PMAI LAB

Linear Algebra: Theory, Computation, and Modern Applications

Lecture 07 • Least Squares and Linear Regression

Course lens
Understand → Model → Compute → Apply → Code → Interpret

Audience
Undergraduate BS Mathematics students

Instructor
Dr. Hira Benish

Least squares replaces an impossible exact fit with the closest model-supported approximation.

1. Why Least Squares?

Real observations rarely satisfy a mathematical model exactly

Consider a system

$$\mathbf{A}\boldsymbol{x} = \boldsymbol{b},$$

where **A** has more rows than columns. Each row supplies an observation or condition. Noise and natural variation often make the system inconsistent, so no vector $\boldsymbol{x}$ satisfies every equation exactly.

Least squares chooses $\widehat{\boldsymbol{x}}$ that minimizes

$$\|\mathbf{A}\boldsymbol{x} - \boldsymbol{b}\|^2.$$

The fitted vector is $\widehat{\boldsymbol{b}} = \mathbf{A}\widehat{\boldsymbol{x}}$ and the residual is

$$\boldsymbol{r} = \boldsymbol{b} - \widehat{\boldsymbol{b}}.$$

Learning outcomes

By the end of this lecture, students should be able to:

- distinguish exact, inconsistent, and least-squares solutions;
- derive and solve the normal equations;
- interpret least squares as orthogonal projection;
- construct a design matrix for simple linear regression;
- compute fitted values, residuals, SSE, RMSE, and R^2;
- recognize multiple and polynomial regression as matrix models; and
- solve and visualize an energy-demand regression problem in Python.

Geometric interpretation

The columns of **A** span the set of outputs the model can produce. If $\boldsymbol{b}$ is outside Col($\mathbf{A}$), least squares projects it onto that column space:

$$\widehat{\boldsymbol{b}} = \text{proj}_{\text{Col}(\mathbf{A})}(\boldsymbol{b}).$$

The residual is orthogonal to every column:

$$\mathbf{A}^\top \boldsymbol{r} = \mathbf{0}.$$

> **Central idea**
> Least squares does not force an exact answer where none exists. It finds the achievable model output with the smallest Euclidean error and preserves the geometry of orthogonal projection.

Where it appears

Calibration, forecasting, curve fitting, image reconstruction, navigation, finance, epidemiology, experimental science, machine learning, and data analytics all use least-squares models.

2. Normal Equations

Orthogonality turns an approximation problem into a solvable linear system

At the best approximation, the residual

$$\boldsymbol{r} = \boldsymbol{b} - \mathbf{A}\widehat{\boldsymbol{x}}$$

is orthogonal to the column space of $\mathbf{A}$. Hence

$$\mathbf{A}^\top(\boldsymbol{b} - \mathbf{A}\widehat{\boldsymbol{x}}) = \mathbf{0}.$$

Rearranging gives the **normal equations**:

$$\boxed{\mathbf{A}^\top\mathbf{A}\widehat{\boldsymbol{x}} = \mathbf{A}^\top\boldsymbol{b}.}$$

If the columns of $\mathbf{A}$ are independent, then $\mathbf{A}^\top\mathbf{A}$ is invertible and

$$\widehat{\boldsymbol{x}} = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\boldsymbol{b}.$$

Projection matrix

The fitted vector can be written as

$$\widehat{\boldsymbol{b}} = \mathbf{P}\boldsymbol{b}, \qquad \mathbf{P} = \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top.$$

The projection matrix is symmetric and idempotent:

$$\mathbf{P}^\top = \mathbf{P}, \qquad \mathbf{P}^2 = \mathbf{P}.$$

A small inconsistent system

Let

$$\mathbf{A} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \qquad \boldsymbol{b} = \begin{pmatrix} 2 \\ 3 \\ 7 \end{pmatrix}.$$

The exact system asks one number x to equal 2, 3, and 7 simultaneously. The normal equation is

$$\begin{bmatrix} 3 \end{bmatrix} \widehat{x} = \begin{bmatrix} 12 \end{bmatrix}, \qquad \widehat{x} = 4.$$

Thus

$$\widehat{\boldsymbol{b}} = \begin{pmatrix} 4 \\ 4 \\ 4 \end{pmatrix}, \qquad \boldsymbol{r} = \begin{pmatrix} -2 \\ -1 \\ 3 \end{pmatrix}, \qquad \mathbf{A}^\top\boldsymbol{r} = 0.$$

The least-squares constant is the arithmetic mean of the observations.

Residual check
Always verify $\mathbf{A}^\top\boldsymbol{r} = \mathbf{0}$. It confirms that no small movement within the model space can further reduce the squared error.

3. Simple Linear Regression

Fit a line to observations using a least-squares design matrix

For observations (x_i, y_i), assume

$$y_i \approx \beta_0 + \beta_1 x_i.$$

The parameter vector and design matrix are

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \qquad \mathbf{A} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \qquad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

The first column of ones creates the intercept; the second column represents the predictor.

Worked data

Consider

$$(1, 2), \qquad (2, 2.5), \qquad (3, 4).$$

Then

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, \qquad \boldsymbol{y} = \begin{pmatrix} 2 \\ 2.5 \\ 4 \end{pmatrix}.$$

The normal equations are

$$\begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix} \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} = \begin{pmatrix} 8.5 \\ 19 \end{pmatrix}.$$

Solving gives

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 5/6 \\ 1 \end{pmatrix} \approx \begin{pmatrix} 0.8333 \\ 1.0000 \end{pmatrix}.$$

Therefore the fitted line is

$$\boxed{\widehat{y} = 0.8333 + 1.0000x.}$$

Interpretation

- $\widehat{\beta}_0 = 0.8333$ is the predicted response when $x = 0$; it is meaningful only if $x = 0$ is relevant to the context.
- $\widehat{\beta}_1 = 1$ means the predicted response increases by one unit when x increases by one unit.

Association is not causation
A fitted slope describes association in the observed data. It does not by itself prove that changing x causes a change in y; experimental design and domain evidence are required.

4. Residuals and Model Evaluation

A fitted equation is incomplete without an examination of error

For the worked line $\widehat{y} = 5/6 + x$, the fitted values are

$$\widehat{\boldsymbol{y}} = \begin{pmatrix} 11/6 \\ 17/6 \\ 23/6 \end{pmatrix} \approx \begin{pmatrix} 1.8333 \\ 2.8333 \\ 3.8333 \end{pmatrix}.$$

The residual vector is observed minus fitted:

$$\boldsymbol{r} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = \begin{pmatrix} 1/6 \\ -1/3 \\ 1/6 \end{pmatrix}.$$

Because an intercept is included, the residuals sum to zero.

Common error measures

$$\begin{aligned} \text{SSE} &= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \|\boldsymbol{r}\|^2, \\ \text{MSE} &= \frac{1}{n}\text{SSE}, \\ \text{RMSE} &= \sqrt{\text{MSE}}, \\ R^2 &= 1 - \frac{\text{SSE}}{\sum_{i=1}^{n}(y_i - \overline{y})^2}. \end{aligned}$$

For this model,

$$\text{SSE} = \frac{1}{6} \approx 0.1667, \quad \text{RMSE} = \sqrt{\frac{1}{18}} \approx 0.2357, \quad R^2 = \frac{12}{13} \approx 0.9231.$$

How to read these values

Quantity	What it measures	Caution
Residual	Error for one observation	Sign depends on observed-minus-fitted convention
SSE	Total squared training error	Grows with data size and response scale
RMSE	Typical error in response units	Sensitive to large errors
R^2	Fraction of variation explained relative to the mean model	High R^2 does not establish causality or good future prediction

Residual diagnostics
Residuals should be examined against fitted values or predictors. Visible curves, changing spread, or extreme points may indicate that the linear model is inadequate.

5. Extensions and Numerical Practice

The same matrix framework supports richer predictive models

Multiple linear regression

With predictors $x_1, \ldots, x_p$,

$$\widehat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Each predictor becomes a column of the design matrix. The least-squares problem remains

$$\min_{\boldsymbol{\beta}} \|\mathbf{A}\boldsymbol{\beta} - \boldsymbol{y}\|^2.$$

Polynomial regression

A curved relationship can still be linear in its parameters:

$$\widehat{y} = \beta_0 + \beta_1 x + \beta_2 x^2.$$

Use the design matrix

$$\mathbf{A} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}.$$

The model is nonlinear in x but linear in the unknown coefficients.

Rank deficiency

If one predictor column is a linear combination of others, $\mathbf{A}^\top\mathbf{A}$ is singular and the coefficient vector is not unique. Redundant features should be removed or handled using suitable regularization.

Numerical stability

The formula $(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\boldsymbol{y}$ is mathematically useful, but explicitly forming an inverse may magnify numerical error. In Python, prefer

`np.linalg.lstsq(A, y, rcond=None)`.

QR factorization or SVD is generally more stable than solving through an explicit inverse.

Regularization idea

Ridge regression modifies the normal equations:

$$(\mathbf{A}^\top\mathbf{A} + \lambda\mathbf{I})\widehat{\boldsymbol{\beta}} = \mathbf{A}^\top\boldsymbol{y}, \qquad \lambda > 0.$$

The additional term discourages excessively large coefficients and can improve stability, but introduces bias.

> **Modelling discipline**
> More predictors or higher-degree polynomials do not automatically create a better model. Compare training and validation performance, examine residuals, and interpret coefficients in context.

6. Application Case Study: Cooling-Energy Demand

Use a regression line to support short-range energy planning

An energy analyst studies how a temperature-above-baseline index x relates to a building's cooling-demand index y:

x	1	2	3	4	5
y	2	3	5	4	6

Matrix model

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix}, \qquad \boldsymbol{y} = \begin{pmatrix} 2 \\ 3 \\ 5 \\ 4 \\ 6 \end{pmatrix}, \qquad \widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix}.$$

Solving the least-squares problem gives

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 1.3 \\ 0.9 \end{pmatrix}, \qquad \boxed{\widehat{y} = 1.3 + 0.9x.}$$

Fitted values and residuals

$$\widehat{\boldsymbol{y}} = \begin{pmatrix} 2.2 \\ 3.1 \\ 4.0 \\ 4.9 \\ 5.8 \end{pmatrix}, \qquad \boldsymbol{r} = \begin{pmatrix} -0.2 \\ -0.1 \\ 1.0 \\ -0.9 \\ 0.2 \end{pmatrix}.$$

The error measures are

$$\text{SSE} = 1.9, \qquad \text{RMSE} = \sqrt{1.9/5} \approx 0.6164, \qquad R^2 = 0.81.$$

Planning prediction

For $x = 6$,

$$\widehat{y} = 1.3 + 0.9(6) = 6.7.$$

The slope suggests an average increase of 0.9 demand-index units for each one-unit increase in the temperature index within the observed operating range.

Application decision
Use 6.7 as a model-based planning estimate for $x = 6$, while retaining an uncertainty allowance because observed demand does not lie exactly on the line.

Responsible limitation
The dataset is small, and energy demand may also depend on humidity, occupancy, insulation, equipment, day of week, and operational policy. Extrapolation far beyond the observed range is not justified.

7. Python Laboratory

Fit the model, verify the normal equations, calculate metrics, and inspect residuals

```python
import numpy as np
import matplotlib.pyplot as plt

x = np.array([1., 2., 3., 4., 5.])
y = np.array([2., 3., 5., 4., 6.])

# Design matrix: intercept column followed by predictor
A = np.column_stack([np.ones_like(x), x])

# Stable least-squares solution
beta, residual_sum, rank, singular_values = np.linalg.lstsq(
    A, y, rcond=None
)
y_hat = A @ beta
residuals = y - y_hat

sse = residuals @ residuals
rmse = np.sqrt(np.mean(residuals**2))
sst = np.sum((y - y.mean())**2)
r_squared = 1 - sse / sst
normal_equation_error = A.T @ residuals

x_new = 6.0
y_new = beta[0] + beta[1] * x_new

print("Coefficients [intercept, slope]:", np.round(beta, 4))
print("Fitted values:", np.round(y_hat, 4))
print("Residuals:", np.round(residuals, 4))
print("SSE:", round(float(sse), 4))
print("RMSE:", round(float(rmse), 4))
print("R-squared:", round(float(r_squared), 4))
print("A.T @ residuals:", np.round(normal_equation_error, 10))
print("Prediction at x=6:", round(float(y_new), 4))

grid = np.linspace(0.5, 6.0, 100)
line = beta[0] + beta[1] * grid

fig, axes = plt.subplots(1, 2, figsize=(10, 3.8))
axes[0].scatter(x, y, s=65, color="#1665D8", label="Observed")
axes[0].plot(grid, line, color="#0C8F8F", label="Least-squares line")
axes[0].scatter([x_new], [y_new], color="#E5A823", label="Prediction")
axes[0].set_xlabel("Temperature-above-baseline index")
axes[0].set_ylabel("Cooling-demand index")
axes[0].legend()

axes[1].axhline(0, color="grey", lw=0.8)
axes[1].scatter(x, residuals, s=65, color="#1665D8")
axes[1].set_xlabel("Temperature index")
axes[1].set_ylabel("Residual")
axes[1].set_title("Residual plot")
plt.tight_layout()
plt.show()
```

Selected output

```
Coefficients [intercept, slope]: [1.3 0.9]
SSE: 1.9    RMSE: 0.6164    R-squared: 0.81
A.T @ residuals: [0. 0.]
Prediction at x=6: 6.7
```

Interpretation

The zero normal-equation error verifies orthogonality. The two plots show both the fitted relationship and the observation-level deviations that the line does not explain.

8. Concept Check and Practice

Common misconceptions

Misconception	Why it fails	Correct practice
Least squares makes every residual zero	It minimizes total squared residuals	Inspect fitted values and residuals
A high R^2 proves causation	R^2 measures fit, not causal design	Use domain and experimental evidence
An explicit matrix inverse is always best	It may be inefficient and unstable	Use a least-squares solver or QR/SVD
More predictors always improve the model	Training fit can improve while generalization worsens	Validate and control complexity

Three concept-check MCQs

1. A least-squares residual satisfies:

A. $\mathbf{A}\boldsymbol{r} = \mathbf{0}$

B. $\mathbf{A}^\top \boldsymbol{r} = \mathbf{0}$

C. $\boldsymbol{r} = \boldsymbol{b}$

D. $\mathbf{A}^\top \boldsymbol{b} = \mathbf{0}$

2. The first column of ones in simple regression represents the:

A. residual

B. slope

C. intercept

D. response

3. Which Python method is preferred here for numerical least squares?

A. `np.linalg.det`

B. `np.linalg.lstsq`

C. `np.sort`

D. `np.trace`

MCQ answers: 1-B, 2-C, 3-B

Practice exercises

1. Find the least-squares constant for observations $2, 5, 8$.
2. Solve the normal equation for $\mathbf{A} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ and $\boldsymbol{b} = (1, 4, 4)^\top$.
3. Construct the design matrix for data $(0, 1), (1, 3), (2, 4)$.
4. Fit a line to $(0, 1), (1, 2), (2, 3)$ and interpret its slope.
5. For observed values $(2, 4)^\top$ and fitted values $(3, 3)^\top$, find residuals, SSE, and RMSE.
6. Write the design matrix for a quadratic model fitted at $x = 1, 2, 3$.
7. Add the energy observation $(6, 7.5)$, refit the model in Python, and compare the slope and R^2.

9. Scenario Challenge, Answers, and Takeaways

Scenario-based problem: machine energy use

A factory records machine-load index $x = (0, 1, 2, 3)^\mathsf{T}$ and energy-use index $y = (1, 2, 2, 4)^\mathsf{T}$. Fit $\widehat{y} = \beta_0 + \beta_1 x$, calculate SSE, and predict energy use at $x = 4$.

Brief answers to practice

Exercise	Brief answer
1	Mean $= (2 + 5 + 8)/3 = 5$
2	$\widehat{x} = 3$ and $\widehat{\boldsymbol{b}} = (3, 3, 3)^\mathsf{T}$
3	$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$ and $\boldsymbol{y} = (1, 3, 4)^\mathsf{T}$
4	$\widehat{y} = 1 + x$; slope 1 means one predicted response unit per input unit
5	Residuals $= (-1, 1)^\mathsf{T}$; SSE $= 2$; RMSE $= 1$
6	$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix}$
7	$\widehat{y} \approx 1.0333 + 1.0143x$ and $R^2 \approx 0.8909$
Scenario	$\widehat{y} = 0.9 + 0.9x$; residuals $= (0.1, 0.2, -0.7, 0.4)^\mathsf{T}$; SSE $= 0.7$; prediction at $x = 4$ is 4.5

Lecture summary

- Least squares minimizes the squared Euclidean residual norm.
- Normal equations express orthogonality between residuals and predictor columns.
- Linear regression is a least-squares model with an intercept and predictor columns.
- Residuals, RMSE, and R^2 describe different aspects of model fit.
- Multiple and polynomial regression use the same design-matrix framework.
- Stable solvers should be preferred over explicitly forming a matrix inverse.

Research and career connection

Least squares supports data analytics, forecasting, econometrics, engineering calibration, computer vision, geodesy, experimental science, epidemiology, finance, optimization, and machine learning.

Key terms: least squares • normal equations • design matrix • intercept • slope • fitted value • residual • SSE • RMSE • R^2 • regression • regularization

Suggested references: Gilbert Strang, Introduction to Linear Algebra; David C. Lay, Steven R. Lay, and Judi J. McDonald, Linear Algebra and Its Applications; NumPy least-squares documentation; Matplotlib scatter documentation.