Linear Regression: Frequentist vs. Bayesian approaches

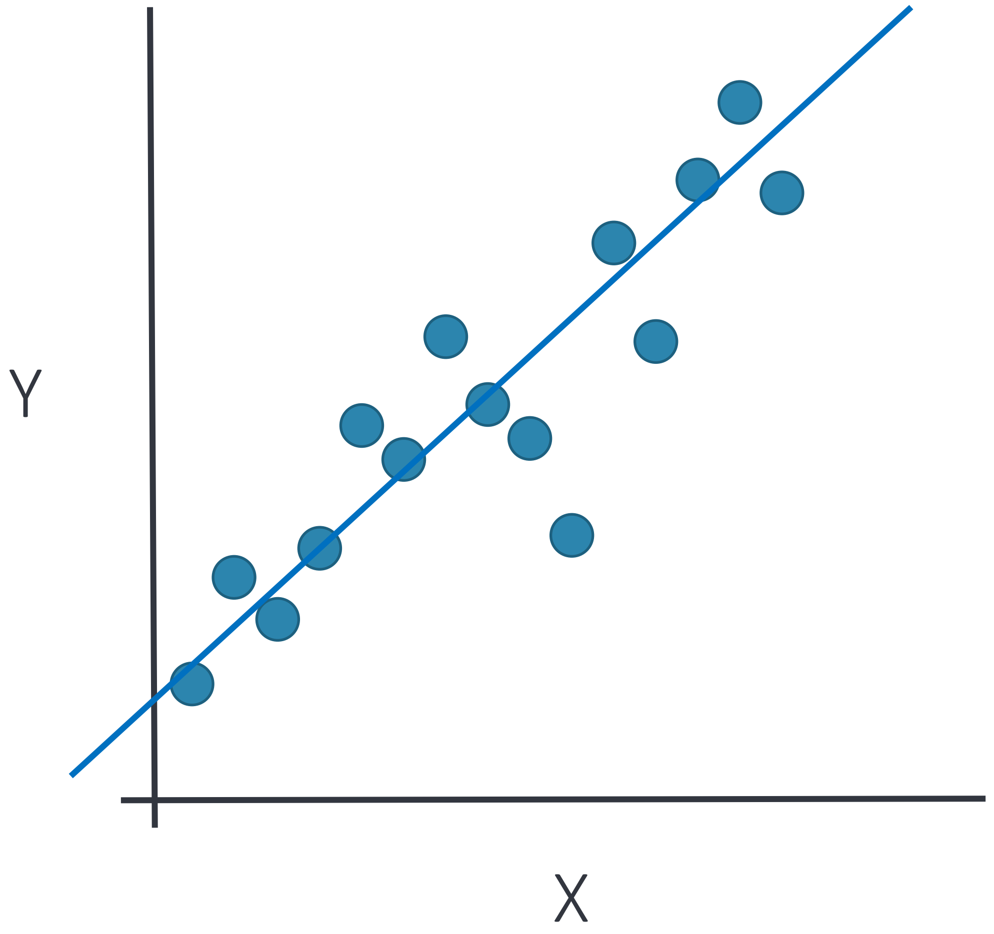
May 6, 2025

Prof. Gwendolyn Eadie

Terminology used in statistics and for regression

- X values can be called...
  - Covariates
  - Predictors
  - Explanatory variables
  - Independent variables
  - Features (computer science/machine learning)

- Y values can be called
  - Response variable
  - Dependent variable
  - Labels (ML)



Regression

- Estimate functional relationships
- Many types!
  - Linear regression, generalized linear models (GLMs), lasso regression, Poisson regression, logistic regression, …
- "standard" regression assumes X is fixed without error
- **Linear regression does not imply we are fitting a line**
  - E.g. "linear" regression means *linear in the parameters*

Concept of Regression

- *Estimating functional relationships*

"expected value" (mean)

"given"

function with parameters

errors

$$E[Y|X] = f(X, \theta) + \epsilon$$

- Note the asymmetry in most regression analysis. This is not a fit to the joint distribution of (Y, X)
- *Homoscedastic errors:* ϵ is an n-vector with σ^2
- *Heteroscedastic errors*: ϵ is an n-vector with σ_i^2 *(known or unknown)*
- Errors-in-Variables models assume X has error as well

Are these models linear or non-linear?

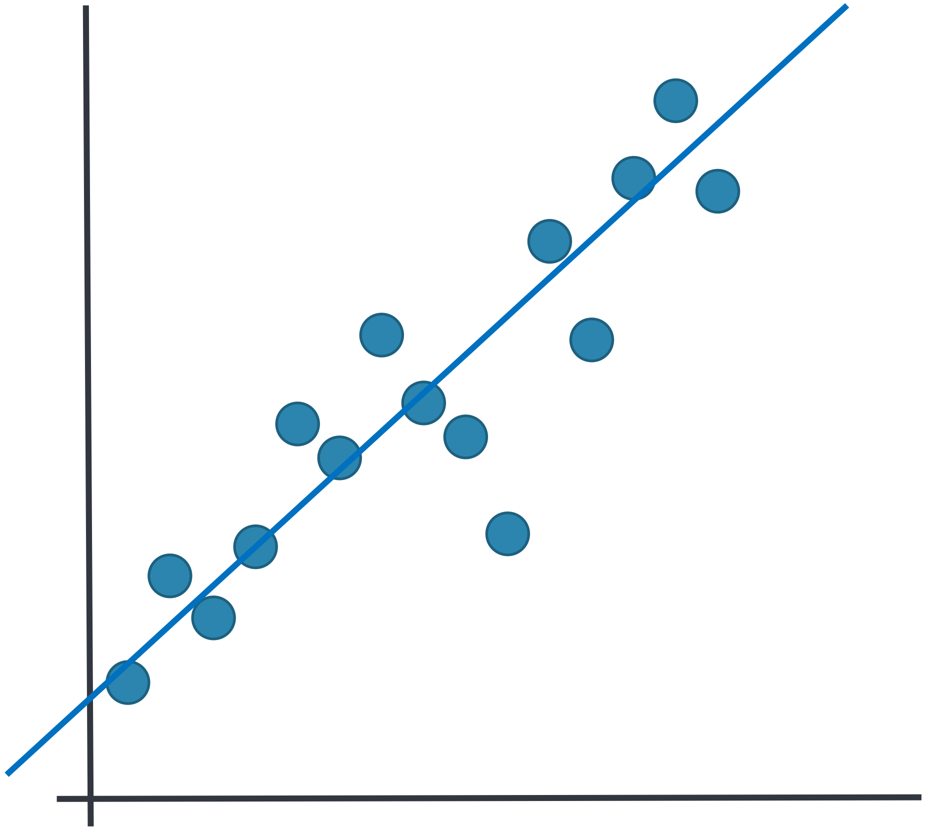
Example	Linear or Non-Linear?
$Y = \underline{\beta_0} + \underline{\beta_1} X + \underline{\beta_2} e^X + \epsilon$	linear
$Y = \left(\frac{X}{\beta_0}\right)^{-\beta_1} + \epsilon$	non-linear
$Y = \beta_0 + \beta_1 \sin(X) + \beta_2 \cos(X) + \epsilon$	linear
$Y = \beta_0 e^{-\beta_1 X} + \epsilon$	non-linear

$y = \beta_0 X + e^{-\beta_1 X}$

Conditions for linear regression

- Nearly normal residuals
  - Residuals should be normally distributed about 0
  - No trends in residuals
  - No major outlier(s) or "influential points" (we'll get to this!)
- Constant variability
  - Variability above/below least squares line shouldn't change as x changes
- Independent observations
  - e.g., typically don't apply to time series data

Frequentist:
Fitting a line using Ordinary Least Squares (OLS) Regression



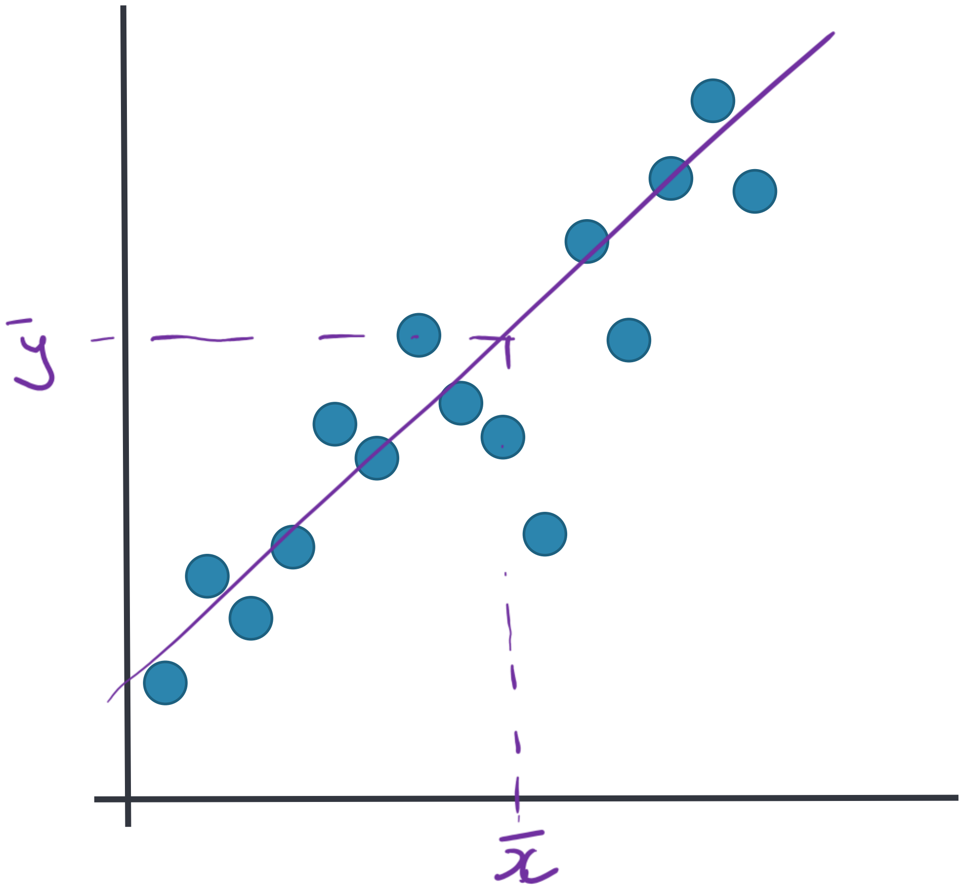
- Useful when the relationship between two quantities can be summarized by a straight line
- Correlation describes the *strength* of the correlation between x and y, whereas the regression line is used to describe the relationship between x and y
- The regression line is a model which follows the equation:

$$y = \underset{\substack{\uparrow}}{\overset{\text{intercept}}{\beta_0}} + \underset{\substack{\uparrow}}{\overset{\text{slope}}{\beta_1}} x + \varepsilon$$

parameters

$$\hat{y} = b_0 + b_1 x$$ (fitted line, $\hat{y}$ predicted expected values)

Frequentist:
Fitting a line using Ordinary Least Squares (OLS) Regression



- A regression line can tell you something about the effect of the predictor or independent variable on the response variable

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

- Slope of a regression line is related to the correlation of the points:

$$b_1 = r \frac{s_y}{s_x}$$

s_y ← sample standard deviation of y

s_x ← " " " " x

r ← correlation

- Intercept of a regression line is:

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

the line passes through $(\bar{x}, \bar{y})$

Relationship between correlation and slope (for simple L.R.)

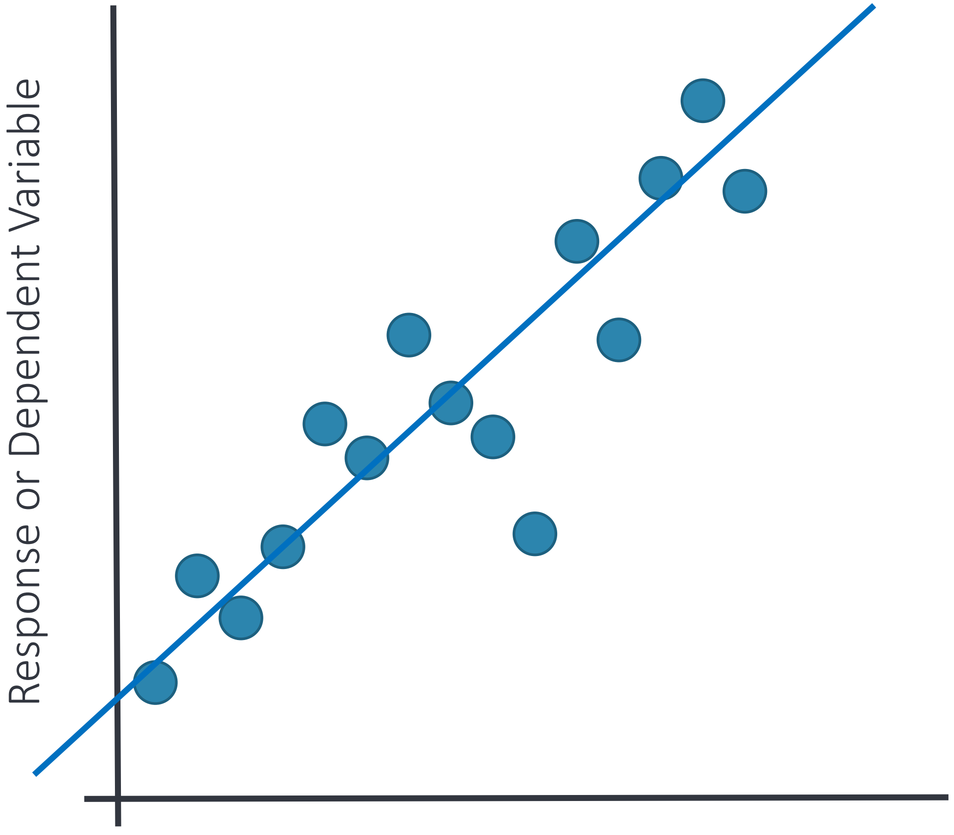
Correlation r between x and y :

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

(always between -1 and 1)

- r tells you about the *strength* of the relationship
- Slope tells you change in Y per unit change in X $\quad b_1 = r \frac{s_y}{s_x}$

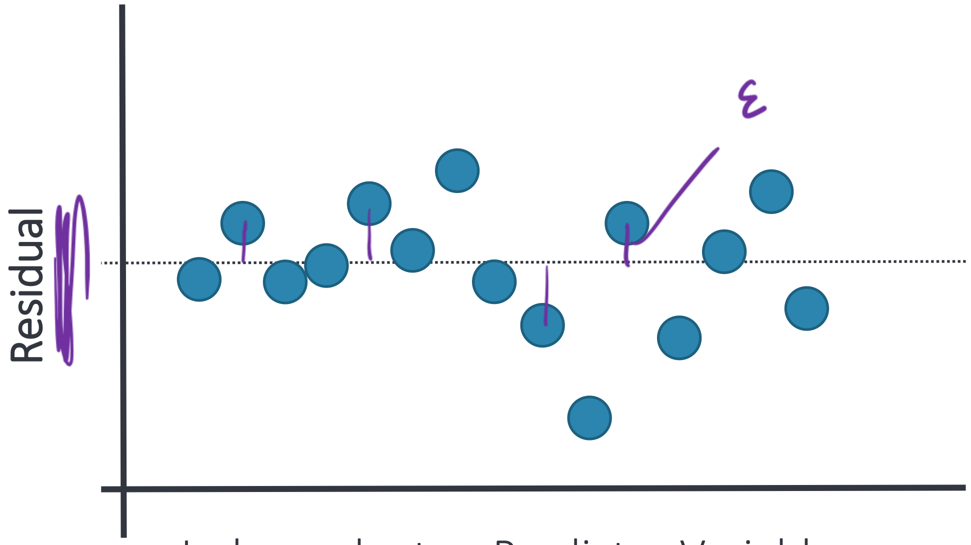
Frequentist:
How we find the least squares line



- The least squares line **minimizes the sum of squared residuals:**

$\leftarrow n$ # of data points

minimize $\sum_{i=1}^{n} \varepsilon_i^2$



Linear Regression more generally

$p \to$ # of parameters

$$Y = X\beta + \epsilon$$

Dimensions of each: $\quad Y: n \times 1 \quad X: n \times p \quad \beta: p \times 1 \quad \epsilon: n \times 1$

It helps to visualize:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \ddots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Equivalently:

More concisely:

Review of Linear Regression

- Model

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$$

- X often called the "design matrix"
- X is an $n \times p$ matrix of covariates/explanatory variables. These could be
  - Measurements
  - Fixed by design
  - Introduced to increase model flexibility
- In practice, the intercept β_0 may be encoded in X:

Least Squares estimators for β

- If the linear regression model is $Y = \beta_0 + \beta_1 X + \epsilon$, then the OLS estimator minimizes the *residual sums of squares (RSS):*

$$\min(RSS) = \min\left[\sum_{i=1}^{n}\big(\underbrace{Y_i - (\beta_0 + \beta_1 X_i)}_{\varepsilon_i}\big)^2\right]$$

- Least squares estimator of β:

$$\hat{\beta}_{LS} = \mathrm{argmin}_{\beta} \sum_{i}^{n} \left(y_i - x_i^T \beta\right)^2$$

Maximum Likelihood Estimator (MLE) for β

Assuming normally-distributed, independent errors

If we assume that $y_i \sim N(\underbrace{x_i^T\beta}_{\text{mean}}, \underbrace{\sigma^2}_{\text{variance}})$, then we can write out the likelihood function

Likelihood function

$f(y_1, y_2, \ldots, y_n | x_1, x_2, \ldots x_n; \beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - x_i^T\beta)^2}$

$y_i | X \sim N(X\beta, \sigma^2 I)$

The likelihood function is

$$L(\underbrace{\beta, \sigma^2}_{\text{parameters}}; \underset{\substack{\uparrow \\ \text{data}}}{y}) = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} \exp\left\{-\frac{1}{2}(y - X\beta)^T(y - X\beta)\right\}$$

The log-likelihood is then

$\log \mathcal{L} = \ldots$

And now we can do maximum likelihood estimation …

Maximum likelihood estimate of β

log likelihood $\rightarrow$ $\ell(\beta, \sigma^2; y) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)$

Take derivative w.r.t β and set equal to zero:

$$\frac{\partial \ell}{\partial \beta} = -\frac{1}{2n^2} X^T(y - X\beta)\Big|_{\hat{\beta}_{MLE}} = 0$$

M.L.E. of β: $\hat{\beta}_{ML} = (X^TX)^{-1}X^TY = \hat{\beta}_{LS}$

Slope and Intercept Estimators

- Under the assumption that ϵ are i.i.d., then the slope and intercept estimators are unbiased and asymptotically distributed as:

true value variance

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

assuming σ is known.

- When the true variance is unknown, the distributions of the estimators are not known (because we don't know β_1 nor σ^2)

Uncertainties in the line

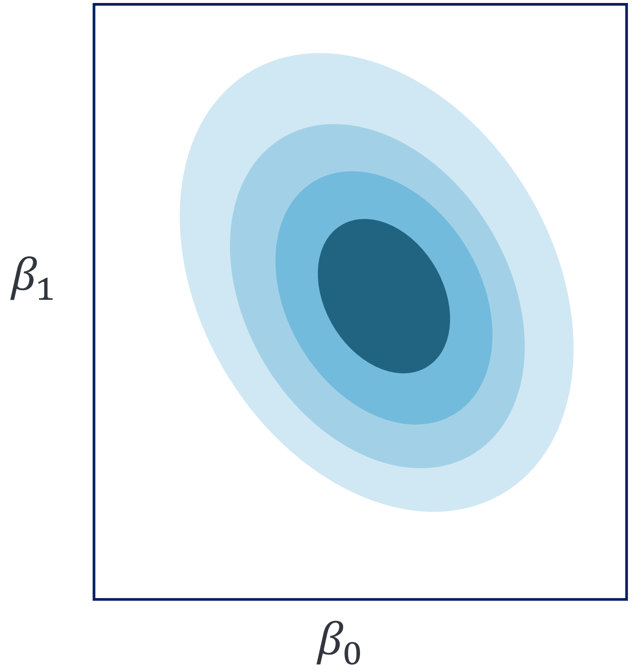
- Because we know that $\widehat{\beta_1} \sim N(\beta_1, \frac{\sigma^2}{s})$, we can use this to construct confidence intervals for Y at each X using

$$Y|x = \hat{\beta}_0 + \hat{\beta}_1 x \pm t_{\alpha/2, n-2}\, s\sqrt{1 + \frac{1}{n} \frac{(x-\bar{x})^2}{S_{xx}}}$$

→ do this for a range of x-values, and you will get confidence bands about the best-fit (OLS) line that are hyperbolas.

Point Estimates vs. Bayesian Inference

- The estimators for β_0, β_1, etc. In previous slides are *point estimates*
- A maximum likelihood estimate is also a point estimate of a parameter
- In Bayesian inference…
  - We try to infer the *distribution* for a parameter
  - We get a *posterior distribution*
  - The posterior distribution encodes all the information from
    - Prior assumptions
    - Model assumptions
    - Data
  - We report the whole posterior distribution in our results
  - We can report credible intervals to express uncertainty



Linear Regression in Bayesian context

- The model for y is

mean variance

$$y \sim N(X\beta, \sigma^2)$$

- The likelihood under this model is

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - x_i\beta)^2}$$

- We must set priors on the parameters

$p(\beta_0, \beta_1) \rightarrow$ joint prior distribution

or $\beta_0 \sim N(0, \sigma_0^2)$