Linear Regression

Andrew Nobel

April, 2020

General Setting: Real-Valued Response

New setting: Jointly distributed pair $(X, Y) \in \mathcal{X} \times \mathbb{R}$

- Feature vector X with values in $\mathcal{X}$ (usually $\mathbb{R}^p$)
- Response Y is real-valued

Ex 1: Marketing (ISL)

- $X =$ money spent on different components of marketing campaign
- $Y =$ gross profits from sales of marketed item

Ex 2: Cost of housing

- $X =$ geographic and demographic features of a neighborhood
- $Y =$ median home price

Predicting the Response from the Features

Basic Components

- Jointly distributed pair $(X, Y) \in \mathcal{X} \times \mathbb{R}$
- Prediction rule is a map $\varphi : \mathcal{X} \to \mathbb{R}$. Idea: $\varphi(X)$ is an estimate of Y
- Squared loss $\ell(y', y) = (y' - y)^2$, error when y' is used to predict y
- Risk of prediction rule φ is expected loss

$$R(\varphi) = \mathbb{E}\ell(\varphi(X), Y) = \mathbb{E}(\varphi(X) - Y)^2$$

Overall goal: Find rule φ to minimize risk $R(\varphi)$

The Regression Function

Fact: Under the squared loss

$$R(\varphi) \;=\; \mathbb{E}[\,\varphi(X) - \mathbb{E}(Y|X)\,]^2 \;+\; \mathbb{E}[\,\mathbb{E}(Y|X) - Y\,]^2$$

Thus optimal prediction rule φ is the *regression function*

$$f(x) = \mathbb{E}(Y|X = x)$$

Signal plus noise model

$$Y = f(X) + \varepsilon \;\text{ where } \varepsilon \perp\!\!\!\perp X, \mathbb{E}\varepsilon = 0, \mathrm{Var}(\varepsilon) = \sigma^2$$

In this case f is the regression function and

$$R(\varphi) \;=\; \mathbb{E}(\varphi(X) - f(X))^2 \;+\; \mathrm{Var}(\varepsilon)$$

Observations, Procedures, and Empirical Risk

Observations: $D_n = (X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathbb{R}$ iid copies of (X, Y)

Definition

- A *regression procedure* is a map $\varphi_n : \mathcal{X} \times (\mathcal{X} \times \mathbb{R})^n \to \mathbb{R}$
- $\hat{\varphi}_n(x) := \varphi_n(x : D_n)$, prediction rule based on observations D_n

Definition: The *empirical risk* or *training error* of a rule φ is given by

$$\hat{R}_n(\varphi) \;=\; \frac{1}{n} \sum_{i=1}^{n} (Y_i - \varphi(X_i))^2$$

Linear Regression

Two sides of the coin

- Linear models: How data is generated
- Linear prediction rules: How data is fit

Linear Regression Model

Model: For some coefficient vector $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^t \in \mathbb{R}^{p+1}$

$$Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon$$

where we assume that

- ε is independent of feature vector X
- $\mathbb{E}\varepsilon = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$

Note: *No assumption* about distribution of feature vector X

Convention: $X = (1, X_1, \ldots, X_p)^t$, so linear model can be written

$$Y = X^t \beta + \varepsilon$$

Flexibility of Linear Model (from ESL)

Flexibility arises from latitude in defining the features $X = (1, X_1, \ldots, X_p)^t$

Features can include

- Any numerical quantity (possibly taking a finite number of values)
- Transformations (square root, log, square) of numerical quantities
- Polynomial ($X_2 = X_1^2$, $X_3 = X_1^3$) or basis expansions of other features
- Dummy variables to code qualitative inputs
- Variable interactions: $X_3 = X_1 \cdot X_2$ or perhaps $X_3 = \mathbb{I}(X_1 \geq 0, X_2 \geq 0)$

Linear Rules and Procedures

Definition: Let $\mathcal{X} = \mathbb{R}^{p+1}$

- *Linear prediction rule* has form $\varphi_\beta(x) = x^t\beta$ for some $\beta \in \mathbb{R}^{p+1}$
- *Linear procedure* φ_n produces linear rules from observations D_n

Notation: Linear rule φ_β fully determined by coefficient vector β

- $R(\beta) \ = \ \mathbb{E}(Y - X^t\beta)^2$
- $\hat{R}_n(\beta) \ = \ n^{-1}\sum_{i=1}^n (Y_i - X_i^t\beta)^2$

Underlying Distributions: Assumptions and Non-Assumptions

Fitting: Fitting linear models

- Data $(x_1, y_1), \ldots, (x_n, y_n)$ fixed (non-random)
- No assumption about underlying distribution

Inference: For coefficients from OLS, Ridge, LASSO

- $y_i = x_i^t \beta + \varepsilon_i$ with x_j fixed and ε_j iid $\sim \mathcal{N}(0, \sigma^2)$
- Conditions on the feature vectors x_j (design matrix)

Assessment: Test error, cross-validation

- Data from iid observations $(X_i, Y_i) \sim (X, Y)$

Ordinary Least Squares (OLS)

Given: Paired observations $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^{p+1} \times \mathbb{R}$ define

- Response vector $y = (y_1, \ldots, y_n)^t$
- Design matrix X with ith row x_i^t

OLS: Identify the vector $\hat{\beta}$ minimizing the residual sum of squares (RSS)

$$n \, \hat{R}_n(\beta) \; = \; ||y - X\beta||^2$$

Motivation: If observations (x_i, y_i) generated from linear model with normal errors, $\hat{\beta}$ is the MLE of true coefficient vector

Least Squares Estimation of Coefficient Vector

Fact: If rank$(X) = p$ then $\hat{R}_n(\beta)$ is strictly convex and has unique minimizer

$$\hat{\beta} = (X^t X)^{-1} X^t y \qquad \text{(normal eqns)}$$

- Minimization problem has closed form solution
- Assumption rank$(X) = p$ ensures $X^t X$ is invertible, requires $n \geq p$
- OLS procedure yields linear rule $\varphi_{\hat{\beta}}(x) = x^t \hat{\beta}$
- Predicted values of response y given by $\hat{y} = X\hat{\beta}$

Inference for OLS Coefficients in Gaussian Setting

Fact: Suppose $y_i = x_i^t \beta + \varepsilon_i$ with x_j fixed and ε_j iid $\sim \mathcal{N}(0, \sigma^2)$

1. $y = X\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$
2. $\hat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2 (X^t X)^{-1})$
3. $||y - X\hat{\beta}||^2 \sim \sigma^2 \chi^2_{n-p-1}$ so one may estimate σ^2 by
$$\hat{\sigma}^2 := ||y - X\hat{\beta}||^2/(n-p-1)$$
4. If $\beta_j = 0$ then $T_j = \hat{\beta}_j / \hat{\sigma}\sqrt{(X^t X)^{-1}_{jj}} \sim t_{n-p-1}$. Use T_j to test if $\beta_j = 0$
5. Approx. 95% confidence interval for β_j is $(\hat{\beta}_j - 1.96\hat{\sigma}, \hat{\beta}_j + 1.96\hat{\sigma})$

Penalized Linear Regression

OLS estimate $\hat{\beta}$ depends on $(X^t X)^{-1}$

- Inverse does not exist if $p > n$
- Small eigenvalues resulting from (near) collinearity among features can lead to unstable estimates, unreliable predictions

Alternative: Penalized regression

- Regularize OLS cost function by adding a term that penalizes large coefficients, shrinking estimates towards zero

Ridge Regression

Given: Paired observations $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^{p+1} \times \mathbb{R}$

- Response vector y, design matrix X

Penalized cost function: For each $\lambda \geq 0$ define

$$\hat{R}_{n,\lambda}(\beta) \;=\; ||y - X\beta||^2 \,+\, \lambda\, ||\beta||^2$$

- $||y - X\beta||^2$ measures fit of the linear model
- $||\beta||^2$ measures magnitude of coefficient vector
- λ controls tradeoff between fit and magnitude

Ridge Regression, cont.

Fact: If $\lambda > 0$ then $\hat{R}_{n,\lambda}(\beta)$ is strictly convex and has unique minimizer

$$\hat{\beta}_\lambda = (X^t X + \lambda I_p)^{-1} X^t y$$

- Eigenvalues of $X^t X + \lambda I_p =$ eigenvalues of $X^t X$ plus λ.
- If $\lambda > 0$ then $X^t X + \lambda I_p > 0$ is invertible so $\hat{\beta}_\lambda$ is well defined
- If $\lambda_1 \leq \lambda_2$ then $||\hat{\beta}_{\lambda_2}|| \leq ||\hat{\beta}_{\lambda_1}||$. Penalty shrinks $\hat{\beta}_\lambda$ towards zero
- Ridge procedure yields linear rule $\varphi_{\hat{\beta}_\lambda}(x) = x^t \hat{\beta}_\lambda$
- Ridge regression is really a *family* of procedures, one for each λ

Ridge Regression as a Convex Program

Recall: $\hat{R}_{n,\lambda}(\beta) \;=\; ||y - X\beta||^2 \,+\, \lambda\,||\beta||^2$

Fact: Minimizing $\hat{R}_{n,\lambda}(\beta)$ is the Lagrangian form of the mathematical program

$$\min f(\beta) = ||y - X\beta||^2 \text{ subject to } \; ||\beta||^2 \leq t,$$

where t depends on λ

Note: Objective function and constraint set of the program are convex.

Selecting Penalty Parameter

Issue: Different parameters λ give different solutions $\hat{\beta}_\lambda$. How to choose λ?

- Fix "grid" $\Lambda = \{\lambda_1, \ldots, \lambda_N\}$ of parameter values

1. Independent training set D_n and test set D_m
   - Find vectors $\hat{\beta}_{\lambda_1}, \ldots, \hat{\beta}_{\lambda_N}$ using training set D_n
   - Select vector $\hat{\beta}_{\lambda_\ell}$ minimizing test error $\hat{R}_m(\beta)$
2. Cross-validation
   - For each $1 \leq \ell \leq N$ evaluate cross-validated risk $\hat{R}^{\text{k-CV}}(\text{Ridge}(\lambda_\ell))$
   - Select vector $\hat{\beta}_{\lambda_\ell}$ for which λ_ℓ minimizes cross-validated risk

Ridge Regression and Gaussian Linear Model

Setting: Suppose $y_i = x_i^t \beta + \varepsilon_i$ with x_j fixed and ε_j iid $\sim \mathcal{N}(0, \sigma^2)$

Ridge estimator $\hat{\beta}_\lambda$ shrinks OLS estimator/MLE $\hat{\beta}$ towards zero. For $\lambda > 0$

- Increased bias $\mathbb{E}\hat{\beta}_\lambda \neq \beta$
- Reduced variance $\mathrm{Var}(\hat{\beta}_\lambda) < \mathrm{Var}(\hat{\beta})$

Appropriate λ can reduce mean-squared error $\mathbb{E}||\hat{\beta}_\lambda - \beta||^2 < \mathbb{E}||\hat{\beta} - \beta||^2$