SDS 387
Linear Models

Fall 2024

Lecture 16 - Thu, Oct 24, 2024

Instructor: Prof. Ale Rinaldo

- Announcement: I will post HW 4 soon (and also post solutions to HW 3)

- Last time: Ordinary Least Squares (OLS) estimator

- Setting:

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \in \mathbb{R}^n$$ vector of response variables

Φ design matrix whose i^{th} row contains the i^{th} observation for the covariates (a point in $\mathbb{R}^d$)

$n \times d$

Notation used in Bach's book

- The notation Φ is non-standard in statistics but reflect the common practice in ML of turning a vector of covariates, say x_i, into a vector of features $\Phi_i \in \mathbb{R}^d$, $\Phi_i = \varphi(x_i)$.

- I will assume throughout that the first column of $\underset{n \times d}{\Phi}$ is a vector of 1's.

  This means that we always fit an intercept to our linear model

- The OLS estimator is obtained as the minimizer of the empirical risk:

$$\beta \to \hat{R}(\beta) = \frac{1}{2n} \| Y - \Phi \beta \|_2^2$$

Assuming that Φ has full-column rank ($\text{rank}(\Phi) = d$) then the solution exists and is unique and given by

$$\begin{aligned} \hat{\beta} &= (\Phi^T \Phi)^{-1} \Phi^T Y \\ &= \hat{\Sigma}^{-1} \frac{\Phi Y}{n} \quad \text{where } \hat{\Sigma} = \frac{\Phi^T \Phi}{n} \end{aligned}$$

Importantly $\hat{\beta}$ satisfies the normal equations

$$\Phi^T \Phi \hat{\beta} = \Phi^T Y$$

Geometric interpretation:

From the formula for $\hat{\beta}$ the vector of fitted values

$$\hat{y} = \Phi\hat{\beta} = \underbrace{\Phi(\Phi^T\Phi)^{-1}\Phi^T}_{H} y$$

is a linear function of y

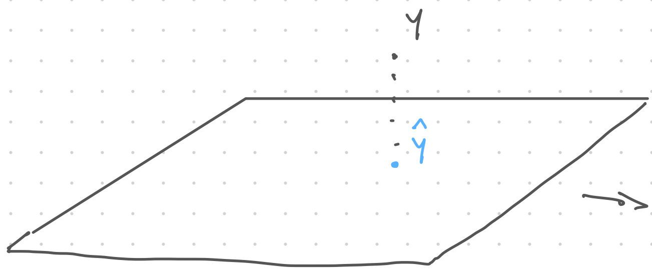
H → = "hat matrix"

= "prediction of the responses given by the model or the estimated value of the regression function"

Note: $\underset{n \times n}{H}$ is a orthogonal projection matrix

$$H^2 = H \quad \text{and} \quad H^T = H$$

H projects y onto the d-dimensional linear subspace of $\mathbb{R}^n$ spanned by the columns of Φ



→ $C(\Phi)$ column space of Φ

$\hat{y}$ is the point in $C(\Phi)$ that is closest to y in $\mathbb{R}^n$

From this we can see that the residuals

$$e = y - \hat{y} = (I - H)\, y \in \mathbb{R}^n$$

are the orthogonal projection of y

onto the orthogonal complement of $C(\Phi)$

<u>Remark</u> $I-H$ is also an orthogonal projection and

$$\langle e, \hat{Y} \rangle = \langle (I-H)Y, HY \rangle = 0$$

$\hookrightarrow$ $$Y = \underbrace{\hat{Y} + e}_{\text{orthogonal}}$$

$\hookrightarrow$ $$\|Y\|^2 = \|\hat{Y}\|^2 + \|e\|^2$$

$\|\hat{Y}\|^2$ = "energy" or variability explained by model

Numerical considerations

How hard is to compute $\hat{\beta}$ numerically?

To compute $\hat{\beta}$ one has to invert $\underbrace{\Phi^T\Phi}_{d \times d}$

typically requires order $O(d^3)$ computations.

Gradient descent (Bach 5.2.1)

Starting from an initial point $\beta_0 \in \mathbb{R}^d$ consider the sequence of updates

$$\beta_t = \beta_{t-1} - \underbrace{\gamma}_{>0 \text{ stepsize}} \nabla \hat{R}(\beta_{t-1})$$

where recall that

$$\hat{R}(\beta) = \frac{1}{2n} \|Y - \Phi\beta\|^2 \quad \text{and}$$

$$\nabla\hat{R}(\beta) = \frac{1}{n}\Phi^T\Phi\beta - \frac{\Phi^T Y}{n}$$

$$\text{He}\,\hat{R}(\beta) = \frac{1}{n}\Phi^T\Phi = \hat{\Sigma}$$

Hessian

Also recall that any minimizer, say $\tilde{\beta}$, of $\hat{R}$ satisfies

$$\hat{\Sigma}\tilde{\beta} = \frac{\Phi^T Y}{n}$$

A solution exists always and is unique if $\Phi^T\Phi$ is invertible. If not, there exist infinitely many solutions

First off, notice that, if solution β^* exists, then

$$\hat{R}(\beta) - \hat{R}(\beta^*) = \frac{1}{2n}\|Y - \Phi\beta\|^2 - \frac{1}{2n}\|Y - \Phi\beta^*\|^2$$

$$= \frac{1}{2n}\|Y - \Phi\beta^* + \Phi(\beta^* - \beta)\|^2 - \frac{1}{2n}\|Y - \Phi\beta^*\|^2$$

$$= \cancel{\frac{1}{2n}\|Y - \Phi\beta^*\|^2} + \frac{1}{2n}\|\Phi(\beta^* - \beta)\|^2 + \frac{\cancel{2}}{\cancel{2}n}\underbrace{\langle Y - \Phi\beta^*, \Phi(\beta^* - \beta)\rangle}_{\substack{= 0 \text{ because} \\ Y - \Phi\beta^* \in C(\Phi)^{\perp}}}$$

$$- \cancel{\frac{1}{2n}\|Y - \Phi\beta^*\|^2}$$

$$= \frac{1}{2n} \| \Phi (\beta^* - \beta) \|^2 = \frac{(\beta - \beta^*)^T \hat{\Sigma} (\beta - \beta^*)}{2}$$

$$= \frac{\| \beta - \beta^* \|^2_{\hat{\Sigma}}}{2}$$

Next, let's look at the gradient iterates:

$$\beta_t = \beta_{t-1} - \gamma \nabla \hat{R}(\beta_{t-1}) = \beta_{t-1} - \gamma \left[\frac{1}{n} \Phi^T (\Phi \beta_{t-1} - Y) \right]$$

$$= \beta_{t-1} - \gamma \hat{\Sigma} (\beta_{t-1} - \beta^*)$$

because $\hat{\Sigma} \beta^* = \frac{\Phi^T Y}{n}$

$\downarrow$

$$\beta_t - \beta^* = (I - \gamma \hat{\Sigma}) (\beta_{t-1} - \beta^*)$$

This implies that

$$\boxed{\| \beta_t - \beta^* \|^2 = (\beta_0 - \beta^*)^T (I - \gamma \hat{\Sigma})^{2t} (\beta_0 - \beta^*)}$$

and

$$\boxed{\hat{R}(\beta_t) - \hat{R}(\beta^*)} = \frac{(\beta_t - \beta^*)^T \hat{\Sigma} (\beta_t - \beta^*)}{2}$$

$$= \frac{1}{2} (\beta_0 - \beta^*)^T (I - \gamma \hat{\Sigma} \underbrace{\hat{\Sigma} (I - \gamma \hat{\Sigma})^t}_{\text{commute}} (\beta_0 - \beta^*)$$

HW!

$$\boxed{= \frac{1}{2} (\beta_0 - \beta^*)^T (I - \gamma \hat{\Sigma})^{2t} \hat{\Sigma} (\beta_0 - \beta^*)}$$

Let's first look at convergence to the minimizer. Assume that $\hat{\Sigma}$ is invertible. The eigenvalues of $(I - \gamma \hat{\Sigma})^{2t}$ are of the form $(1 - \gamma \lambda)^{2t}$ where λ is an eigenvalue of $\hat{\Sigma}$. So all the eigenvalues of $(I - \gamma \hat{\Sigma})^{2t}$ are less than

$$\max_{\lambda_{min}(\hat{\Sigma}) \leq \lambda \leq \lambda_{max}(\hat{\Sigma})} |1 - \gamma \lambda|^{2t}$$

Now let's choose γ to be $\frac{1}{\lambda_{max}(\hat{\Sigma})}$. Then the expression above is equal to

$$\left(1 - \frac{\lambda_{min}(\hat{\Sigma})}{\lambda_{max}(\hat{\Sigma})}\right)^{2t} = \left(1 - \frac{1}{\kappa}\right)^{2t} \qquad \kappa = \frac{\lambda_{max}}{\lambda_{min}}$$

$\downarrow$ condition number of $\hat{\Sigma}$

Putting everything together:

$$\|\beta_t - \beta^*\|^2 \leq \left(1 - \frac{1}{\kappa}\right)^{2t} \|\beta_0 - \beta^*\|^2$$

$$\leq e^{-2t/\kappa} \|\beta_0 - \beta^*\|^2$$

exponential / geometric / linear convergence

If $\kappa = \infty$ (ie $\lambda_{min}(\hat{\Sigma}) = 0$) this will simply say that $\|\beta_t - \beta^*\|^2 \leq \|\beta_0 - \beta^*\|^2$ all t.

- Let's look at the convergence of the objective function $\hat{R}$

$$\hat{R}(\beta_t) - \hat{R}(\beta^*) = \frac{1}{2} (\beta_0 - \beta^*)^T (I - \gamma \hat{\Sigma})^{2t} \hat{\Sigma} (\beta_0 - \beta^*)$$

$x^T A x = \mathrm{tr}(A x x^T)$

$$= \frac{1}{2} \mathrm{tr}\left((I - \gamma \hat{\Sigma})^{2t} \hat{\Sigma} (\beta_0 - \beta^*)(\beta_0 - \beta^*)^T \right)$$

$$\leq \frac{1}{2} \|(I - \gamma \hat{\Sigma})^{2t}\|_{op} \, \mathrm{tr}\left(\hat{\Sigma} (\beta_0 - \beta^*)(\beta_0 - \beta^*)^T \right)$$

because

$\mathrm{tr}(AB) \leq \|A\|_{op} \mathrm{tr}(B)$

psd

$$= \frac{1}{2} \underbrace{\max_{\substack{\lambda \\ \text{eig. of } \hat{\Sigma}}} |1 - \gamma \lambda|^{2t}}_{\text{if } \gamma = \frac{1}{\lambda_{max}(\hat{\Sigma})}} \underbrace{\left[(\beta_0 - \beta^*)^T \hat{\Sigma} (\beta_0 - \beta^*) \right]}_{\hat{R}(\beta_0) - \hat{R}(\beta^*)}$$

$$\leq \exp\left\{ - \frac{2t}{\kappa} \right\}$$

$$\leq \frac{e^{-2t/\kappa}}{2} \left[\hat{R}(\beta_0) - \hat{R}(\beta^*) \right]$$

- What if $\kappa = \infty$ (i.e. $\hat{\Sigma}$ is not invertible)?

$$\hat{R}(\beta_t) - \hat{R}(\beta^*) \leq \frac{1}{2} \| (I - \gamma \hat{\Sigma})^{2t} \hat{\Sigma} \|_{op} \|\beta_0 - \beta^*\|^2$$

setting $\gamma = \frac{1}{\lambda_{max}(\hat{\Sigma})}$

$$\leq \max_{\lambda} \lambda (1 - \gamma \lambda)^{2t} \|\beta_0 - \beta^*\|^2$$

$\vdots$

$$\leq \frac{\lambda_{max}}{8t} \|\beta_0 - \beta^*\|^2$$

$\downarrow$

convergence is polynomial in t