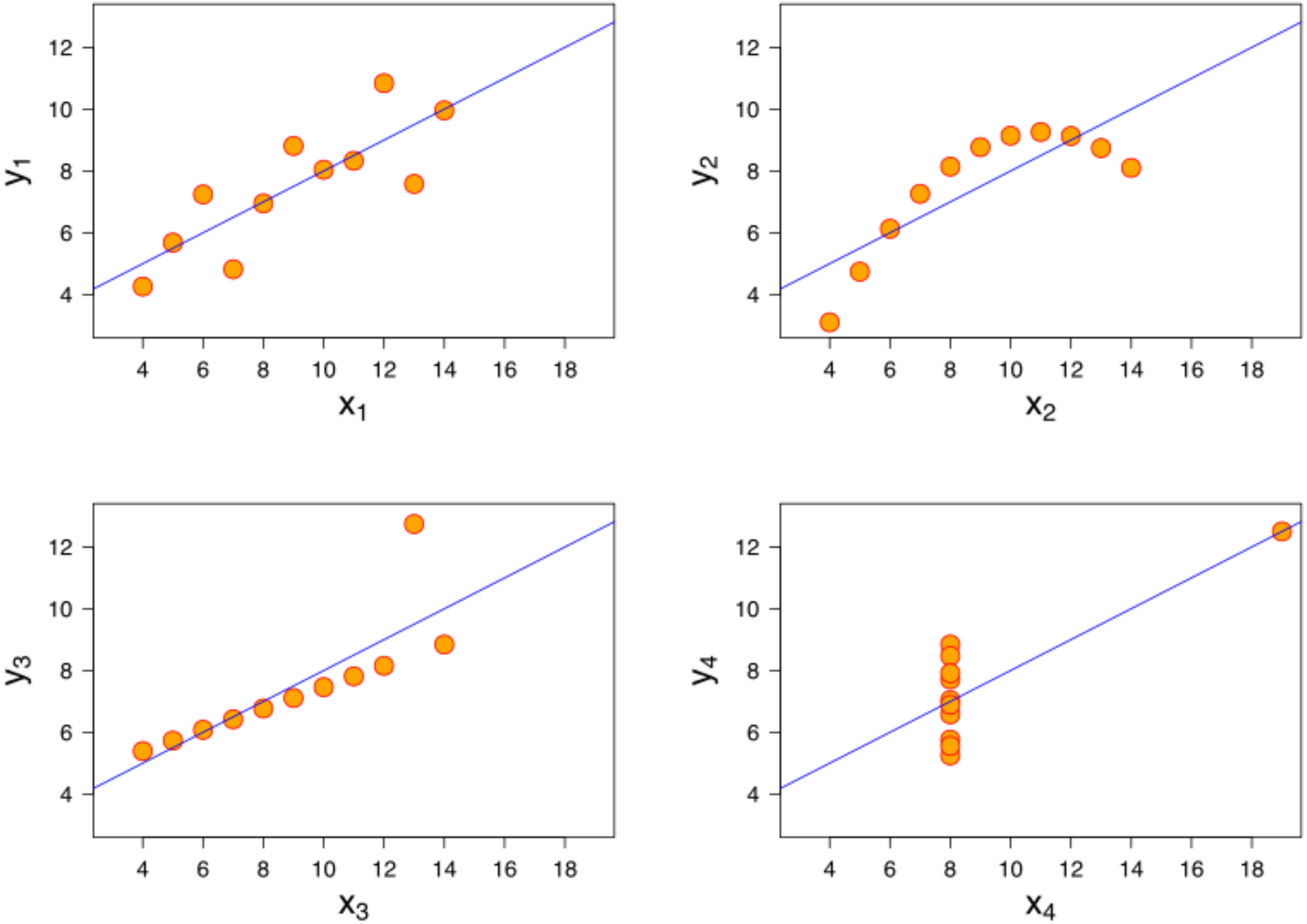
Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient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

https://en.wikipedia.org/wiki/Anscombe%27s_quartet and original references therein

Anscombe's quartet

Four datasets with similar characteristics:

But very different fit of a linear regression model:

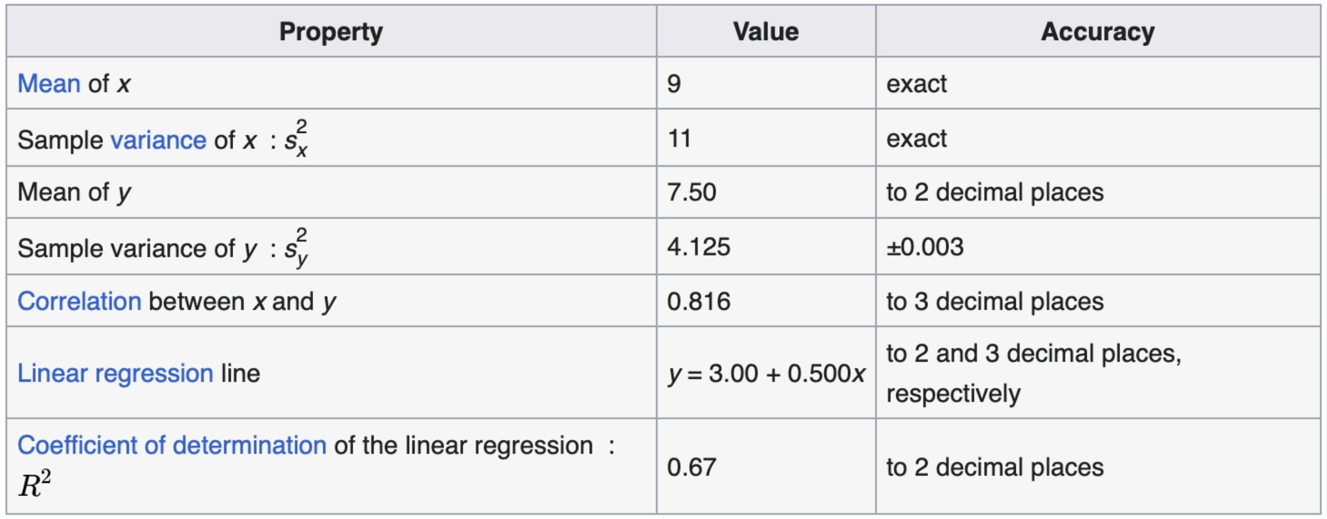


https://en.wikipedia.org/wiki/Anscombe%27s_quartet and original references therein

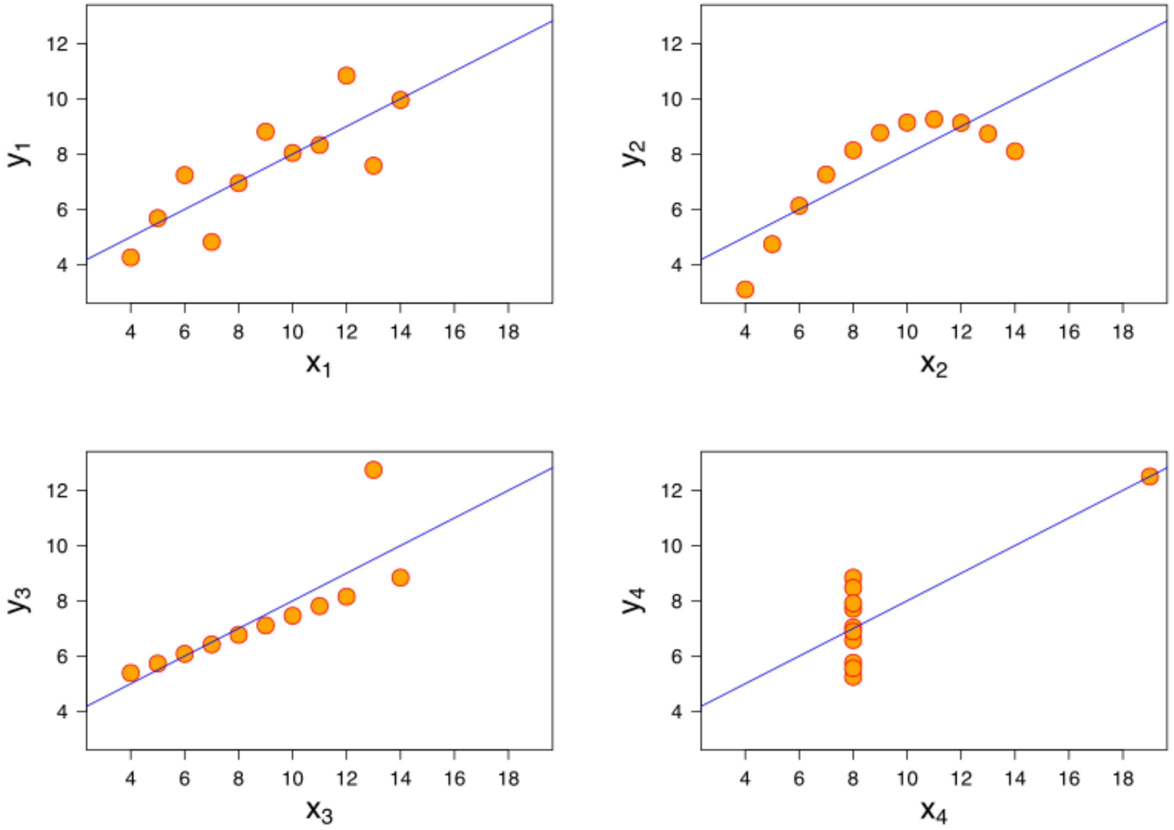
Anscombe’s quartet

Four datasets with similar characteristics:

Message: Always visualise your data!



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 and original references therein

Anscombe's less has been picked up by ML and DS community

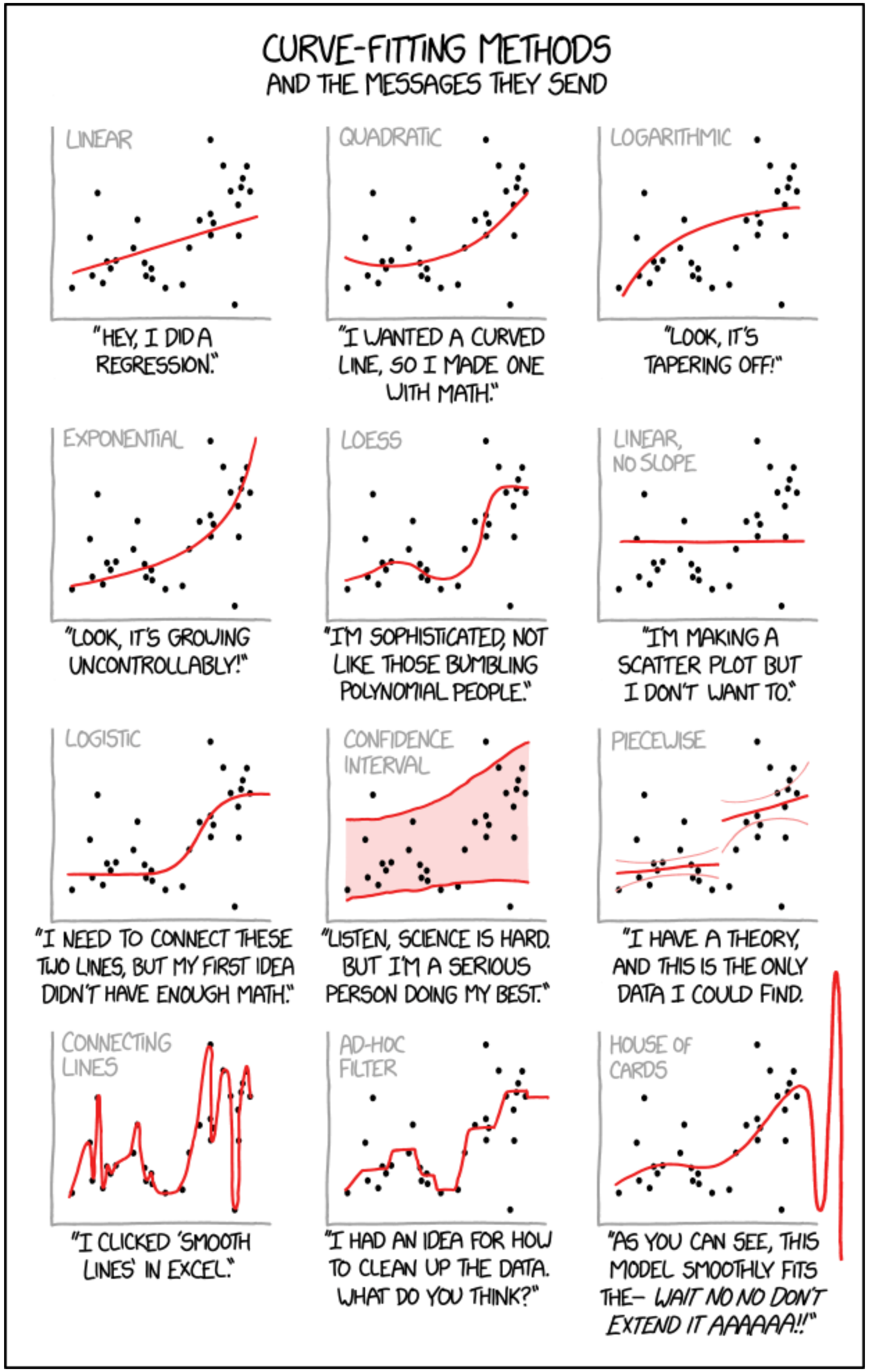
"…four datasets that were intentionally created to describe the importance of data visualisation and how any regression algorithm can be fooled by the same.

Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model."

https://towardsdatascience.com/importance-of-data-visualization-anscombes-quartet-way-a325148b9fd2
https://rawdatastudies.com/2021/06/19/anscombes-quartet-escapee/

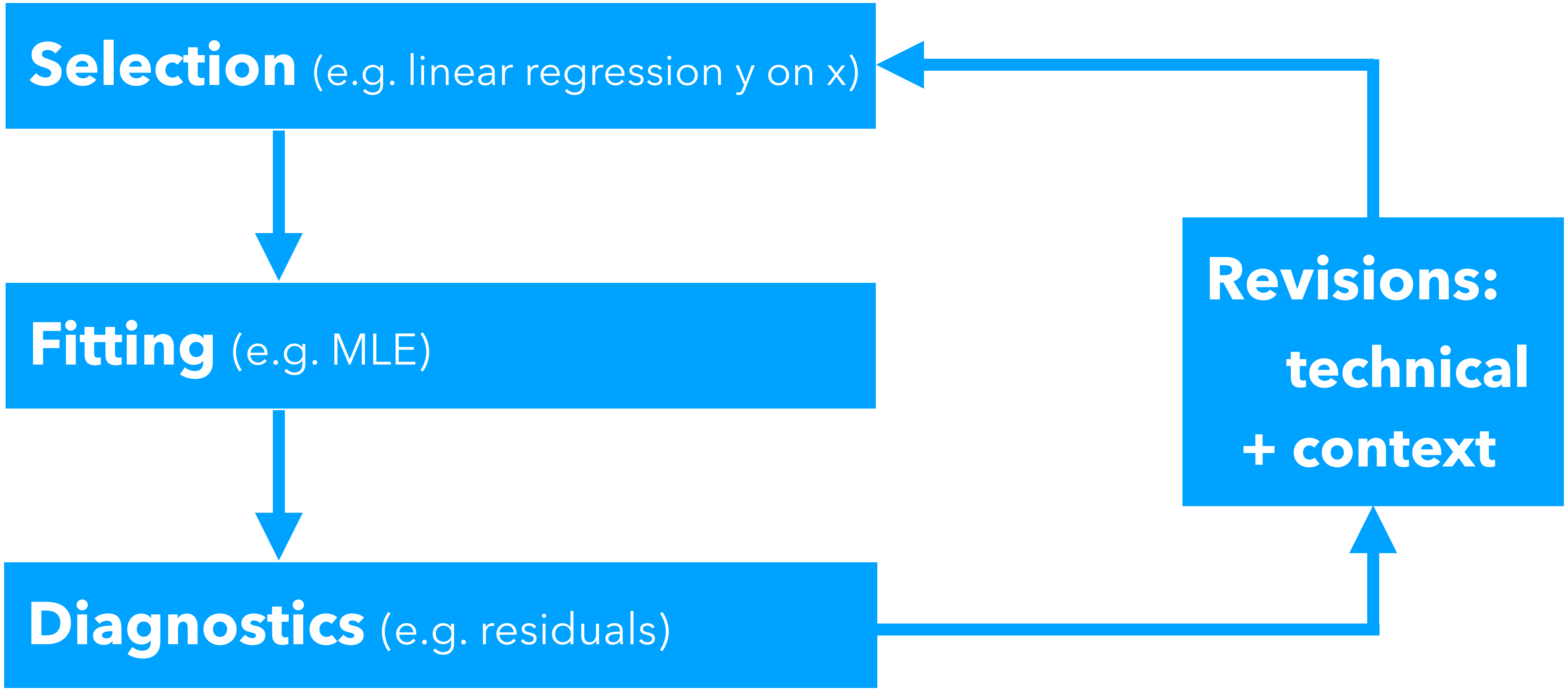
Cartoon version

Highly recommended: xkcd


CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND
LINEAR
"HEY, I DID A REGRESSION."
QUADRATIC
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."
LOGARITHMIC
"LOOK, IT'S TAPERING OFF!"
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY!"
LOESS
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."
LINEAR, NO SLOPE
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."
LOGISTIC
"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."
CONFIDENCE INTERVAL
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."
PIECEWISE
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND."
CONNECTING LINES
"I CLICKED 'SMOOTH LINES' IN EXCEL."
AD-HOC FILTER
"I HAD AN IDEA FOR HOW TO CLEAN UP THE DATA. WHAT DO YOU THINK?"
HOUSE OF CARDS
"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE- WAIT NO NO DON'T EXTEND IT AAAAAA!!"

https://www.explainxkcd.com/wiki/images/2/24/curve_fitting.png

Does my model fit and how can I improve it?



Selection (e.g. linear regression y on x)

Fitting (e.g. MLE)

Diagnostics (e.g. residuals)

Revisions:

technical

+ context

*"All **models are wrong**, some models are **useful**."*

G Box

https://en.wikipedia.org/wiki/All_models_are_wrong

George Box

George Edward Pelham Box FRS[1] (18 October 1919 – 28 March 2013) was a British statistician, who worked in the areas of quality control, time-series analysis, design of experiments, and Bayesian inference. He has been called "one of the great statistical minds of the 20th century".[3][4][5][6]

Education and early life

He was born in Gravesend, Kent, England. Upon entering university he began to study chemistry, but was called up for service before finishing. During World War II, he performed experiments for the British Army exposing small animals to poison gas. To analyze the results of his experiments, he taught himself statistics from available texts. After the war, he enrolled at University College London and obtained a bachelor's degree in mathematics and statistics. He received a PhD from the University of London in 1953, under the supervision of Egon Pearson.[2][7]

https://en.wikipedia.org/wiki/George_E._P._Box