Linear Regression

Linear Algebra and its Applications

Henry Kirveslahti

AI511 University of Southern Denmark

Oct 27 2025

House Keeping

- Tentative Schedule for the Rest of the Course

Theory – Exercises	Week	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Linear Regression – Lecture	44	Oct 27	Oct 28	Oct 29	Oct 30	Oct 31	Nov 01	Nov 02
Linear Regression – Exercises	45	Nov 03	Nov 04	Nov 05	Nov 06	Nov 07	Nov 08	Nov 09
Markov Chains – Lecture	46	Nov 10	Nov 11	Nov 12	Nov 13	Nov 14	Nov 15	Nov 16
Markov Chains – Exercises	47	Nov 17	Nov 18	Nov 19	Nov 20	Nov 21	Nov 22	Nov 23
Assignment 2 DL								
Assignment 3 DL								

Regression

Regression:

- Key: Variation
- Regression: Relate variation in one variable (x) to the variation in another variable (y)

Regression

Regression:

- Key: Variation
- Regression: Relate variation in one variable (x) to the variation in another variable (y)
- Example: Taller people typically weigh more - Tall figure (x) explains (to a degree) individual's weight (y)

Regression

Regression:

- Key: Variation
- Regression: Relate variation in one variable (x) to the variation in another variable (y)
- Example: Taller people typically weigh more - Tall figure (x) explains (to a degree) individual's weight (y)
- Point: We are better off guessing someone's weight if we know their height

Regression

Regression:

- Key: Variation
- Regression: Relate variation in one variable (x) to the variation in another variable (y)
- Example: Taller people typically weigh more - Tall figure (x) explains (to a degree) individual's weight (y)
- Point: We are better off guessing someone's weight if we know their height
- Another way to say this is that there is an *association* between x and y

Regression

Regression:

- Key: Variation
- Regression: Relate variation in one variable (x) to the variation in another variable (y)
- Example: Taller people typically weigh more - Tall figure (x) explains (to a degree) individual's weight (y)
- Point: We are better off guessing someone's weight if we know their height
- Another way to say this is that there is an *association* between x and y
- *All models are wrong, but some are useful* – George Box

Association and Causality

- An association need not imply causality!
- For example, towns with more churches tend to have more criminal activity

Association and Causality

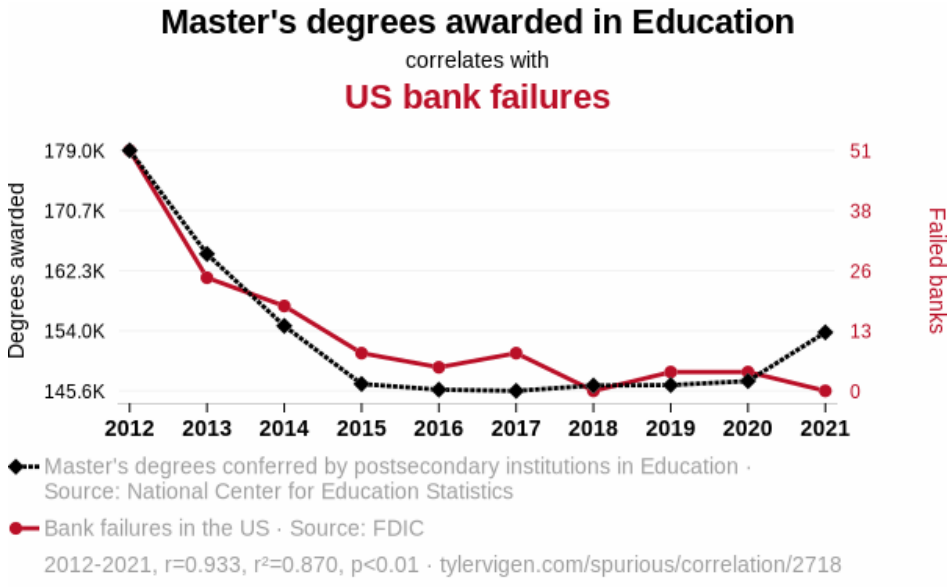
- An association need not imply causality!
- For example, towns with more churches tend to have more criminal activity
- Another association: Daily ice cream sales are correlated with drowning accidents

Association and Causality

- An association need not imply causality!
- For example, towns with more churches tend to have more criminal activity
- Another association: Daily ice cream sales are correlated with drowning accidents
- What might explain these associations?

Spurious Correlations

As the number of Master's degrees awarded in Education decreased, there were fewer people able to comprehend the concept of "fractional reserve banking." This led to a decrease in risky financial practices and ultimately contributed to a lower rate of US bank failures. After all, you can't spell "financial stability" without "STEM education" - or so the bankers now realize!



Credit goes to Tyler Vigen – See more at:
`www.tylervigen.com/spurious-correlations`

Relating variables to each other

Three approaches

1. $y = f(x)$, mathematical. Too mechanistic

Relating variables to each other

Three approaches

1. $y = f(x)$, mathematical. Too mechanistic
2. $y \approx f(x)$, an approximation (curve fitting, machine learning, topic of today)

Relating variables to each other

Three approaches

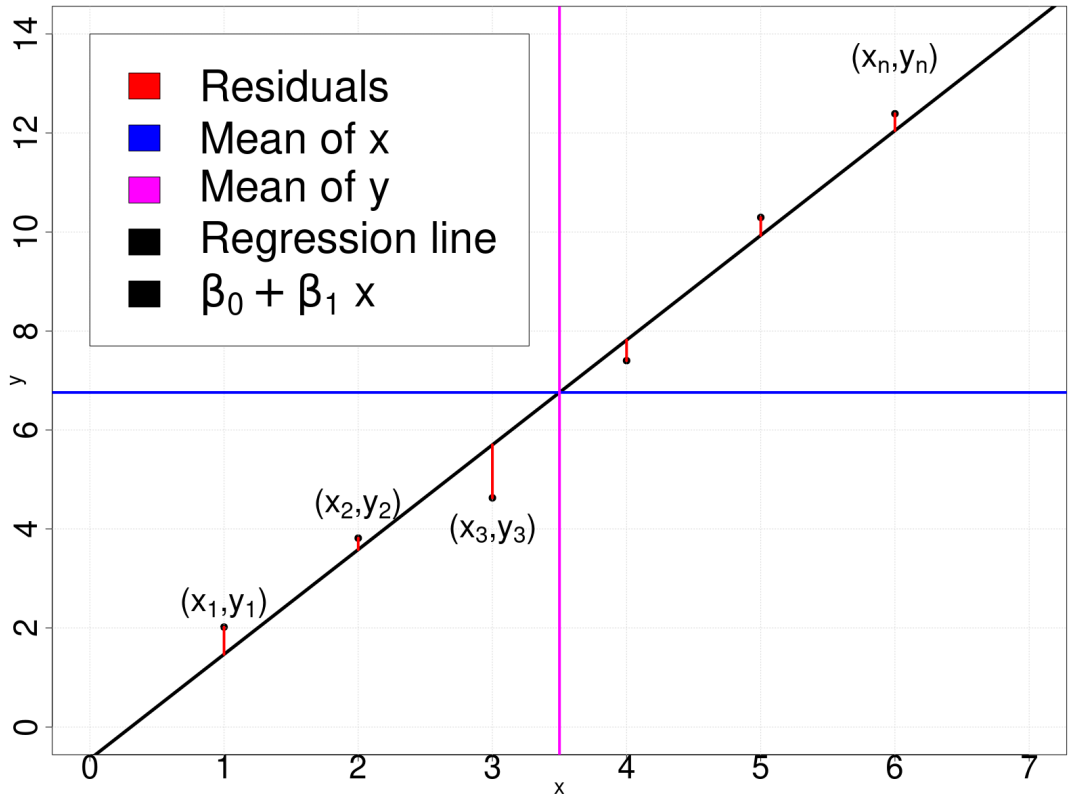
1. $y = f(x)$, mathematical. Too mechanistic
2. $y \approx f(x)$, an approximation (curve fitting, machine learning, topic of today)
3. $y_i = f(x_i) + \epsilon_i, \epsilon_i \sim D$, (statistical: very interesting, but we only get two lectures on this topic)

   NB: $\epsilon_i \sim D$ means the errors ϵ_i follow some distribution D

Simple Linear Regression

Fitting a line to explain observed y_i with observed x_i:

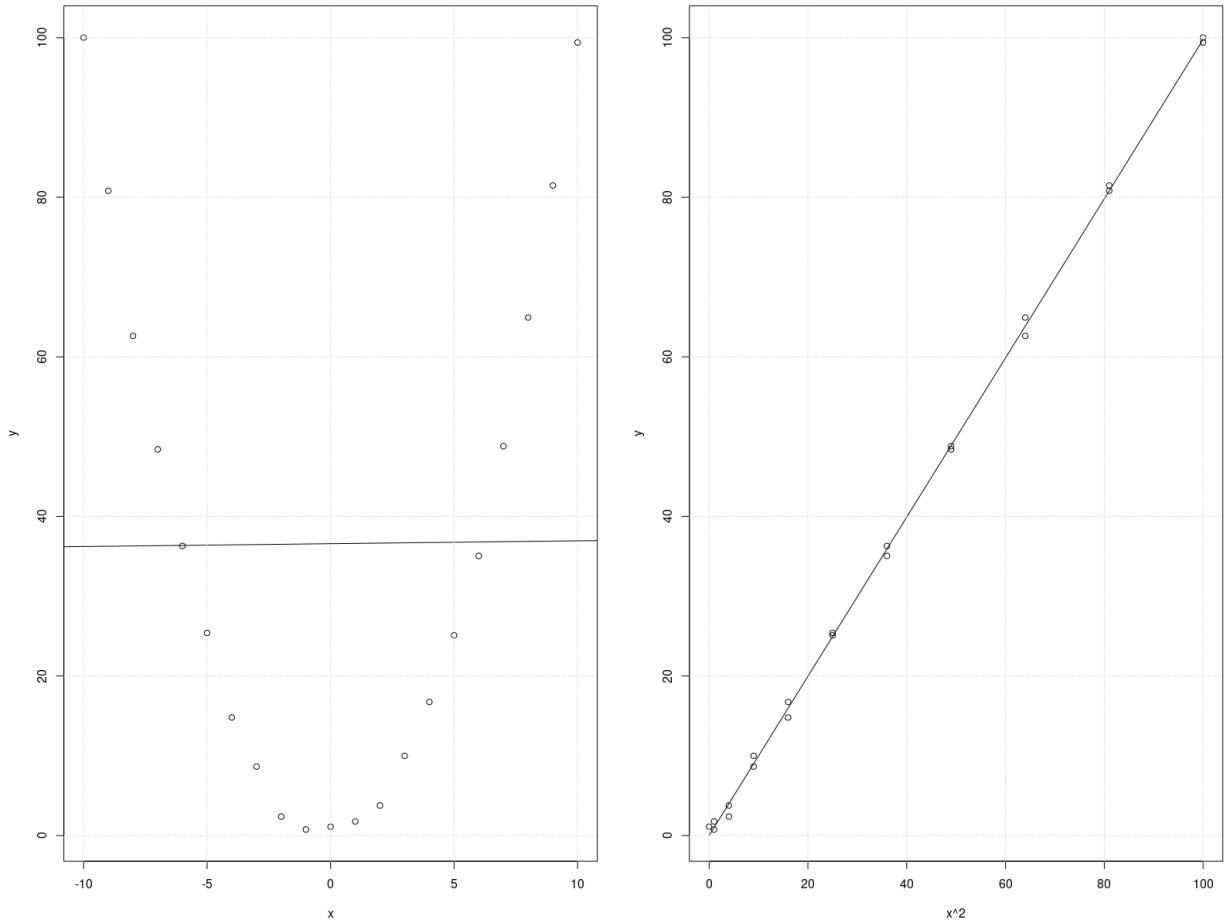
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$



Notation: Greek: Unknown quantities. Latin: Observed quantities.

What if my data is not linear?

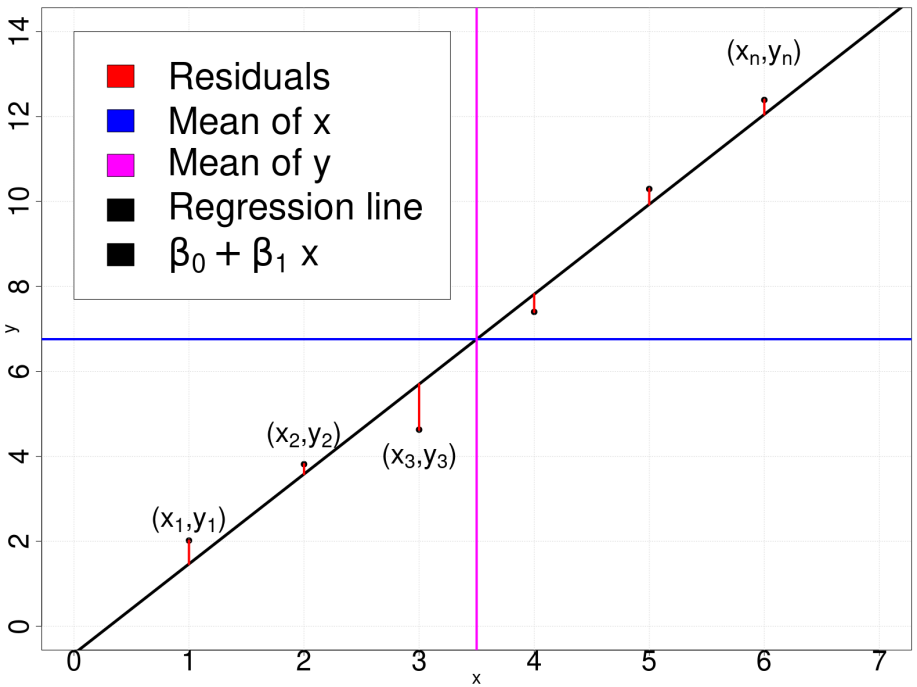
You can transform it, e.g. $z_i = x_i^2$: $\boxed{y_i = \beta_0 + \beta_1 z_i + \epsilon_i}$



Similar extensions make linear models very useful

Simple Linear Regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$



- $\beta_0 + \beta_1 x_i$ is the systemic part of the model (a line)
- The residuals[a] ϵ_i are what is left unexplained by the systemic part:
  $\epsilon_i = y_i - \beta_0 - \beta_1 x_i$

[a] Pedantly: estimates of the errors are called residuals.

The Math Approach - Minimize the RSS

Figure: C.F.Gauss proposed the OLS

Sum of squared residuals (RSS):

$$\text{argmin}_{\beta_0,\beta_1} \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Solving this gives the *OLS-estimates*[a]:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x};$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})} =: \frac{SXY}{SXX}.$$

For our purposes, these are just point estimates (numbers) – No statistics involved

[a] OLS: Ordinary Least Squares

Derivation of OLS Estimates in Simple Linear Regression

We consider the model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

Derivation of OLS Estimates in Simple Linear Regression

We consider the model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

The Ordinary Least Squares (OLS) method minimizes the sum of squared residuals:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Derivation of OLS Estimates in Simple Linear Regression

We consider the model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

The Ordinary Least Squares (OLS) method minimizes the sum of squared residuals:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Take partial derivatives and set them to zero:

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

Derivation of OLS Estimates in Simple Linear Regression - continued

From the first equation:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad \Rightarrow \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Derivation of OLS Estimates in Simple Linear Regression - continued

From the first equation:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad \Rightarrow \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Substitute into the second equation and simplify:

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Derivation of OLS Estimates in Simple Linear Regression - continued

From the first equation:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad \Rightarrow \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Substitute into the second equation and simplify:

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Thus, the OLS estimators are:

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

Why Study Linear Models?

- **Simplicity:** Linear models are easy to understand, implement, and interpret.
  - Linear relationship: output is a weighted sum of inputs.
- **Foundation:** Linear models are the building blocks of more complex models.
  - Nonlinear models often rely on linear approximations.
  - Many machine learning methods (e.g., neural networks, kernel methods) extend linear ideas.
- **Computational Efficiency:** Solving linear systems is fast and well-understood.
  - Exploits tools from linear algebra: matrix factorization, projection, orthogonality. (As we will see)
- **Interpretability:** Coefficients β_j reveal the influence of input variables x_j.

A Brief History of Linear Regression

- **Carl Friedrich Gauss (c. 1795):**
  - Developed the *method of least squares* to analyze astronomical data.
  - Introduced the idea of minimizing the sum of squared errors.
- **Adrien-Marie Legendre (1805):**
  - Independently published the method of least squares.
  - Used it to fit orbits of celestial bodies.
- **Francis Galton (1886):**
  - Studied heredity; coined the term *regression* in the phrase "regression toward mediocrity."
  - Observed that children's heights tend to regress toward the mean of the population.
- **Ronald A. Fisher (1920s):**
  - Formalized linear regression in statistics.
  - Introduced the concept of *analysis of variance (ANOVA)*.

Digression: Interpreting coefficients

- Let $y = \beta_0 + \beta_1 x + \epsilon_i$, and the epsilons are iid noise
- Statistically speaking, $\beta_0 + \beta_1 x$ is the *conditional expectation* of y at $X = x$ – it is the population mean
- In particular, β_0 is the population average of y at $x = 0$
- Same way, β_1 tells how much y changes, on average, for one unit increase in x
- In the earlier parlance, β_1 is the *association* between x and y: $\beta_1 \approx 0$ means no association, positive means positive association
- Remember: association does not imply causality. This gets even more nuanced if the model includes additional variables (or if there are variables that are not in the model but affect the relationship)

From Simple to Multiple Linear Regression

- More often than not, we want to understand how many variables $x_1, x_2, \ldots x_p$ explain variation in response y

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$

for $i = 1, \ldots, n$.

From Simple to Multiple Linear Regression

- More often than not, we want to understand how many variables $x_1, x_2, \ldots x_p$ explain variation in response y

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$

for $i = 1, \ldots, n$.

This can be more conveniently expressed with matrices:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Short recap of Linear Algebra

Matrix as a Linear Transformation:

$$A \in \mathbb{R}^{n \times p} \quad \text{defines a linear map} \quad A : \mathbb{R}^p \to \mathbb{R}^n$$

- For any vector $x \in \mathbb{R}^p$, the product $Ax \in \mathbb{R}^n$ is the image of x under the linear transformation A – In other words, A takes a p-vector, and spits out an n-vector

Matrix Product: Composition of Linear Maps

$$C = AB \quad \text{means} \quad C(x) = A(Bx)$$

- If $B : \mathbb{R}^k \to \mathbb{R}^p$ and $A : \mathbb{R}^p \to \mathbb{R}^n$, then $AB : \mathbb{R}^k \to \mathbb{R}^p \to \mathbb{R}^n$
- Matrix multiplication is composition of linear maps - this is in some sense the modern formulation of matrix products!

Matrices as Linear Maps - Continued

Let $A : \mathbb{R}^p \to \mathbb{R}^n$ be a linear map (matrix)

Transpose: Dual (Adjoint) Map

$$A^T : \mathbb{R}^n \to \mathbb{R}^p$$

- Reverses the direction of the map, flips rows and columns
- Also, if $B : \mathbb{R}^n \to \mathbb{R}^p$ and $A : \mathbb{R}^p \to \mathbb{R}^q$, then $AB : \mathbb{R}^n \to \mathbb{R}^p \to \mathbb{R}^q$
- Then $(AB)^T : \mathbb{R}^q \to \mathbb{R}^p \to \mathbb{R}^n$ is given by $= B^T A^T$ – this is easy to remember when one considers the composition and the fact that the transpose reverses the arrow!
- In particular, If X is $n \times p$ matrix, then $X^T X$ is an $p \times p$ matrix - a map $\mathbb{R}^p \to \mathbb{R}^p$ that factors through $\mathbb{R}^n$

Matrices as Linear Maps - Continued

Inverse (if it exists): Undoing the Map

$$A^{-1} : \mathbb{R}^n \to \mathbb{R}^p \quad \text{s.t.} \quad A^{-1}A = I$$

- Only square, full-rank matrices are invertible (so must have $p = n$)
- Interpreted as reversing the transformation
- We still have $(AB)^{-1} = B^{-1}A^{-1}$, provided this makes sense

Solving the OLS estimate geometrically

Goal: Solve the least squares problem

$$\hat{\beta} = \arg\min_{\beta} \|X\beta - y\|_2^2$$

Observation: Whatever the value of β is, $X\beta$ has to live in the column space of X.
Another Observation:

$$\min_{\beta} \|X\beta - y\|_2^2$$

is the squared distance of y (which is a vector in $\mathbb{R}^n$) from the subspace of $\mathbb{R}^n$ defined by the column space of X.

The shortest distance has to be realized when ϵ is orthogonal to the column space of X

$$y = X\beta + \varepsilon$$

Geometric View of Projection onto col(X)

The error vector ε is orthogonal to the column space of X if it is orthogonal to every column of X

$$\begin{aligned} y &= X\beta + \epsilon && \mid \text{multiply from left with } X^T \\ X^T y &= X^T X\beta + \underbrace{X^T \epsilon}_{=0} \end{aligned}$$

An optimal solution $\hat{\beta}$ has to satisfy the *normal equations*

$$X^T y = X^T X \hat{\beta}.$$

If $X^T X$ is invertible, we get

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

When is $\mathbf{X}^\top\mathbf{X}$ Invertible?

Condition for invertibility:

$X^\top X$ is invertible $\iff$ the columns of X are linearly independent.

That is, no column of X can be written as a linear combination of the others.
Why? Because if X is of row rank $q < p$, the vector $X\beta$ lives in a q-dimensional subspace of $\mathbb{R}^n$: You cannot possibly recover $\mathbb{R}^p$ by flattening it to a lower dimensional space.

When is $\mathbf{X}^\top\mathbf{X}$ Invertible?

Condition for invertibility:

$X^\top X$ is invertible $\iff$ the columns of X are linearly independent.

That is, no column of X can be written as a linear combination of the others.

Why? Because if X is of row rank $q < p$, the vector $X\beta$ lives in a q-dimensional subspace of $\mathbb{R}^n$: You cannot possibly recover $\mathbb{R}^p$ by flattening it to a lower dimensional space.

Equivalently:

- $\text{rank}(X) = p$ (full column rank)
- $X^\top X$ is positive definite:

$$\mathbf{z}^\top(X^\top X)\mathbf{z} > 0 \quad \forall\, \mathbf{z} \neq \mathbf{0}$$

When is $\mathbf{X}^\top\mathbf{X}$ Invertible?

Condition for invertibility:

$X^\top X$ is invertible $\iff$ the columns of X are linearly independent.

That is, no column of X can be written as a linear combination of the others.
Why? Because if X is of row rank $q < p$, the vector $X\beta$ lives in a q-dimensional subspace of $\mathbb{R}^n$: You cannot possibly recover $\mathbb{R}^p$ by flattening it to a lower dimensional space.
Equivalently:

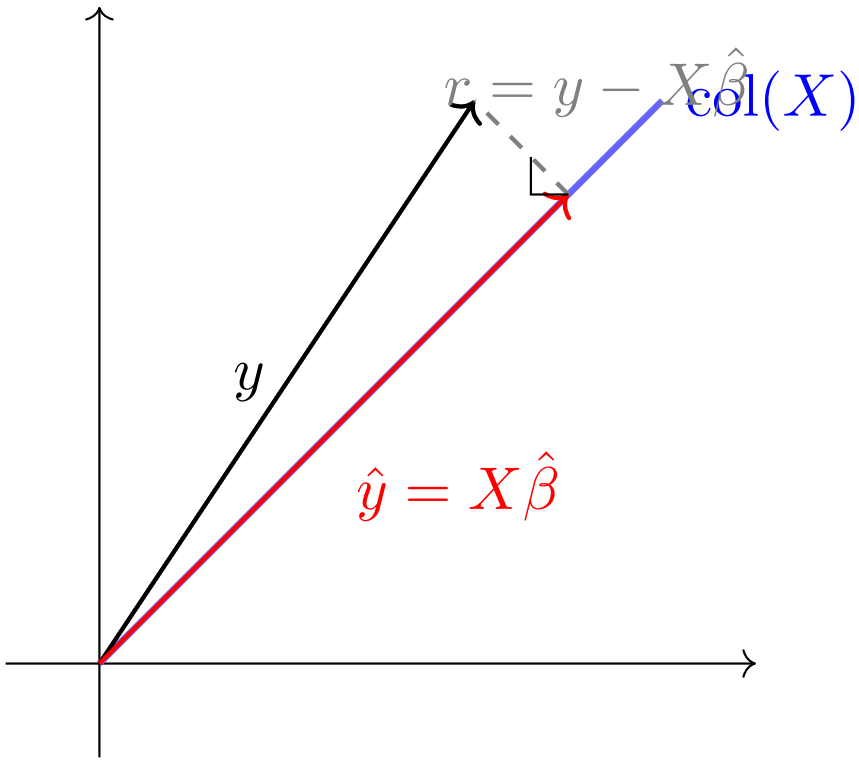
- $\text{rank}(X) = p$ (full column rank)
- $X^\top X$ is positive definite:

$$\mathbf{z}^\top(X^\top X)\mathbf{z} > 0 \quad \forall\, \mathbf{z} \neq \mathbf{0}$$

If not invertible:

- Columns of X are linearly dependent (multicollinearity)
- $X^\top X$ is singular (not full rank)
- OLS solution is not unique

Geometric View of Projection onto col(X)



- $X\hat{\beta}$ is the closest point in col(X) to y.
- The residual vector r is orthogonal to col(X).

The Projection Matrix

Orthogonal projection onto $\text{col}(X)$:

$$P = X(X^T X)^{-1} X^T \in \mathbb{R}^{n \times n}$$

- $Py = X\hat{\beta}$ is the projection of y onto $\text{col}(X)$
- This is also playfully called the *hat matrix* $\hat{y} = Py$, it gives the fitted values (systemic part of the model)

Properties of the Projection Matrix P**:**

- **Symmetric:** $P^T = P$
- **Idempotent:** $P^2 = P$
- **Orthogonal complement:** $I - P$ is also a projection matrix that projects onto the residual space (orthogonal to $\text{col}(X)$)

Interpretation:

- P "filters out" the component of y that lies in $\text{col}(X)$
- $(I - P)y$ is the residual (error) vector

The Gauss–Markov Theorem

We end by stating an interesting and important theorem about the OLS estimators. For:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \text{with } \mathbb{E}[\boldsymbol{\varepsilon}] = 0,\ \mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$$

where X is an $n \times p$ matrix of full column rank.

The Gauss–Markov Theorem

We end by stating an interesting and important theorem about the OLS estimators. For:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \text{with } \mathbb{E}[\boldsymbol{\varepsilon}] = 0,\ \mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$$

where X is an $n \times p$ matrix of full column rank.
Goal: Estimate $\boldsymbol{\beta}$ using a linear function of $\mathbf{y}$:

$$\hat{\boldsymbol{\beta}} = A\mathbf{y}, \quad \text{for some matrix } A.$$

The Gauss–Markov Theorem

We end by stating an interesting and important theorem about the OLS estimators. For:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \text{with } \mathbb{E}[\boldsymbol{\varepsilon}] = 0,\ \mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$$

where X is an $n \times p$ matrix of full column rank.
Goal: Estimate $\boldsymbol{\beta}$ using a linear function of $\mathbf{y}$:

$$\hat{\boldsymbol{\beta}} = A\mathbf{y}, \quad \text{for some matrix } A.$$

Among all linear unbiased estimators (those satisfying $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$), the OLS estimator

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (X^\top X)^{-1} X^\top \mathbf{y}$$

has the **minimum variance**.
In short: The OLS estimator is the most efficient linear unbiased estimator of $\boldsymbol{\beta}$, when the errors are independent and identically distributed.