STAT 587: Data Science I

Winter 2026

Dorcas Ofori-Boateng (PhD)

Linear Regression Models

- Linear regression is a foundational method; Provides *interpretability*, *efficiency*, and *statistical grounding*.
- Serves as a *baseline* for more complex models. Many modern methods **extend** or **regularize** linear models.
- Assume response depends *linearly* on predictors:

$$f(X) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \quad ; \quad \beta = (\beta_0, \beta_1, ..., \beta_p)$$

- Parameters β_j quantify feature influence.
- Choose β to minimize residual sum of squares:

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

Linear Regression Models

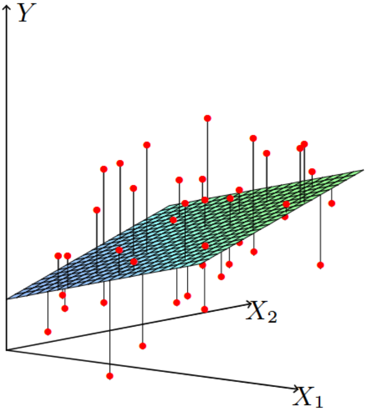

FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

Linear Regression Models

General conventions:

- lower case bold: vector of length n
- lower case normal font: vectors that are not of length n
- upper case bold: matrices
- all vectors are column vectors unless specified otherwise
- x^T denotes transpose of x

Define:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} = [\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p]$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Linear Regression Models

- $Y_i = x_i^T\beta + \epsilon_i = f(x_i) + \epsilon_i$, $f(x) = E(Y|X = x) = x^T\beta$
- $\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$
- rank of $\mathbf{X}$ is full, i.e., $(\mathbf{X}^T\mathbf{X})^{-1}$ exists.
- $\hat{\beta}$ — estimator of β
- $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$ — fitted response vector
- $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$ — residual vector
- Predicted response when $x = x_0$: $\hat{Y}_0 = x_0^T\hat{\beta} = \hat{f}(x_0)$

Linear Regression Models

As before: Minimize $\sum_{i=1}^{n} e_i^2 = \mathbf{e}^T\mathbf{e}$ with respect to β to get $\hat{\beta}$

- Least squares estimator: $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$
- Minimum value of $\sum_{i=1}^{n} e_i^2$ is $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = (\mathbf{Y} - \hat{\mathbf{Y}})^T(\mathbf{Y} - \hat{\mathbf{Y}}) = SS_{\text{ERR}}$ — **error (or residual) sum of squares**

Properties:

- $\hat{\beta}$ is *linear* in $\mathbf{Y}$
- Unbiased, i.e., $E(\hat{\beta}) = \beta$
- $\text{var}(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$
- $\text{var}(\hat{\beta}_0) = \sigma^2 \times$ first diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$
- $\text{var}(\hat{\beta}_j) = \sigma^2 \times (j+1)$th diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$
- $\hat{\sigma}^2 = SS_{\text{ERR}}/(n-p-1)$ — unbiased for σ^2.

Risk: Population vs. Empirical approximation

- Least squares estimators achieve the minimum variance among all linear unbiased estimators.
  - **Gauss–Markov Theorem:** If we have any other linear estimator $\tilde{\theta} = \mathbf{c}^T\mathbf{y}$ that is unbiased for $a^T\beta$, then:

$$\boxed{\mathrm{Var}(a^T\hat{\beta}) \leq \mathrm{Var}(\mathbf{c}^T\mathbf{y})}$$

- Unbiasedness is *not always optimal*, motivating the use of biased methods such as Ridge regression & LASSO.

Shrinkage & Regularization

1. Reduce variance by constraining coefficients
2. Allow small bias to improve prediction

Ridge Regression:

- Adds ℓ_2 penalty:

$$\sum (y_i - x_i^T \beta)^2 + \lambda \sum \beta_j^2$$

- Shrinks coefficients toward zero; Effective under multicollinearity.

LASSO:

- Adds ℓ_1 penalty:

$$\sum (y_i - x_i^T \beta)^2 + \lambda \sum |\beta_j|$$

- Produces sparse solutions; Performs variable selection.

Practicals

Applications in Python

- Simple Linear Regression
- Multiple Linear Regression
- Multivariate Goodness-of-fit
- Interaction terms
- Categorical/Qualitative predictors
- Non-linear transformation of polynomial*

Classification methods

Python coding exercises

- Logistic regression
- Discriminant analysis (LDA, QDA); RDA is given separately.
- KNN
- Naive Bayes