Linear Regression

Andrew Nobel

October, 2021

Regression: Prediction with a Real-Valued Response

Setting: Jointly distributed pair (X, Y) with $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$

- X is a feature vector, often high dimensional
- Y is a real-valued response

Goals

- Predict Y from X
- Identify the components of X that most affect Y

Regression: Prediction with a Real-Valued Response

Ex 1: Marketing (ISL)

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

Ex 2: Housing

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

Regression: Statistical Framework

1. Jointly distributed pair $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$

2. Prediction rule $\varphi : \mathbb{R}^p \to \mathbb{R}$. Regard $\varphi(X)$ as an estimate of Y

3. Squared loss $\ell(y', y) = (y' - y)^2 =$ error when y' used to predict y

4. Risk of prediction rule φ is its expected loss

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

Overall goal: Find a prediction rule φ with small risk $R(\varphi)$

Optimal Prediction and the Regression Function

Fact: Under the squared loss the risk of any fixed rule φ is

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

Thus optimal prediction rule φ is the *regression function*

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

Signal Plus Noise Model: Assume for some function $f : \mathbb{R}^p \to \mathbb{R}$

$$Y = f(X) + \varepsilon \;\text{ where }\; \mathbb{E}\,\varepsilon = 0 \;\text{ and }\; \varepsilon \perp\!\!\!\perp X$$

In this case f is the regression function, and for every prediction rule φ

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

Regression Procedures and Empirical Risk

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

Definition

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

Encompasses assumptions about data generation and prediction

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

Linear Regression Model

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

$$Y \;=\; \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon \;=\; \langle \beta, X \rangle + \varepsilon$$

where we assume that

- ε is independent of augmented feature vector $X = (1, X_1, \ldots, X_p)^t$
- $\mathbb{E}\varepsilon = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$

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

Flexibility of Linear Model (from ESL)

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

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, e.g., $X_3 = X_1 \cdot X_2$ or $X_3 = \mathbb{I}(X_1 \geq 0, X_2 \geq 0)$

Linear Rules and Procedures

Definition

- *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. Write

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

Different Settings, Different Assumptions

Fitting: Fitting linear models

- Data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ is fixed, non-random
- No assumption about underlying distribution(s)

Inference: Concerning coefficients from OLS, Ridge, LASSO

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

Assessment: Test error, cross-validation

- Observations (X_i, Y_i) are iid copies of (X, Y)

Ordinary Least Squares (OLS)

Given: Paired observations $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \in \mathbb{R}^{p+1} \times \mathbb{R}$ define

- Response vector $\mathbf{y} = (y_1, \ldots, y_n)^t$
- Design matrix $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$ with ith row $\mathbf{x}_i^t$

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

$$n\,\hat{R}_n(\beta) \;=\; \sum_{i=1}^{n} (y_i - \mathbf{x}_i^t \beta)^2 \;=\; ||\mathbf{y} - \mathbf{X}\beta||^2$$

Interpretation: Projecting $\mathbf{y}$ onto subspace of $\mathbb{R}^n$ spanned by columns of $\mathbf{X}$, which correspond to features of the data

Least Squares Estimation of Coefficient Vector

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

$$\hat{\beta} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{y} \quad \text{(normal equations)}$$

- Minimization problem has closed form solution
- Assumption $\text{rank}(\mathbf{X}) = p$ ensures $\mathbf{X}^t \mathbf{X}$ is invertible, requires $n \geq p$
- Solution $\hat{\beta}$ yields linear prediction rule $\varphi_{\hat{\beta}}(x) = \langle \hat{\beta}, x \rangle$
- Fitted value of the response $\mathbf{y}$ is the projection $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$

Gaussian Linear Model

Gaussian Linear Model: Assume feature vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are fixed and that responses y_i follow linear model with normal errors

$$y_i = \mathbf{x}_i^t \beta + \varepsilon_i \text{ with } \varepsilon_i \text{ iid } \mathcal{N}(0, \sigma^2)$$

Model can be written in vector form $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$

Fact: Estimate $\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$ has following properties

1. $\mathbb{E}\hat{\beta} = \beta$ and $\mathrm{Var}(\hat{\beta}) = (\mathbf{X}^t\mathbf{X})^{-1}\sigma^2$
2. $\hat{\beta}$ is multivariate normal

Inference for Gaussian Linear Model

1. Can show $||\mathbf{y} - \mathbf{X}\hat{\beta}||^2 \sim \sigma^2 \chi^2_{n-p-1}$. Estimate noise variance σ^2 by

$$\hat{\sigma}^2 \; = \; \frac{||\mathbf{y} - \mathbf{X}\hat{\beta}||^2}{n - p - 1}$$

2. Let $v_j = (\mathbf{X}^t \mathbf{X})^{-1}_{jj}$. If $\beta_j = 0$ then the t-type statistic

$$T_j \; = \; \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}} \; \sim \; t_{n-p-1}$$

We can use T_j to test if $\beta_j = 0$. Approximate 95% confidence interval for β_j is

$$(\hat{\beta}_j - 1.96\sqrt{v_j}\,\hat{\sigma}, \hat{\beta}_j + 1.96\sqrt{v_j}\,\hat{\sigma})$$

Penalized Linear Regression

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

- 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

Setting: Paired observations $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$

- Centering: Assume $\sum_{i=1}^n \mathbf{x}_i = 0$ and $\sum_{i=1}^n y_i = 0$
- Design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$
- Response vector $\mathbf{y} \in \mathbb{R}^n$

Ridge Regression, cont

Given: Design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ and response vector $\mathbf{y} \in \mathbb{R}^n$

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

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

- $||\mathbf{y} - \mathbf{X}\beta||^2$ measures fit of the linear model
- $||\beta||^2$ measures magnitude of coefficient vector
- λ controls tradeoff between fit and magnitude
- OLS is special case $\lambda = 0$

Ridge Regression, cont.

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

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

- Eigenvalues of $\mathbf{X}^t\mathbf{X} + \lambda I_p =$ eigenvalues of $\mathbf{X}^t\mathbf{X}$ plus λ.
- If $\lambda > 0$ then $\mathbf{X}^t\mathbf{X} + \lambda I_p > 0$ 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 $\varphi_{\hat{\beta}_\lambda}(\mathbf{x}) = \langle \mathbf{x}, \hat{\beta}_\lambda \rangle$
- Ridge regression is really a *family* of procedures, one for each λ

Ridge Regression as a Convex Program

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

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

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

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

Approach 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 with different λ
- Select vector $\hat{\beta}_{\lambda_\ell}$ minimizing test error $\hat{R}_m(\beta) = m^{-1} \sum_{j=1}^m (Y_j - X_j^t \beta)^2$

Approach 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 $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ with $\mathbf{X}$ fixed and $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$

Ridge estimate $\hat{\beta}_\lambda$ shrinks OLS estimate $\hat{\beta}$ 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 choice of λ can reduce overall mean-squared error, that is,

$$\mathbb{E}||\hat{\beta}_\lambda - \beta||^2 < \mathbb{E}||\hat{\beta} - \beta||^2$$