MA477: Data Science

Lesson 6 Outline — 21 January 2026

United States Military Academy, West Point
Instructor: MAJ Patrick Kuiper

1 Administrative

- Quiz
- PSET 1 Discussion
- Project 1 Discussion
- Student review
- Lecture
- Linear Regression / CV exercise

2 Regression Lesson Objectives

- Understand the linear regression model (parameters and estimation) and its usage (see important questions in 3.2.2)
- Assess model accuracy using common metrics.

3 Student Review

4 Key Terms

- The **Sum of Square Error** (SSE) is a measure of error for your model to the data $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- The **Sum of Square Total** (SST) is a measure of un-normalized deviation of the data SST $= \sum_{i=1}^{n}(y_i - \bar{y})^2$
- The **Standard Deviation** measures the typical spread of the data around the mean. $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$
- The **The Coefficient of Determination** (R^2) is the percentage of the total observed variation in the response variable that is accounted for by changes in the explanatory variable $R^2 = 1 - \frac{SSE}{SST}$

5 Three Perspectives on Linear Regression

Three Perspectives on Linear Regression (Annotated Outline)

Let $X \in \mathbb{R}^{n\times p}$ be the design matrix, $\mathbf{y} \in \mathbb{R}^n$ the response vector, and $\boldsymbol{\beta} \in \mathbb{R}^p$ the coefficient vector.

—

1. Geometric (Projection) Perspective

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

This gives the coefficients that place the prediction in the span of the columns of X.

$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}}$$

The fitted values are a linear combination of the columns of X.

$$\hat{\mathbf{y}} = X(X^\top X)^{-1}X^\top \mathbf{y}$$

Substituting the coefficient formula shows predictions depend linearly on $\mathbf{y}$.

$$P = X(X^\top X)^{-1}X^\top$$

This matrix maps any vector onto the column space of X.

$$X^\top(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{0}$$

The residual is orthogonal to every predictor direction.

Interpretation: Linear regression projects $\mathbf{y}$ onto the subspace spanned by the predictors.

—

2. Statistical (Probabilistic) Perspective

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

The response is modeled as a linear signal plus random noise.

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

The noise is assumed independent, mean zero, and Gaussian.

$$\ell(\boldsymbol{\beta}) \propto -\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$$

The log-likelihood decreases with squared prediction error.

$$\hat{\boldsymbol{\beta}} = \arg\max \ell(\boldsymbol{\beta})$$

We choose coefficients that make the observed data most likely.

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

Maximizing likelihood yields the least-squares solution.

Interpretation: Least squares is maximum likelihood estimation under Gaussian noise.

—

3. Optimization (Loss Minimization) Perspective

$$L(\boldsymbol{\beta}) = \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$$

Define a loss measuring total squared prediction error.

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta})$$

The goal is to find coefficients that minimize this loss.

$$\nabla L(\boldsymbol{\beta}) = -2X^\top(\mathbf{y} - X\boldsymbol{\beta})$$

Compute the gradient to locate stationary points.

$$X^\top X\hat{\boldsymbol{\beta}} = X^\top \mathbf{y}$$

Setting the gradient to zero gives the normal equations.

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

Solving the normal equations yields the least-squares estimator.

Interpretation: Linear regression minimizes squared error over all linear models.

—

Summary

- **Geometric:** projection of $\mathbf{y}$ onto $\mathrm{Col}(X)$
- **Statistical:** maximum likelihood under a Gaussian noise model
- **Optimization:** minimization of squared prediction error

All three perspectives describe the same estimator $\hat{\boldsymbol{\beta}}$ from different viewpoints.