36-755: Advanced Statistical Theory **Fall 2016**

Lecture 8: September 26

Lecturer: Alessandro Rinaldo *Scribes: Alden Green*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

8.1 Linear Regression

We assume $Y = X\beta^* + \epsilon$, where X is a fixed nxd design matrix and $\epsilon_1, ..., \epsilon_n \overset{ind}{\sim} SG(\sigma^2)$. Let $\hat{\beta} = f(Y)$. The following two tasks are of interest:

- *Mean Estimation.* Let $\tilde{Y}$ be an independent draw with the same distribution as Y. Then, we seek to minimize the mean squared predictive error, which is defined as

$$\frac{1}{n}\mathbb{E}\left[||\tilde{Y} - X\hat{\beta}||^2\right] \tag{8.1}$$

  Alternatively, we could seek to minimize the mean square error,

$$\frac{1}{n}\mathbb{E}\left[||X\left(\beta^* - \hat{\beta}\right)||^2\right] \tag{8.2}$$

- *Parameter Estimation.* Here, we seek to minimize the expected ℓ_2 norm between the vector of estimated parameters and true parameters,

$$\frac{1}{n}\mathbb{E}\left[||\left(\beta^* - \hat{\beta}\right)||^2\right] \tag{8.3}$$

8.1.1 Least Squares Estimator

To define the least square estimator $\hat{\beta}^{LS}$, we first need a generalized notion of matrix inverses known as the matrix psuedoinverse.

Definition 8.1 (Pseudoinverse of a matrix) *Let A be an nxm matrix. Then, A^+ is a **psuedoinverse** of A if it satisfies*

$$AA^+A = A, (AA^+)^T = AA^+ \tag{8.4}$$

$$A^+AA^+ = A^+, (A^+A)^T = A^+A \tag{8.5}$$

Note that if A is square and invertible, A^{-1} is a pseudoinverse of A. Also, note that in general the pseudoinverse is not unique.

Now, take the objective function $\frac{1}{n}||Y - X\beta||^2$, and minimize it. Setting the gradient to zero, we have

$$\nabla_B \left(||Y_X\beta||^2\right) = 0 \rightarrow \tag{8.6}$$
$$X^T X\beta = X^T Y \tag{8.7}$$

and by the convexity of the objective function, any beta which satisfies the above condition will achieve the minimum.

Definition 8.2 (Least Squares Estimator) *The* ***least squares estimator*** $\hat{\beta}^{LS}$ *is defined in general to be*

$$\hat{\beta}^{LS} := (X^T X)^+ X^T Y \tag{8.8}$$

for some psuedoinverse $(X^T X)^+$. Note that if $d < n$ and $X^T X$ is invertible, we recover $\hat{\beta}^{LS} := (X^T X)^{-1} X^T Y$. Also, note that in general, if $\hat{\beta}^{LS}$ is a least squares estimator $\delta \in$ Kernel(X) then $\hat{\beta}^{LS} + \delta$ is also a least squares estimator.

The least squares estimator turns out to have good mean estimation properties.

Theorem 8.3 (Mean Estimation using Least Squares Estimator) *Assume $(\epsilon_1, ..., \epsilon_n) \in SG_n(\sigma^2)$. Let $r = dim(column\ space(X))$ and $\hat{\beta} = \hat{\beta}^{LS}$ Then, $\exists C > 0$ such that*

$$\frac{1}{n}\mathbb{E}\left[||X\left(\beta^* - \hat{\beta}\right)||^2\right] \leq C\frac{\sigma^2 r}{n}, \text{ and} \tag{8.9}$$
$$\mathbb{P}\left(\frac{1}{n}||X\left(\beta^* - \hat{\beta}\right)||^2 \leq C\frac{\sigma^2 r + \log(\frac{1}{\delta})}{n})\right) \geq 1 - \delta \tag{8.10}$$

Proof: By the optimality of $\hat{\beta}$,

$$||Y - X\hat{\beta}||^2 \leq ||Y - X\beta^*||^2 = ||\epsilon||^2 \tag{8.11}$$

Also, we have that,

$$||\left(Y - X\hat{\beta}\right)||^2 = ||X\left(\hat{\beta} - \beta^*\right)||^2 + ||\epsilon||^2 - 2\left\langle \epsilon, X\left(\hat{\beta} - \beta^*\right)\right\rangle \tag{8.12}$$

Putting these two together yields

$$||X\left(\hat{\beta}-\beta^*\right)||^2 \leq 2\left\langle \epsilon, X\left(\hat{\beta}-\beta^*\right)\right\rangle \to \tag{8.13}$$

$$||X\left(\hat{\beta}-\beta^*\right)|| \leq 2\left\langle \epsilon, \frac{X\left(\hat{\beta}-\beta^*\right)}{||X\left(\hat{\beta}-\beta^*\right)||}\right\rangle \tag{8.14}$$

where the second line comes from dividing both sides by $||X\left(\hat{\beta}-\beta^*\right)||$. To bound the RHS, we note that since $r = \dim(\text{column space}(X))$, there exists some projection matrix Φ into $\mathbb{R}^r$ and a unit vector $v \in \mathbb{S}^{r-1}$

$$\frac{X\left(\hat{\beta}-\beta^*\right)}{||X\left(\hat{\beta}-\beta^*\right)||} = \Phi v, \to \tag{8.15}$$

$$\left\langle \epsilon, \frac{X\left(\hat{\beta}-\beta^*\right)}{||X\left(\hat{\beta}-\beta^*\right)||}\right\rangle = \langle \tilde{\epsilon}, v\rangle \tag{8.16}$$

where $\tilde{\epsilon} = \epsilon^T \Phi$

We therefore have that

$$||X\left(\hat{\beta}-\beta^*\right)||^2 \leq 4 \max_{v \in \mathbb{S}^{r-1}} \left(\tilde{\epsilon}^T v\right)^2 \tag{8.17}$$

Since Φ is a projection matrix (i.e. it has orthonormal columns), we have that $\tilde{\epsilon} \in SG_r(\sigma^2)$. Therefore, by Cauchy-Schwarz, we have that

$$\leq 4 \max_{v \in \mathbb{S}^{r-1}} \left(\tilde{\epsilon}^T v\right)^2 \leq 4 \sum_{j=1}^{r} \mathbb{E}\left[\tilde{\epsilon_j}\right] \leq 16\sigma^2 r \tag{8.18}$$

To show the bound in probability, we use our standard discretization argument. Let $\mathcal{N}_{1/2}$ be a minimal 1/2-covering of S^{r-1}.

$$\max_{v \in \mathbb{S}^{r-1}} \left(\tilde{\epsilon}^T v\right) \leq 2 \max_{z \in \mathcal{N}_{1/2}} \left(\tilde{\epsilon}^T z\right) \to \tag{8.19}$$

$$\mathbb{P}\left(\max_{z \in \mathbb{V}_{1/2}} \left(\tilde{\epsilon}^T z\right)^2 \geq t\right) \leq |\mathcal{N}_{1/2}| \exp\left(\frac{-t}{8\sigma^2}\right) \tag{8.20}$$

$$\leq 6^r \exp\left(\frac{-t}{8\sigma^2}\right) \tag{8.21}$$

Setting the above equal to d and solving for t yields the desired result. ■

Also, note that

$$||\hat{\beta} - \beta^\star||^2 \lambda_{min}^2(X) \leq ||X\left(\hat{\beta} - \beta^*\right)||^2 \tag{8.22}$$

which gives us a meaningful (though not necessarily optimal) bound on $||\hat{\beta} - \beta^\star||^2$ if $\lambda_{min}^2(X) > 0$. This does not help us, of course, when $d > n$ as in that case $\lambda_{min}^2(X) = 0$ always holds.

8.2 Penalized Regression and Lasso

Assume the same model for Y. Now, instead of the least squares estimator, consider the penalized regression estimator.

Definition 8.4 (Penalized Least Squares Estimator) *Let $\lambda_n > 0$, and choose a penalty function $f(\beta) \geq 0$. Then, the corresponding* ***penalized least squares*** *estimator $\hat{\beta}^{PLS}$ satisfies*

$$\hat{\beta}^{PLS} \in \underset{\beta}{argmin}\left\{\frac{1}{2n}||Y - X\beta||^2 + \lambda_n f(\beta)\right\} \tag{8.23}$$

The LASSO estimator $\hat{\beta}^{LASSO}$ is the penalized least squares estimator with the ℓ_1 norm as penalty function, $f\beta = ||\beta||_1$. There are several equivalent formulations of the LASSO problem.

Proposition 8.5 (Equivalent Statements of LASSO) *The following three statements lead to equivalent solution paths, over λ_n, B and R respectively:*

$$\underset{\beta}{argmin}\frac{1}{2n}||Y - X\beta||^2 + \lambda_n||\beta||_1 \tag{8.24}$$

$$\underset{\beta}{argmin}||\beta||_1 s.t. \frac{1}{2n}||Y - X\beta||^2 \leq B^2 \tag{8.25}$$

$$\underset{\beta}{argmin}\frac{1}{2n}||Y - X\beta||^2 s.t. ||\beta||_1 \leq R \tag{8.26}$$

The LASSO also has good mean estimation properties. The following theorem is proved in next class.

Theorem 8.6 (Mean Estimation using LASSO) *If $\lambda_n \geq ||\frac{X^T\epsilon}{n}||_\infty$, then any LASSO solution satisfies*

$$\frac{||X\left(\hat{\beta} - \beta^*\right)||^2}{n} \leq 4||\beta^*||_1\lambda_n \tag{8.27}$$