Simple Linear Regression

Fonya Irvine, PhD candidate

BIOL 801: Pedagogical Training

About

- BSc (Marine Biology)
- MSc (Paleoclimatology)

→ UNB EST. 1785 UNIVERSITY OF NEW BRUNSWICK

- Parks Canada

→ Canada

- Fisheries and Oceans

→ Fisheries and Oceans Canada

- Environmental Consulting

→ AJM ENVIRONMENTAL INC

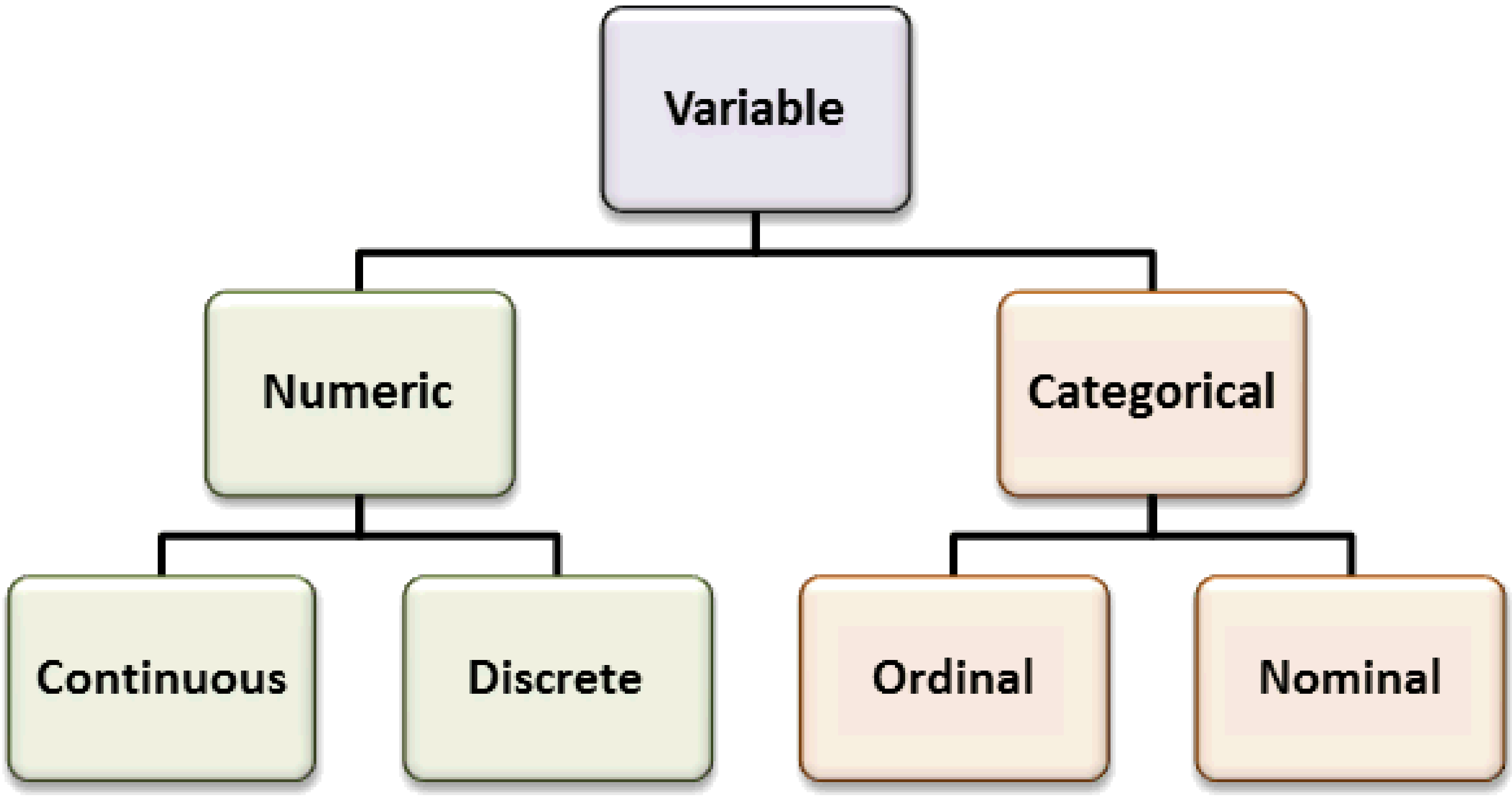
- PhD candidate (Biostatistics)

→ UNIVERSITÉ Concordia UNIVERSITY

Outline:

1) Introduction to Regression Analysis
2) Types of Regression Models
3) Simple Linear Regression Components
4) Fitting a Regression Model
5) Model Significance
6) Interpreting Results
7) Key Assumptions
8) Limitations

Data Types


Variable
Numeric
Categorical
Continuous
Discrete
Ordinal
Nominal

Data Exploration

Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2010, **1**, 3–14 doi: 10.1111/j.2041-210X.2009.00001.x

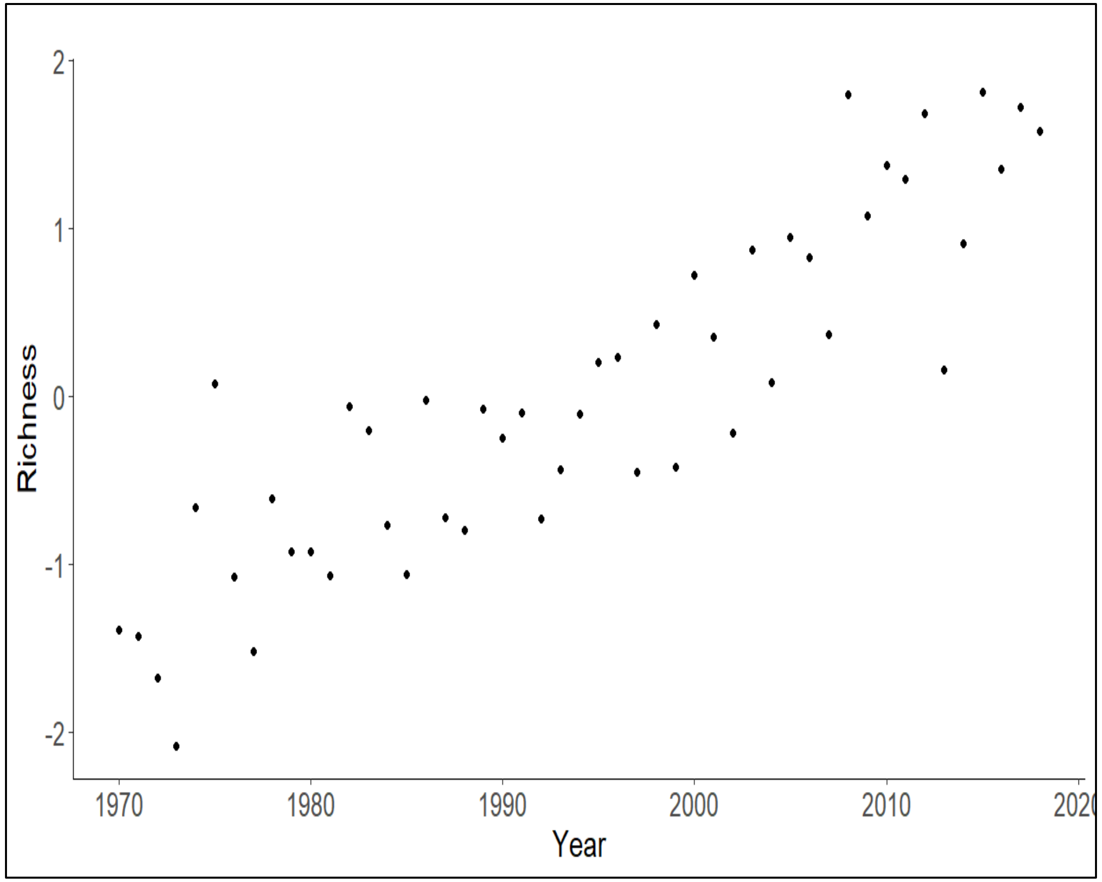
A protocol for data exploration to avoid common statistical problems

Alain F. Zuur*[1,2], Elena N. Ieno[1,2] and Chris S. Elphick[3]

[1]Highland Statistics Ltd, Newburgh, UK; [2]Oceanlab, University of Aberdeen, Newburgh, UK; and [3]Department of Ecology and Evolutionary Biology and Center for Conservation Biology, University of Connecticut, Storrs, CT, USA

1. Introduction to Regression Analysis

- **What is Regression?**
  - Method for exploring the relationship between two continuous variables.

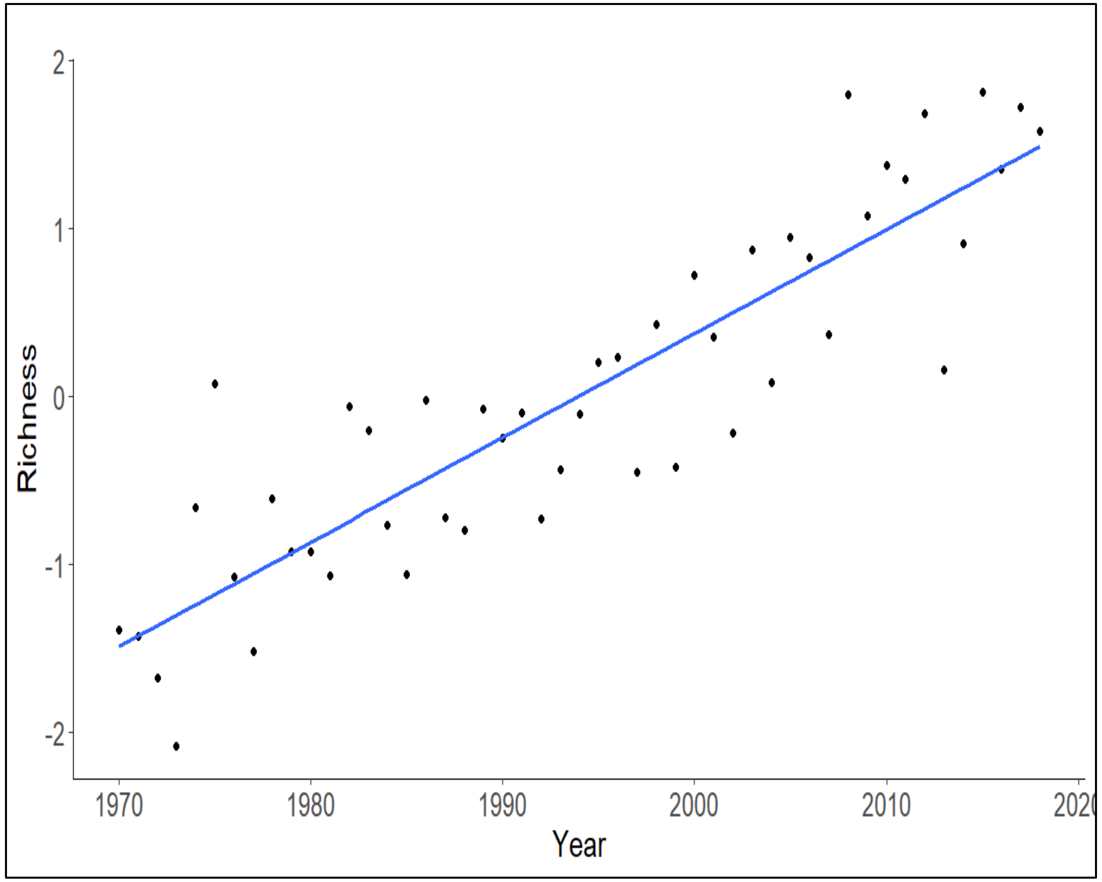


1. Introduction to Regression Analysis

- **What is Regression?**
  - Method for exploring the relationship between two continuous variables.

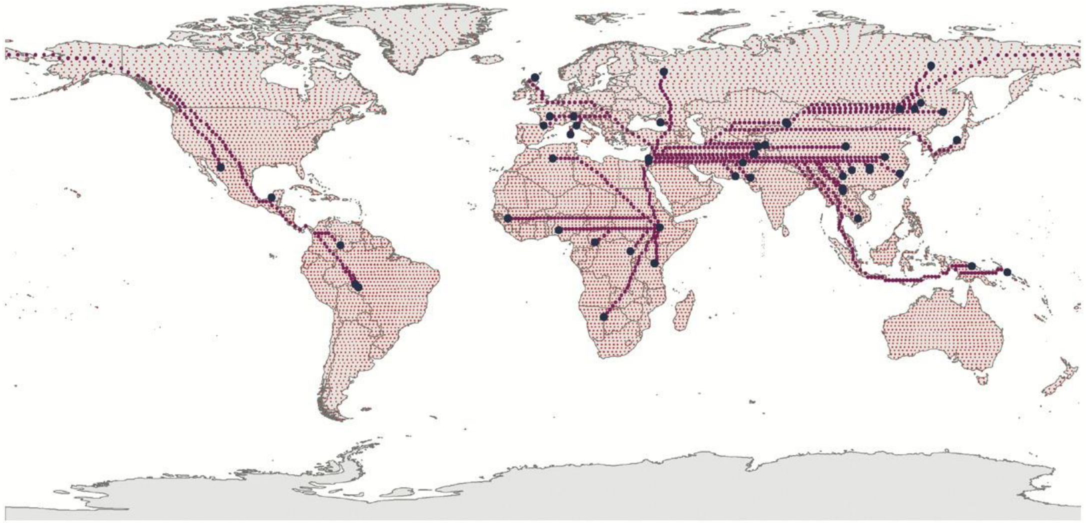
- The predictor variable, X **predicts** the **response** of the response variable, Y.

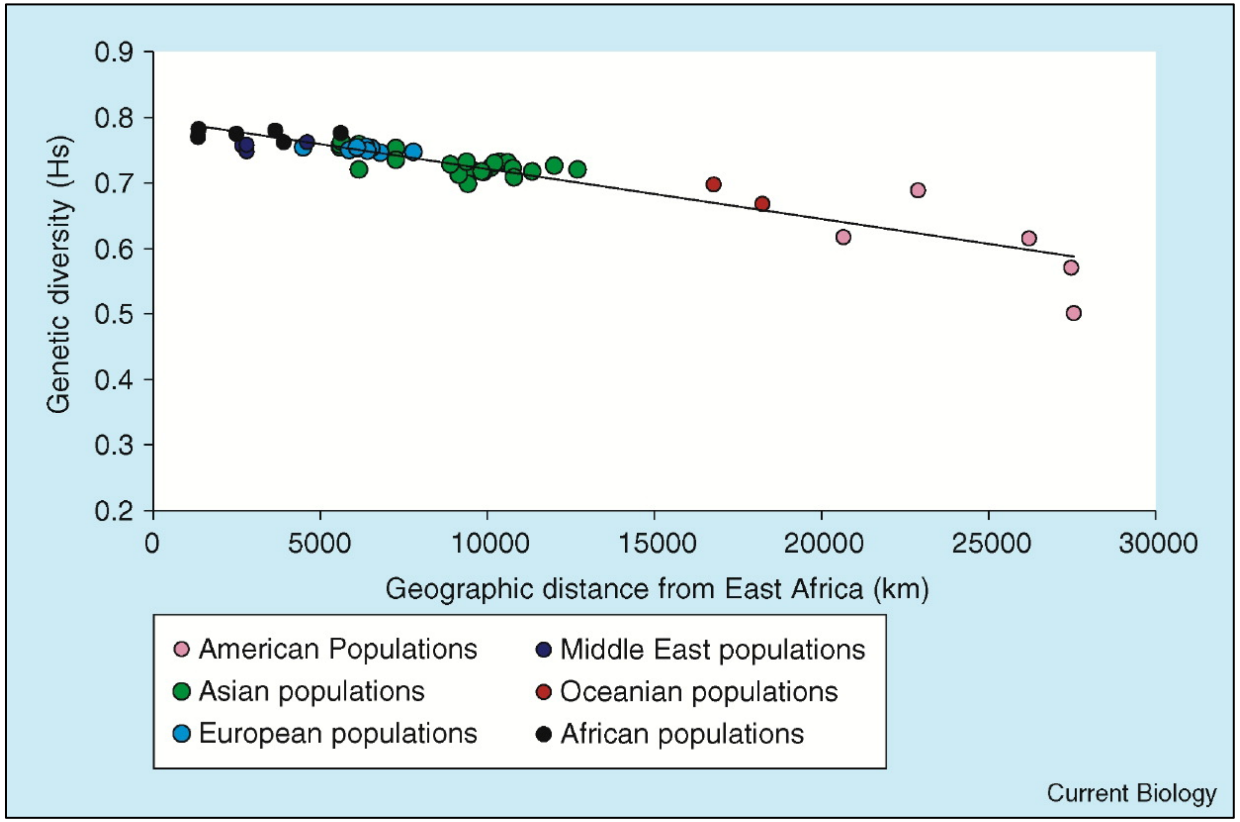
- The regression line is the **“best fit”**



1. Introduction to Regression Analysis

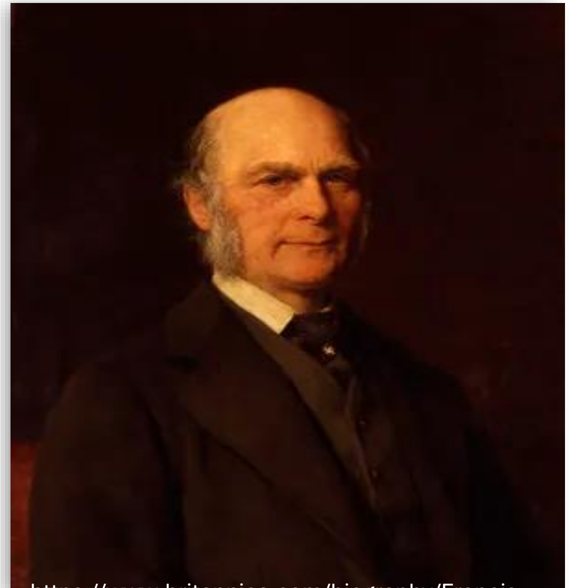
- Example: Genetic diversity vs. geographic distance from Africa (Prugnolle et al., 2005)



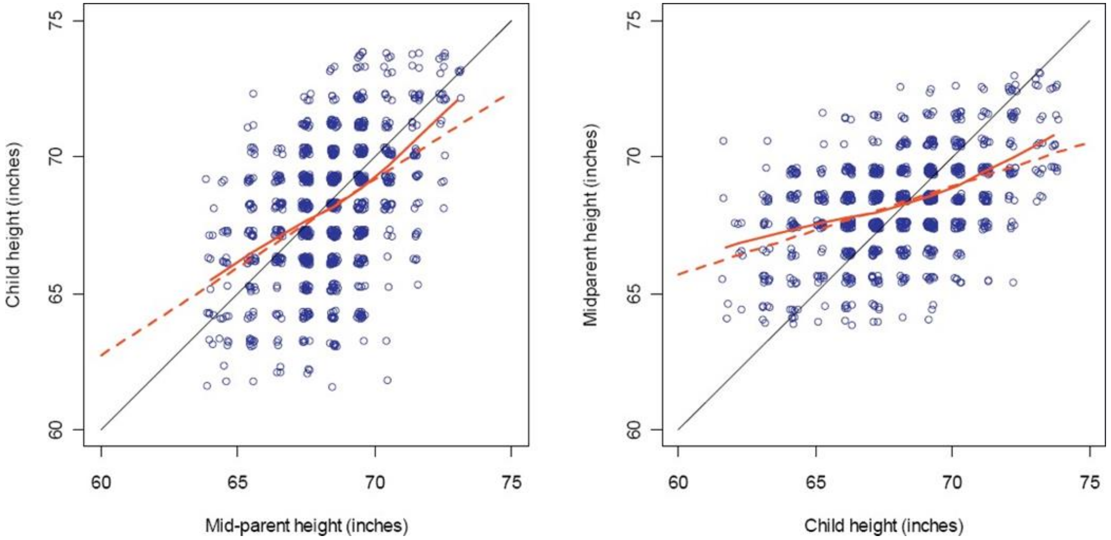


1. Introduction to Regression Analysis

- Regression *noun*
- Why the term “Regression”?
    - Historical context from Francis Galton’s work on height between fathers and sons (regression toward mediocrity)



https://www.britannica.com/biography/Francis



Significance, Volume: 8, Issue: 3, Pages: 124-126, First published: 25 August 2011, DOI: (10.1111/j.1740-9713.2011.00509.x)

2. Types of Regression Models

Simple Linear Regression

- Model the relationship between a predictor variable, X, and a response variable, Y.

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

Multiple Linear Regression

- The average value of the response variable, Y, is assumed to be a linear combination of the predictor variables, X_1, $X_{2, \ldots,}$ X_n

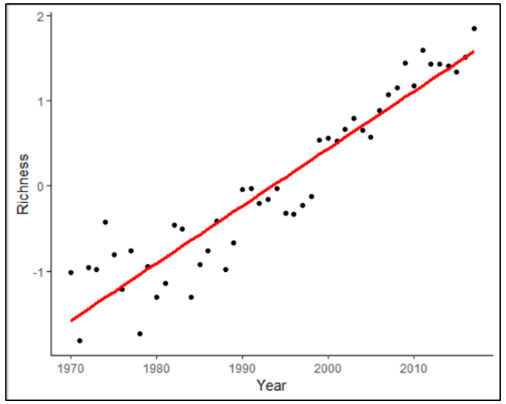
$$Y_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_n X_{n,i} + \varepsilon_i$$

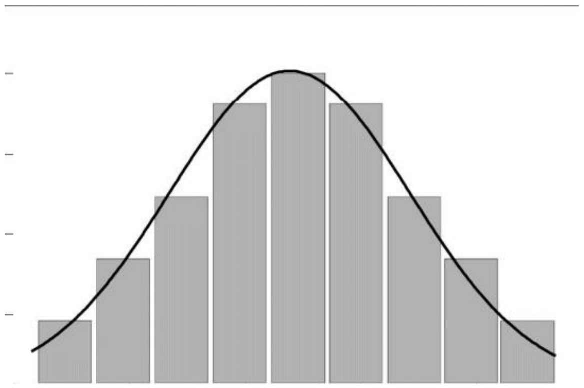
$$\varepsilon_i \sim N(0, \sigma^2)$$

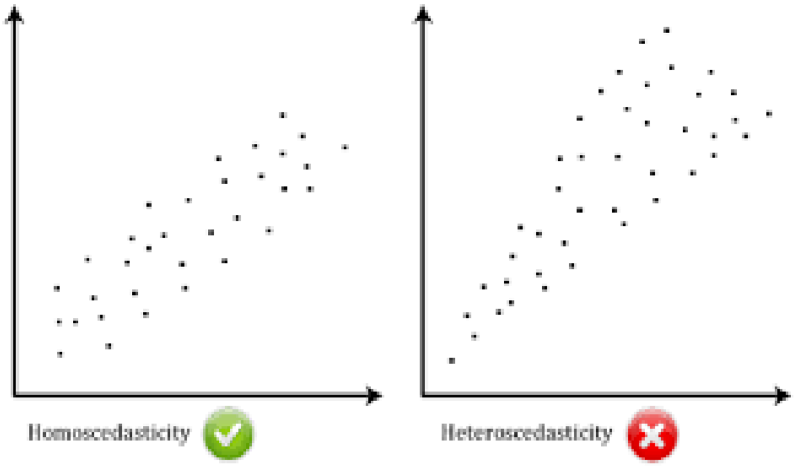
2. Types of Regression Models

Quadratic Regression: $$Y_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i}^2 + \varepsilon_i$$

Polynomial Regression: $$Y_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i}^2 + \beta_2 X_{3,i}^3 + \beta_n X_{n,i}^n + \varepsilon_i$$







3. Simple Linear Regression Components

Response Variable

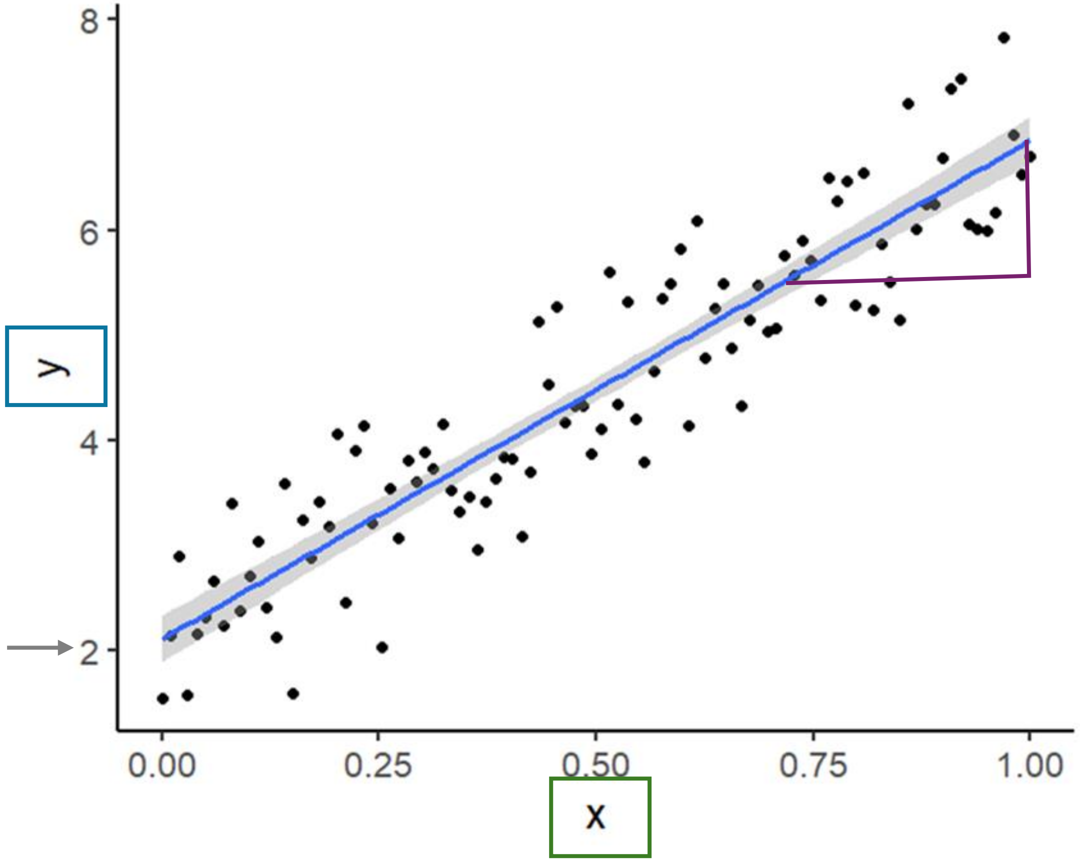
Predictor Variable

Intercept

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Slope

Residual Error

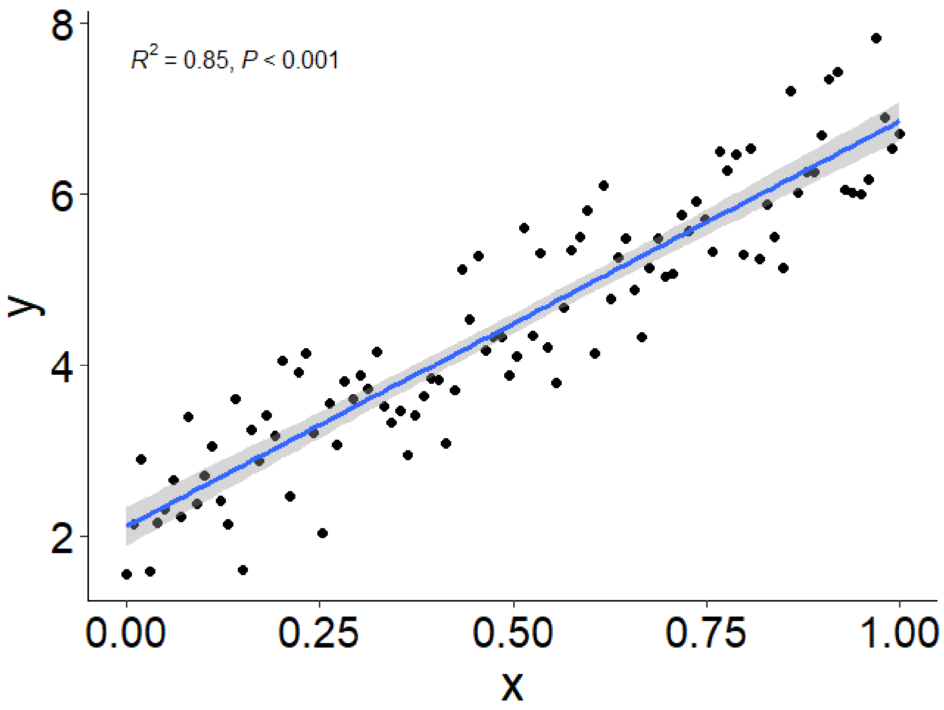


Linear Model

$$Y_i = \alpha + \beta X_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$$

Equations for each observation

$$y_1 = \alpha + \beta x_1 + \varepsilon_1 \qquad \varepsilon_1 \sim N(0, \sigma^2)$$
$$y_2 = \alpha + \beta x_2 + \varepsilon_2 \qquad \varepsilon_2 \sim N(0, \sigma^2)$$
$$\vdots$$
$$y_n = \alpha + \beta x_n + \varepsilon_n \qquad \varepsilon_3 \sim N(0, \sigma^2)$$



Linear Model

$$\boldsymbol{Y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}$$

Response Vector

Vector of Parameters

Design Matrix

Error Vector

$$\boldsymbol{Y} = \begin{pmatrix} y1 \\ y2 \\ y3 \\ \ldots \\ y\mathrm{n} \end{pmatrix} \quad \underbrace{\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x1 \\ 1 & x2 \\ 1 & x3 \\ 1 & \ldots \\ 1 & x\mathrm{n} \end{pmatrix}}_{\text{Multiply matrix } \mathbf{X} \text{ by vector } \boldsymbol{\beta}} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon 1 \\ \varepsilon 2 \\ \varepsilon 3 \\ \ldots \\ \varepsilon \mathrm{n} \end{pmatrix}$$

Linear Model

$$\mathbf{Y} = \boxed{\boldsymbol{\beta}\mathbf{X}} + \boldsymbol{\varepsilon}$$

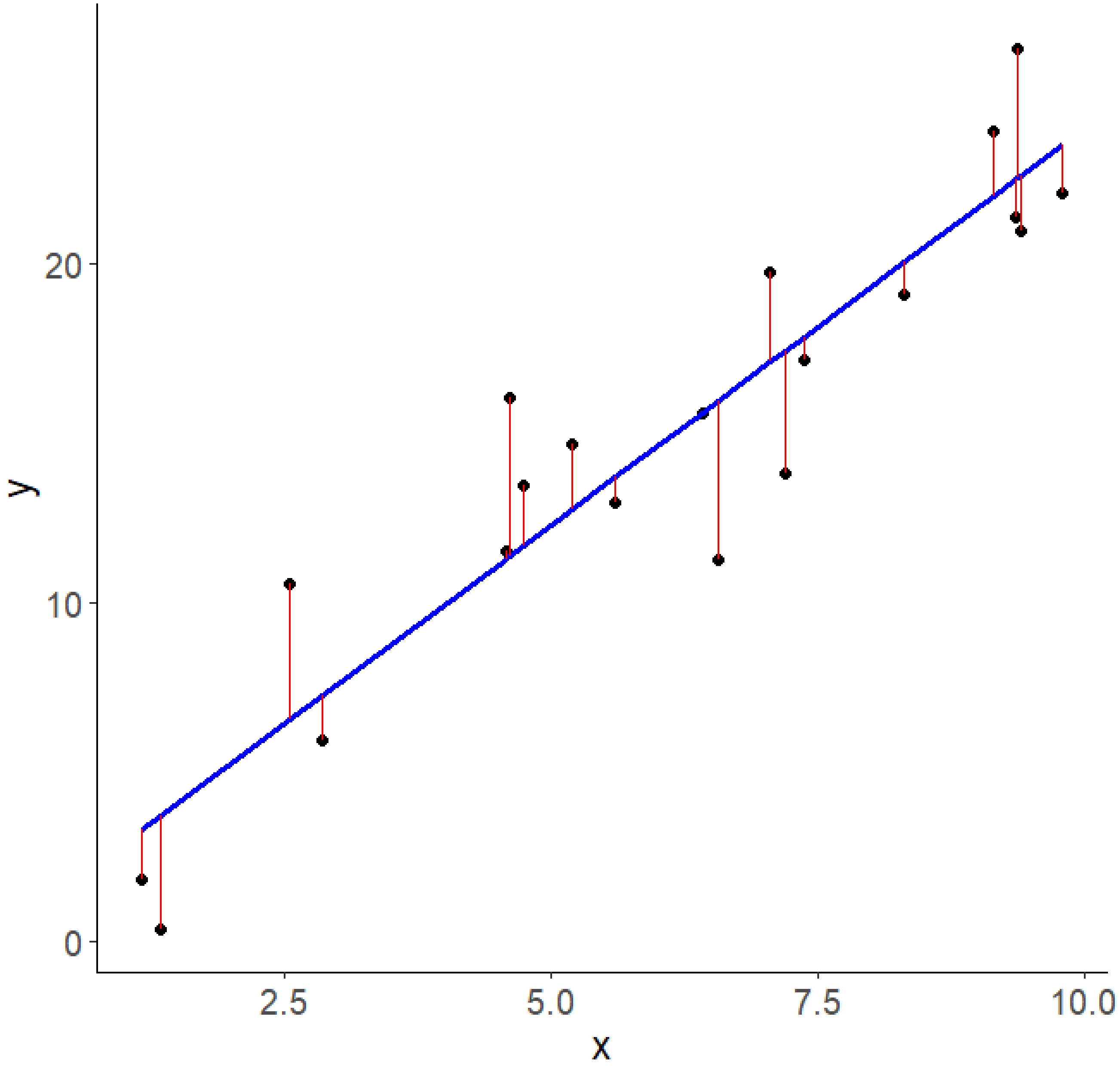
$$\mathbf{Y} = \begin{pmatrix} y1 \\ y2 \\ y3 \\ \dots \\ y\mathrm{n} \end{pmatrix} \quad \boldsymbol{\beta X} = \begin{pmatrix} \alpha & + & \beta \cdot x1 \\ \alpha & + & \beta \cdot x2 \\ \alpha & + & \beta \cdot x3 \\ \alpha & + & \dots \\ \alpha & + & \beta \cdot x\mathrm{n} \end{pmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon 1 \\ \varepsilon 2 \\ \varepsilon 3 \\ \dots \\ \varepsilon \mathrm{n} \end{pmatrix}$$

Residuals

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

- Residuals are the difference between **observed** and **predicted** values

- Describes what is not explained by the model

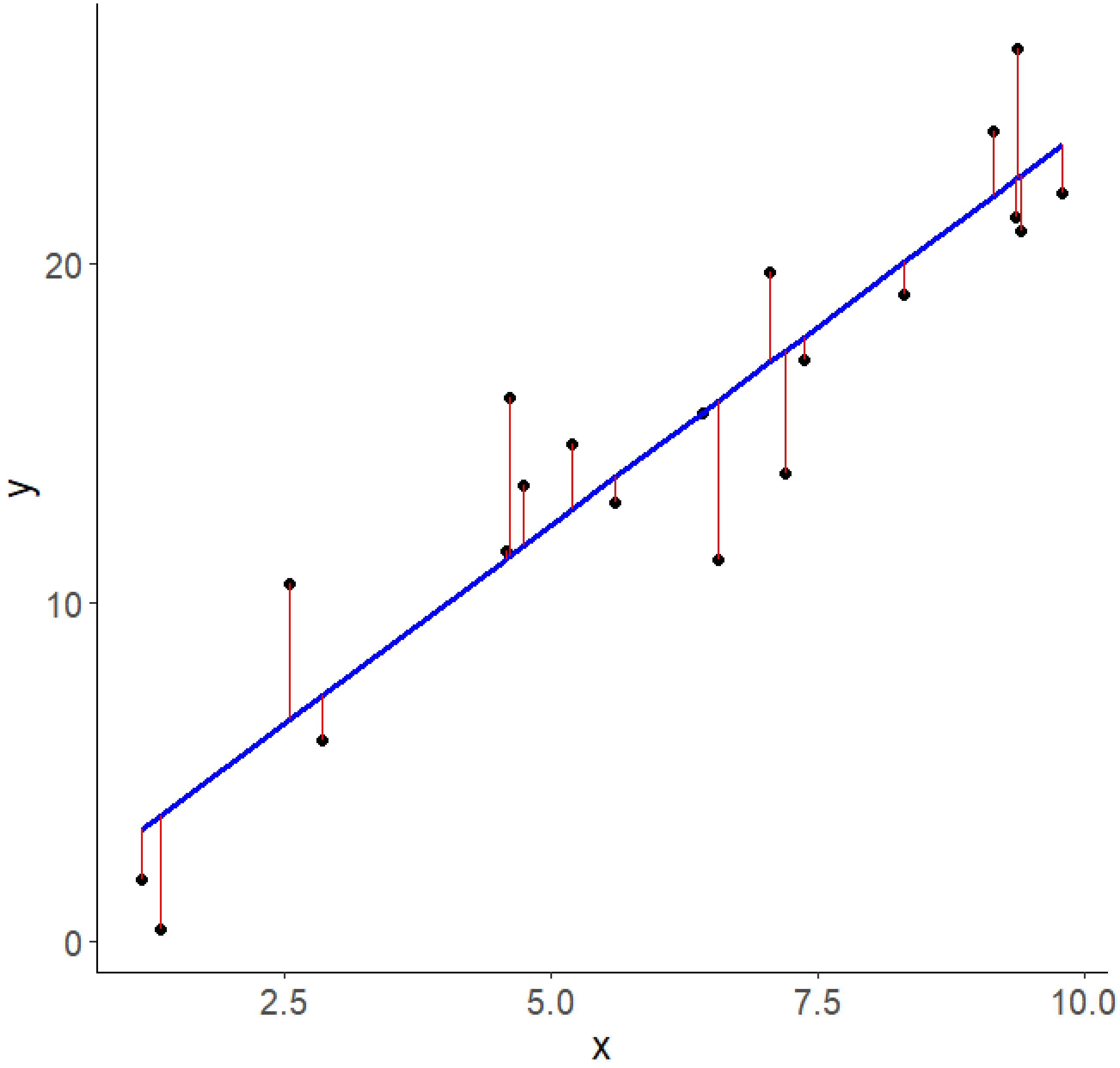


Residuals

$$\varepsilon_i \sim N(0, \sigma^2)$$

- Variance, σ^2, describes the variation of observations around the regression line

- Standard Deviation, σ, describes the average deviation from the regression line



4. Fitting a Regression Model

Residuals: $\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{X} \times \boldsymbol{\beta}$

$$\sum_i \varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = [\varepsilon_1 \quad \varepsilon_2 \quad \varepsilon_3 \cdots \varepsilon_n] \times \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \cdots \\ \varepsilon_n \end{bmatrix}$$

Sum Squared Residuals in Matrix Form

$$\sum_i \left(Y_i - (\alpha + \beta \cdot x_i)\right)^2 = (\mathbf{Y} - \mathbf{X} \cdot \beta)^t \times (\mathbf{Y} - \mathbf{X} \cdot \beta)$$

Optimize with Ordinary Least Squares (OLS)

$$\frac{d}{d\beta}\left((\boldsymbol{Y} - \mathbf{X} \cdot \beta)^t(\boldsymbol{Y} - \mathbf{X} \cdot \boldsymbol{\beta})\right) = -2\mathbf{X}^t(\boldsymbol{Y} - \mathbf{X} \cdot \beta)$$

Take derivative with respect to β

$$-2\mathbf{X}^t(\boldsymbol{Y} - \mathbf{X} \cdot \beta) = \mathbf{0}$$

Set to zero and solve for β

$$\mathbf{X}^{\mathrm{t}}\mathbf{Y} = (\mathbf{X}^t\mathbf{X})\beta$$

Equation to estimate parameters

$$\hat{\beta} = (\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{t}}\mathbf{Y}$$

Equation to solve for estimated parameters

4. Fitting a Regression Model

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^t \times \mathbf{X})^{-1} \times \mathbf{X}^t \times \mathbf{Y}$$

$$\widehat{\boldsymbol{y}} = \mathbf{X} \times \widehat{\boldsymbol{\beta}}$$

$$\widehat{\boldsymbol{y}} = \boxed{\boldsymbol{X} \times (\mathbf{X}^t \times \mathbf{X})^{-1} \times \mathbf{X}^t} \times \mathbf{Y}$$

$$\widehat{\boldsymbol{y}} = \mathbf{H} \times \mathbf{Y}$$

H is the hat matrix

Linear Model: Identity Matrix (n x n)

$$I = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}$$

Assumptions:

- Diagonal elements equal 1 and specify that the variance of each residual is 1 times σ^2
- Off-diagonal elements equal 0 and specify that the covariance between different residuals is 0
- Correlations are zero

Linear Model: Variance-Covariance Matrix

$$\sigma^2\{\boldsymbol{X}\} = \begin{bmatrix} \sigma^2\{x_1\} & \cdots & \sigma^2\{x_1, x_n\} \\ \vdots & \ddots & \vdots \\ \sigma^2\{x_n, x_1\} & \cdots & \sigma^2\{x_n\} \end{bmatrix}$$

$$\sigma^2\{\boldsymbol{\varepsilon}\} = Cov\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \sigma^2 I = \begin{bmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{bmatrix}$$

Linear Model: Residuals

$$\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$$

$$\widehat{\boldsymbol{y}} = \mathbf{H} \times \boldsymbol{Y}$$

$$\boldsymbol{e} = \boldsymbol{y} - \mathbf{H} \times \boldsymbol{Y}$$

$$\boldsymbol{e} = (\mathbf{I} - \mathbf{H}) \times \boldsymbol{Y}$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \times \mathbf{I})$$

Maximum Likelihood Estimation (MLE)

- MLE finds the “best fit” through the data using the log-likelihood function:

$$\ln L\boxed{(\alpha, \beta_1, \sigma^2)} = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\boxed{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

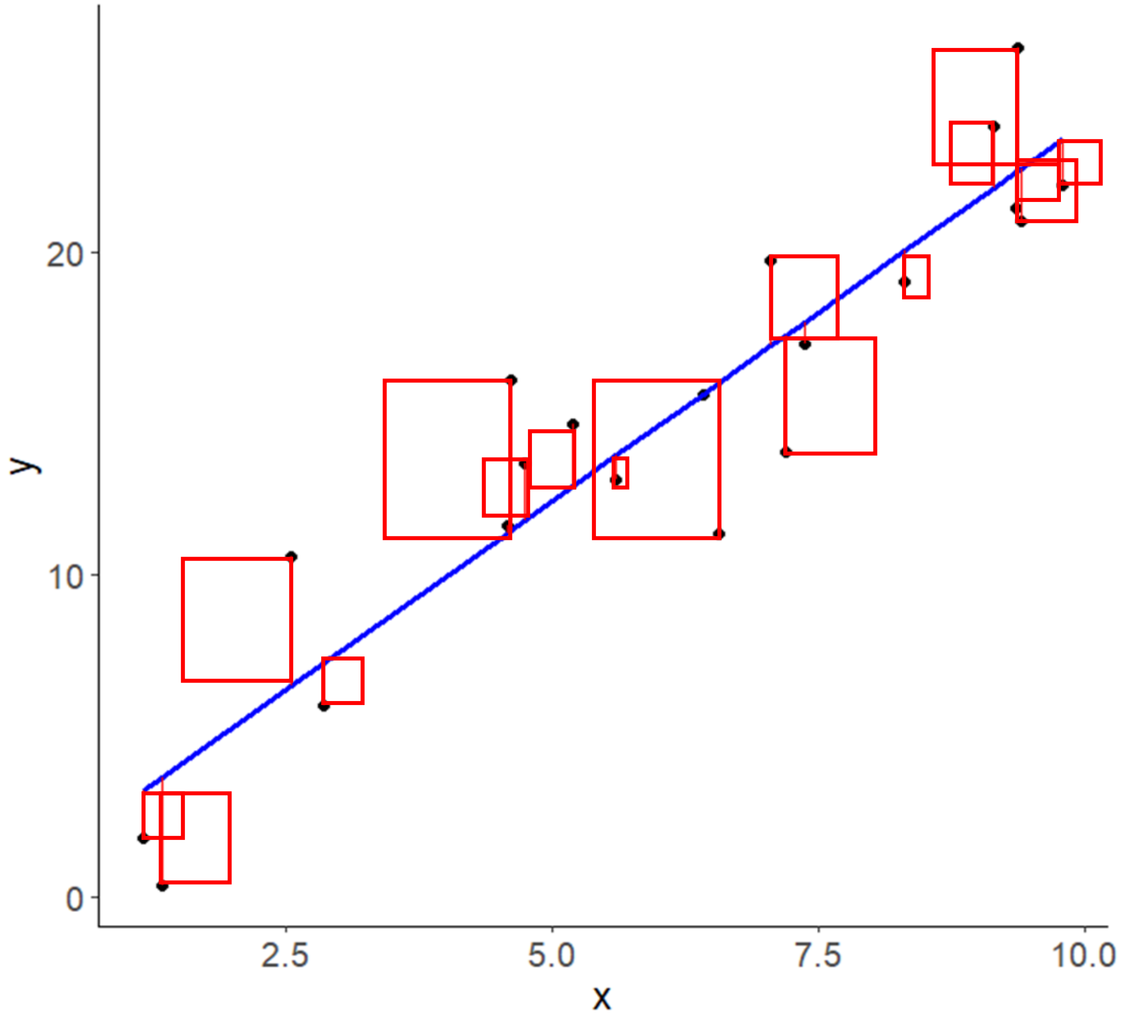
- How?
  - Maximizing the log-likelihood function by minimizing the Sum of Squared Errors:

$$SSE = \boxed{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

How’s the fit?

- Sum of squared errors (SSE) is a measure of unexplained variability.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$



How’s the fit?

- Sum of squares for regression (SSR) is a measure of explained variability.

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- Total sum of squares (SST) is a measure of total variability.

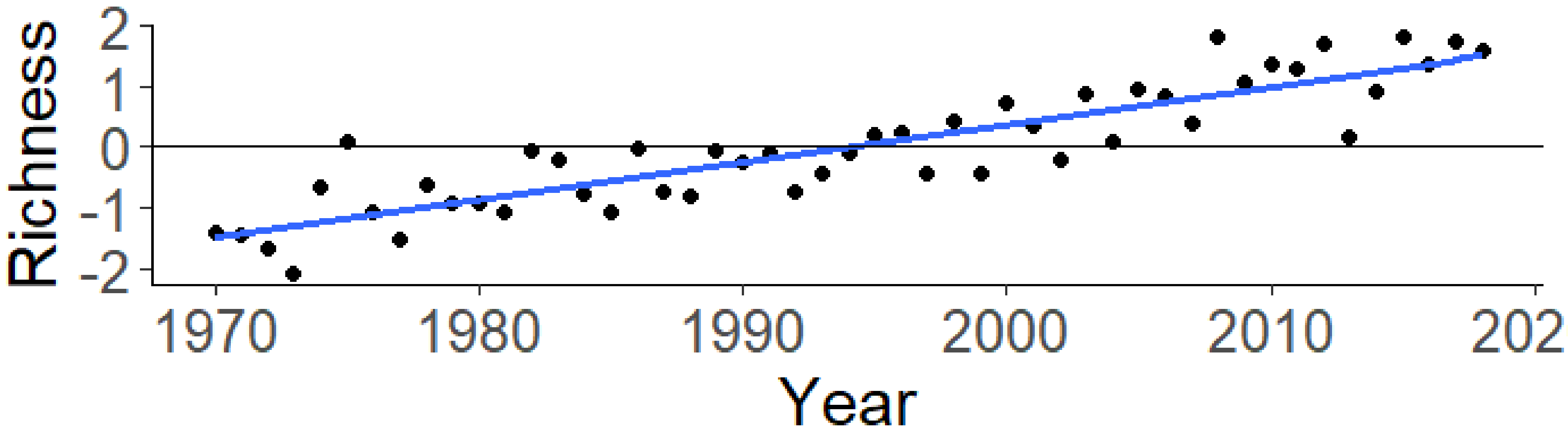
$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$\boxed{\boldsymbol{SST = SSR + SSE}}$$

5. Model Significance

- Model Outputs:
  - Test whether the slope of the relationship is zero or not

$$H_0\colon\ \beta_1 = 0 \qquad\qquad H_1\colon\ \beta_1 \neq 0$$



5. Model Significance

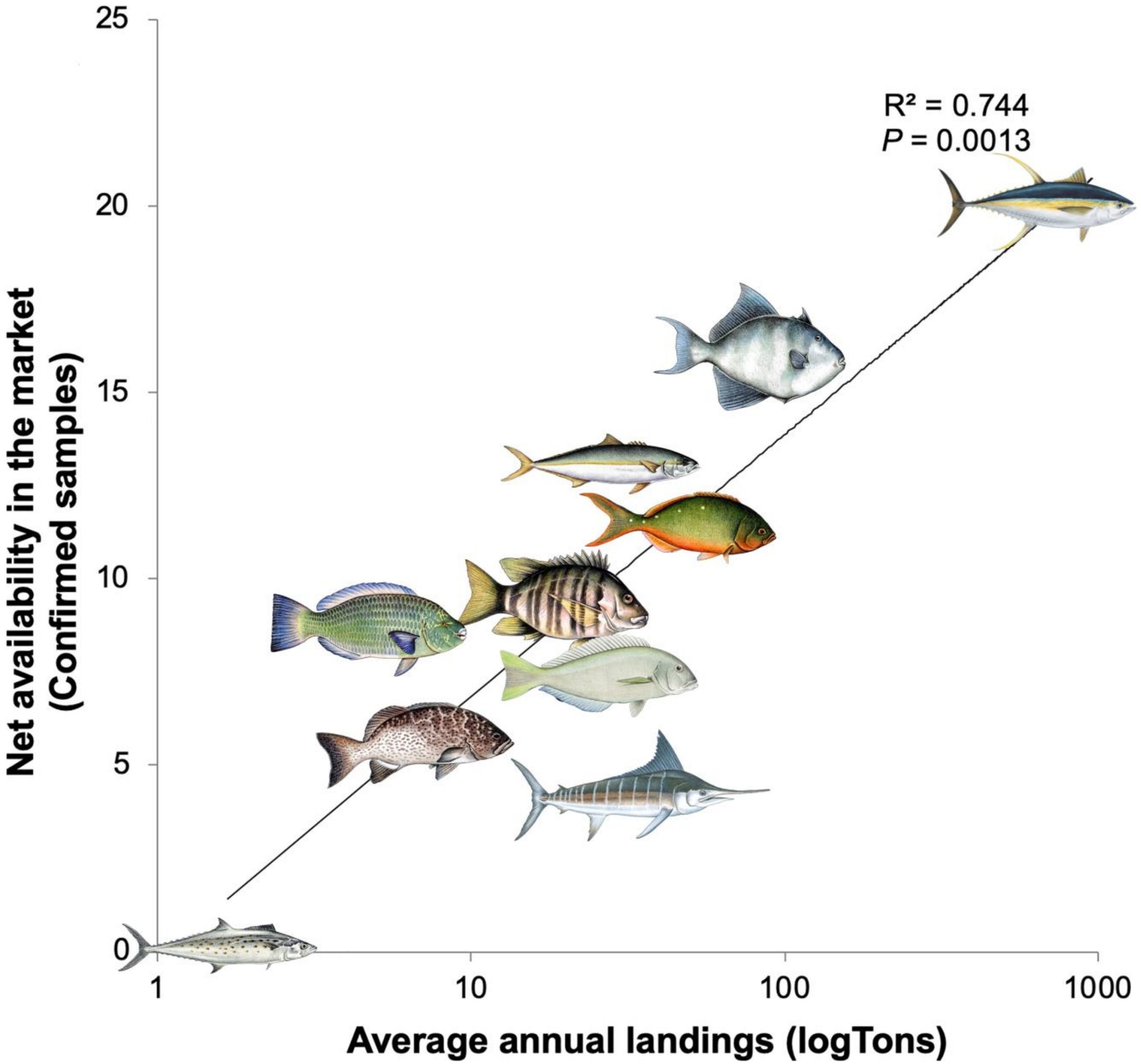
- Positive relationship between Richness and Year

- Strong evidence against null hypothesis that slope = 0

```
Call:
lm(formula = Richness ~ Year, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.02461 -0.33602  0.03834  0.28930  1.25274 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.237e+02  9.429e+00  -13.12   <2e-16 ***
Year         6.202e-02  4.729e-03   13.12   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4681 on 47 degrees of freedom
Multiple R-squared:  0.7854,	Adjusted R-squared:  0.7809 
F-statistic: 172.1 on 1 and 47 DF,  p-value: < 2.2e-16
```

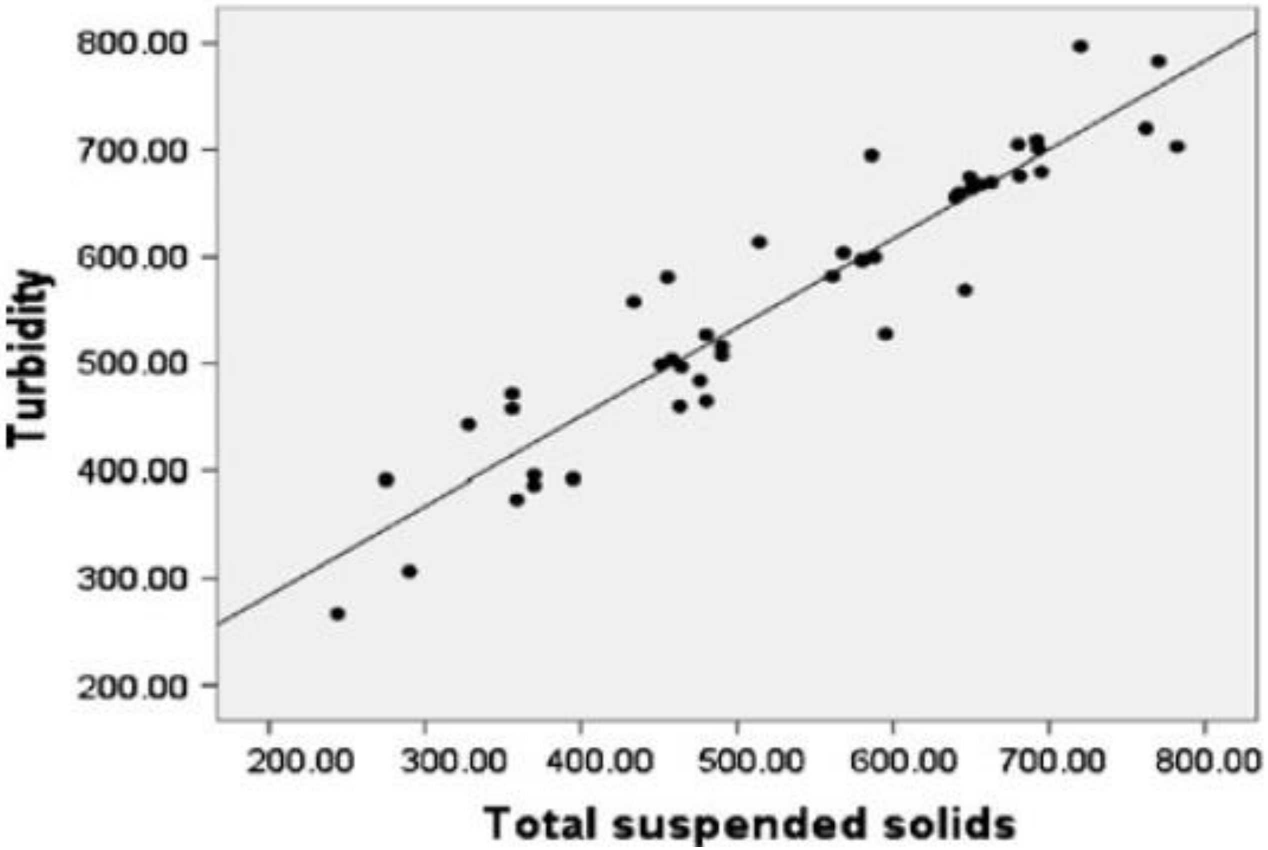
6. Interpreting Results

- Example: Net availability in market vs. average annual landings (Munguia-Vega et al. 2020)
- R^2=0.744
- *p*-value = 0.0013



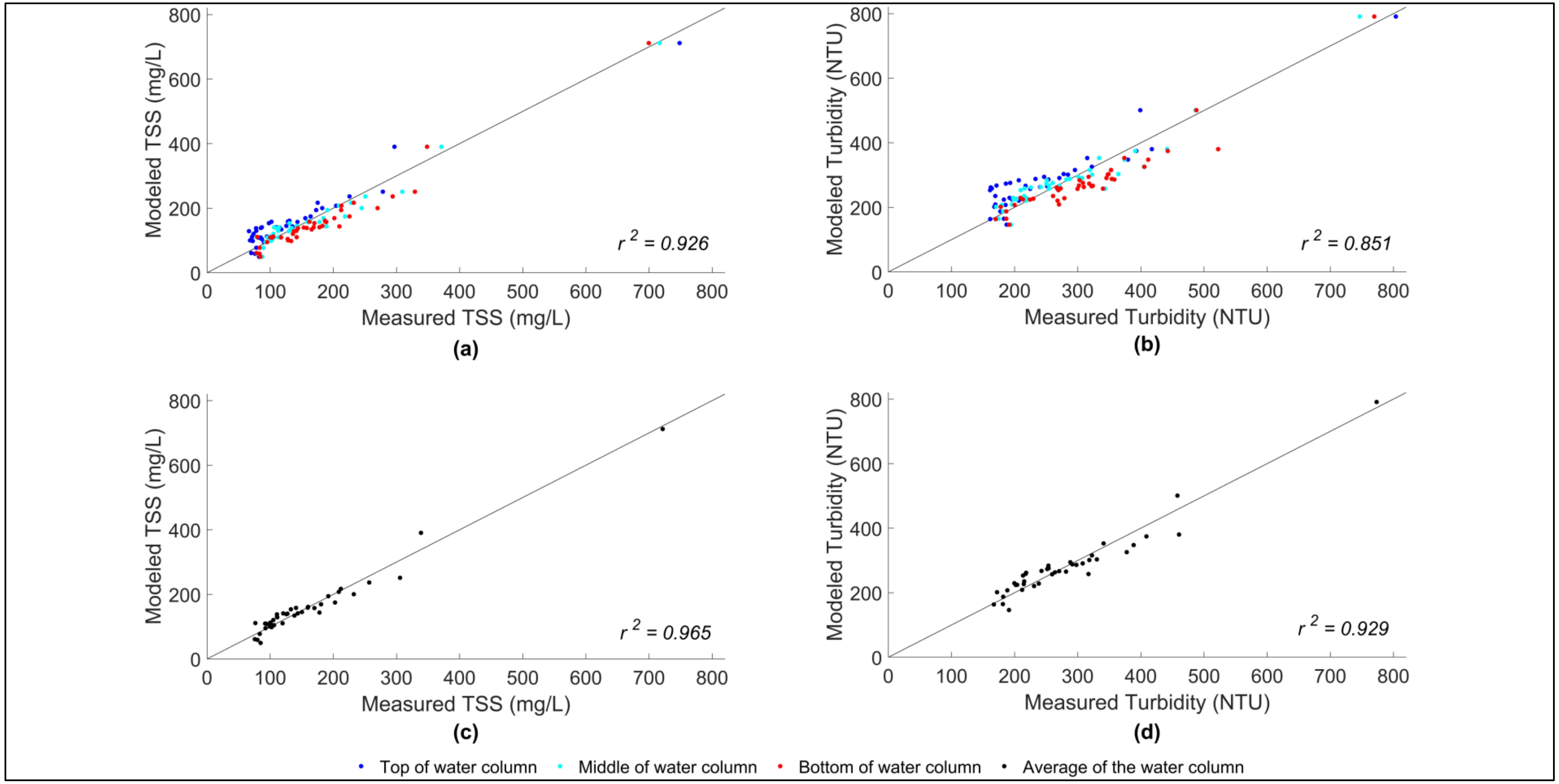
6. Interpreting Results

- Example: Turbidity vs. Total Suspended Solids (Mumtaz et al., 2011)
- $R^2 = 0.88$
- *p-value* < 0.05
- *Turbidity = 118.08 + 0.832*TSS*



6. Interpreting Results

Example: Modelled vs. measured values of TSS and NTU (Prior et al., 2020)



6. Interpreting Results

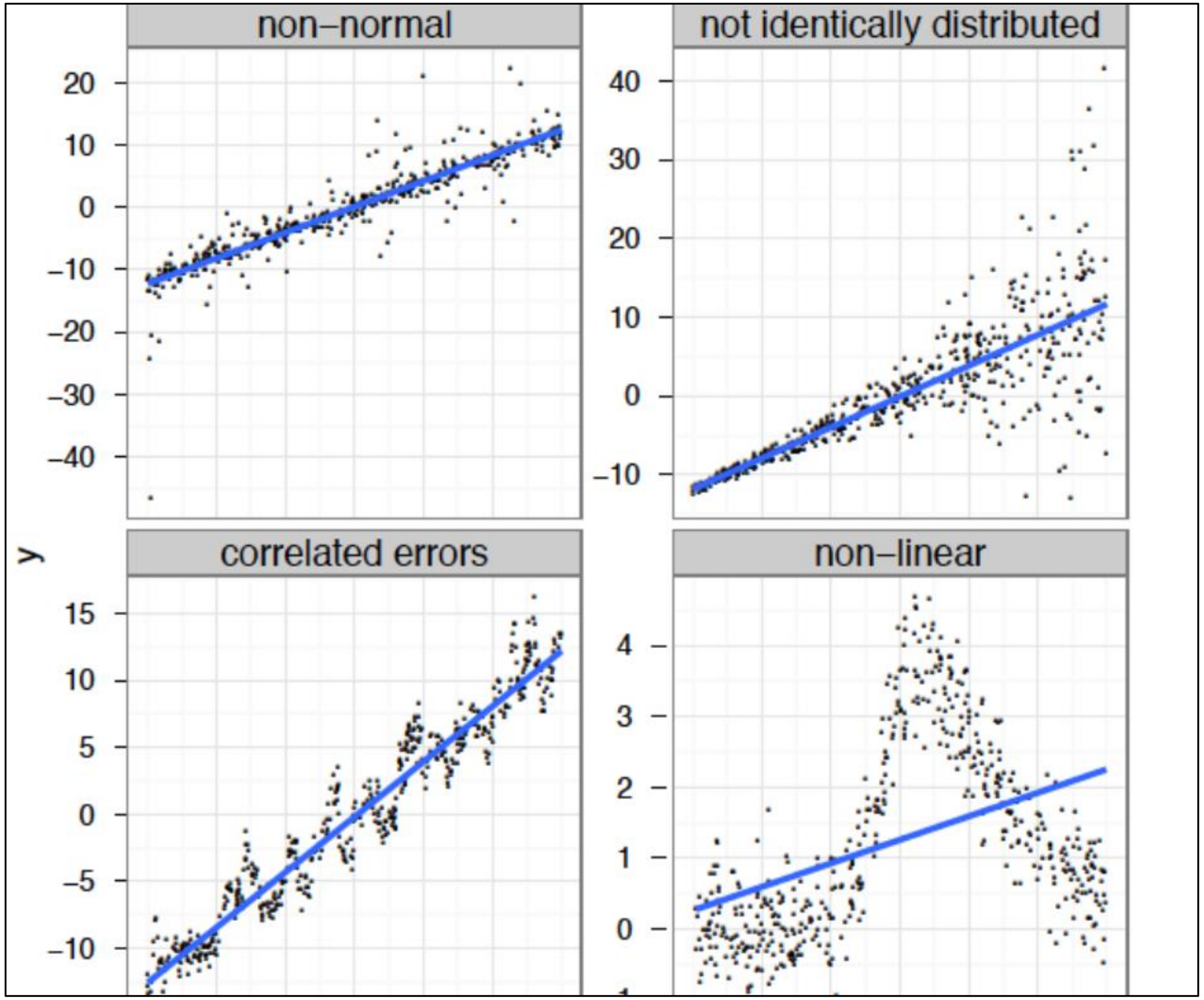
Example: Modelled vs. measured values of TSS and NTU (Prior et al., 2020)

Averaging improves accuracy (i.e., higher R^2), model performance, and bias

	Intercept	Sample Size, n	r^2	RMSE	RPD	MNB (%)
TSS	–319.760	60	0.93	30.7	3.6	4.2
	Intercept	Sample Size, n	r^2	RMSE	RPD	MNB (%)
Averaged TSS	–319.775	20	0.97	21.8	5.0	1.5
	Intercept	Sample Size, n	r^2	RMSE	RPD	MNB (%)
Turbidity	–328.016	60	0.85	44.6	2.5	2.9
	Intercept	Sample Size, n	r^2	RMSE	RPD	MNB (%)
Averaged Turbidity	–328.013	20	0.93	30.9	3.5	1.2

7. Key Assumptions

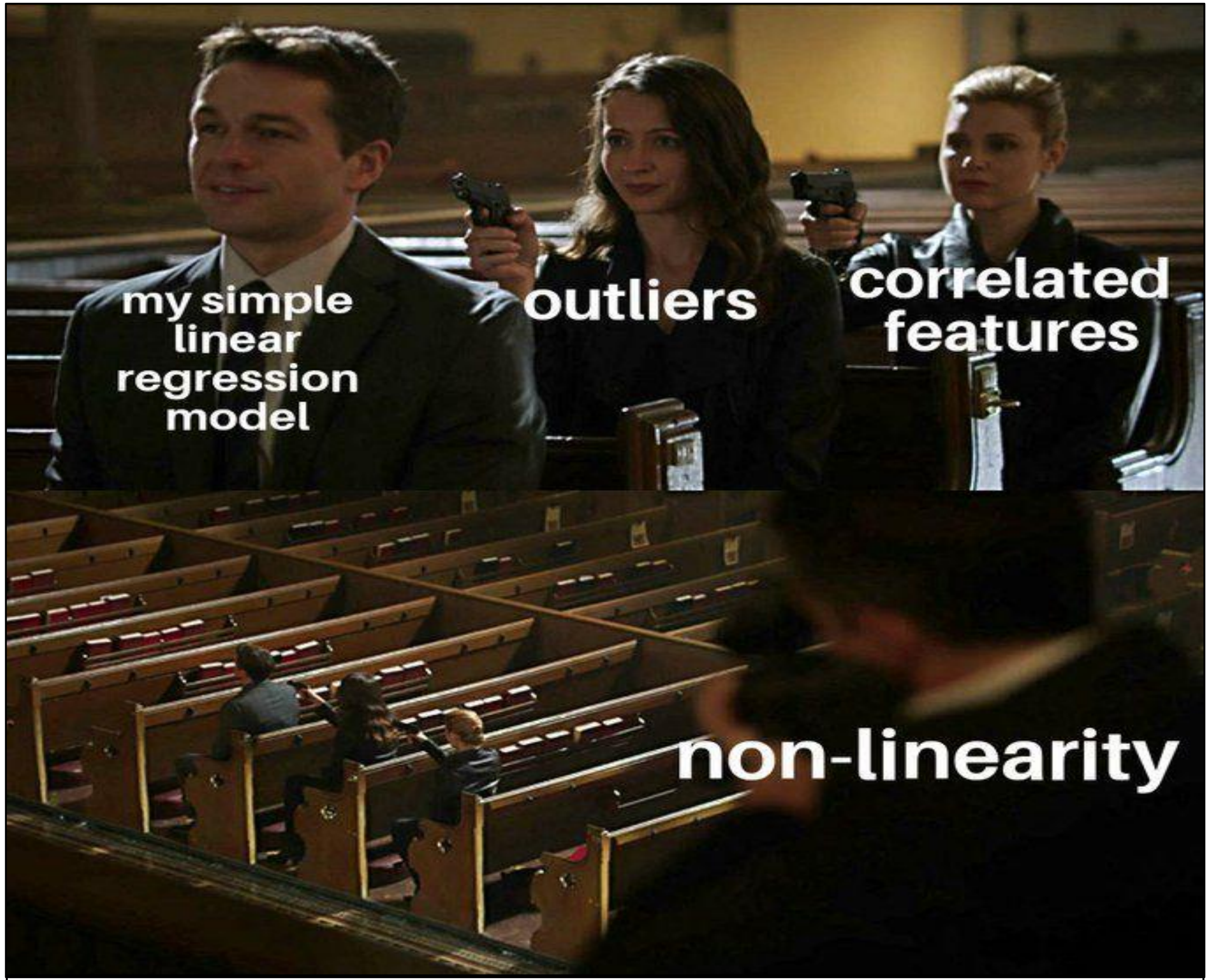
- Linear relationship between X and Y
- Errors are independent
- Error is normally distributed
- Homoscedasticity (equal variance)



https://r.qcbs.ca/workshop08/book-en/intro-linear-models.html

8. Limitations and Practical Considerations

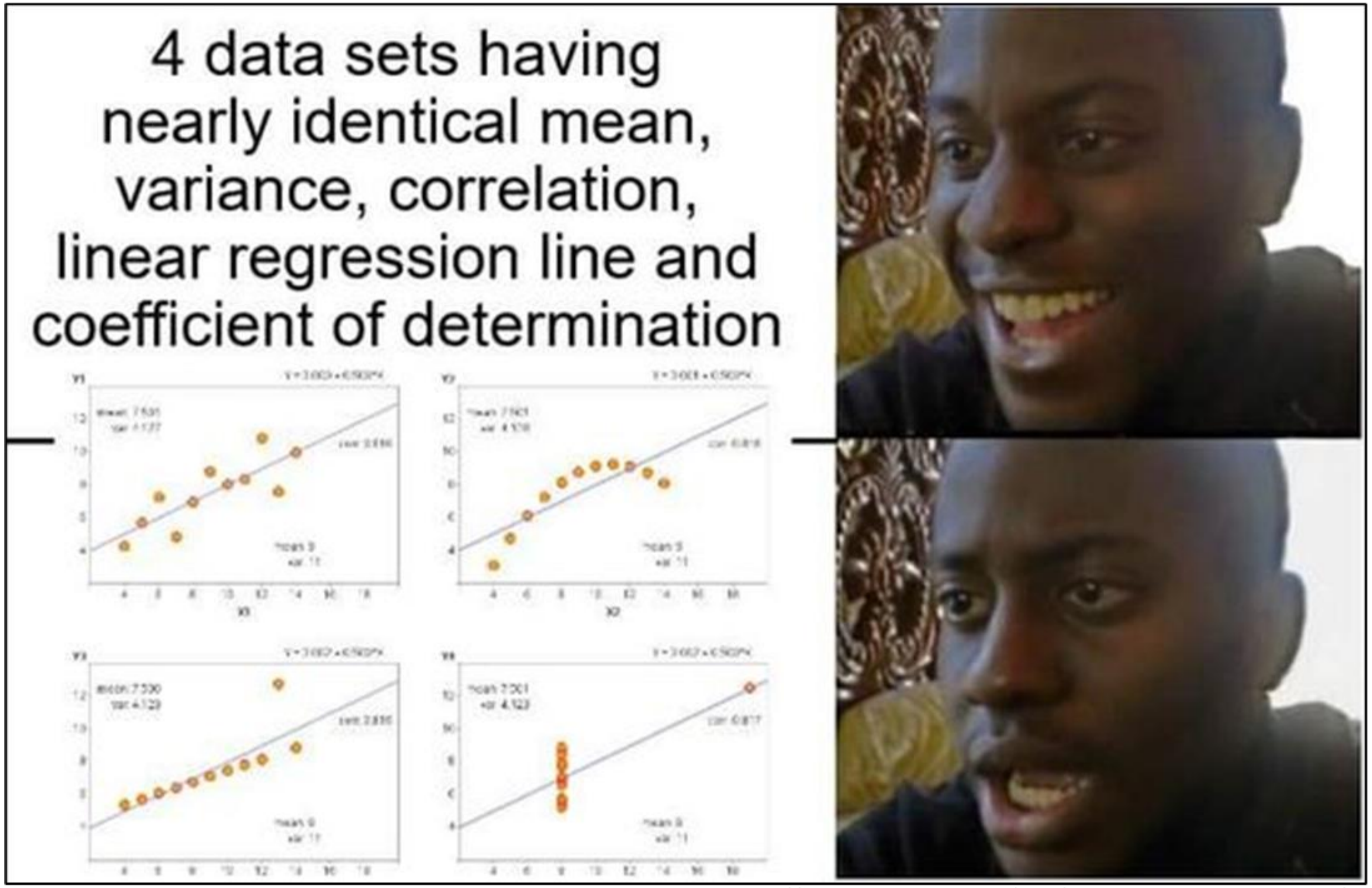
- Extrapolation issues:
  - Difficult to make predictions beyond the range of observed data
- Influence of outliers:
  - Can influence the intercept and slope
- Interpretation pitfalls:
  - Correlation not causation



https://www.kaggle.com/discussions/general/431681

8. Limitations and Practical Considerations

- Anscombe's Quartet
  - Importance of visualizing data



https://in.pinterest.com/pin/729935052119746525/

Quiz

- Why do we square the errors?
  - To account for positive and negative deviations that could potentially cancel each other out
- What is the mean value of y when x equals zero?
  - The estimate of the intercept
- What is the difference between simple linear regression and linear regression?
  - Simple linear regression models the relationship between a single X and single Y variable.
  - Linear regression can model the relationship between a single X and multiple Y variables.

Questions?