8.1 Linear Regression

We assume \( Y = X\beta^* + \epsilon \), where \( X \) is a fixed \( n \times d \) design matrix and \( \epsilon_1, \ldots, \epsilon_n \overset{\text{iid}}{\sim} SG(\sigma^2) \). 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

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

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

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

- **Parameter Estimation.** Here, we seek to minimize the expected \( \ell_2 \) norm between the vector of estimated parameters and true parameters,

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

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 pseudoinverse.

**Definition 8.1 (Pseudoinverse of a matrix)** Let \( A \) be an \( n \times m \) matrix. Then, \( A^+ \) is a pseudoinverse of \( A \) if it satisfies

\[
AA^+A = A, (AA^+)^T = AA^+ \quad (8.4)
\]
\[
A^+AA^+ = A^+, (A^+A)^T = A^+A \quad (8.5)
\]
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

$$\nabla_B \left( ||Y\beta||^2 \right) = 0 \rightarrow X^TX\beta = X^TY \quad (8.6)$$

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

$$\hat{\beta}^{LS} := (X^TX)^+X^TY \quad (8.8)$$

for some pseudoinverse $(X^TX)^+$. Note that if $d < n$ and $X^TX$ is invertible, we recover $\hat{\beta}^{LS} := (X^TX)^{-1}X^TY$. Also, note that in general, if $\hat{\beta}^{LS}$ is a least squares estimator $\delta \in \text{Kernel}(X)$ 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 $r = \text{dim}(\text{column space}(X))$ and $\hat{\beta} = \hat{\beta}^{LS}$ Then, $\exists C > 0$ such that

$$\frac{1}{n} \mathbb{E} \left[ ||X(\beta^* - \hat{\beta})||^2 \right] \leq C\frac{\sigma^2 r}{n}, \text{ and} \quad (8.9)$$

$$\mathbb{P} \left( \frac{1}{n} ||X(\beta^* - \hat{\beta})||^2 \leq C\frac{\sigma^2 r + \log \left( \frac{1}{\delta} \right)}{n} \right) \geq 1 - \delta \quad (8.10)$$

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

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

Also, we have that,

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

Putting these two together yields
\[||X(\hat{\beta} - \beta^*)||^2 \leq 2 \left\langle \epsilon, X(\hat{\beta} - \beta^*) \right\rangle \rightarrow \] (8.13)

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

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

\[
\frac{X(\hat{\beta} - \beta^*)}{||X(\hat{\beta} - \beta^*)||} = \Phi v, \rightarrow \] (8.15)

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

where \(\tilde{\epsilon} = \epsilon^T \Phi\)

We therefore have that

\[||X(\hat{\beta} - \beta^*)||^2 \leq 4 \max_{v \in S^{r-1}} (\tilde{\epsilon}^Tv)^2 \] (8.17)

Since \(\Phi\) 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

\[
\leq 4 \max_{v \in S^{r-1}} (\tilde{\epsilon}^Tv)^2 \leq 4 \sum_{j=1}^r \mathbb{E}[\tilde{\epsilon}_j]^2 \leq 16\sigma^2r \] (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 \(S^{r-1}\).

\[
\max_{v \in S^{r-1}} (\tilde{\epsilon}^Tv) \leq 2 \max_{z \in \mathcal{N}_{1/2}} (\tilde{\epsilon}^Tz) \rightarrow \] (8.19)

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

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

Setting the above equal to \(d\) and solving for \(t\) yields the desired result. 

Also, note that
\[ ||\hat{\beta} - \beta^*||^2 \lambda^2_{\text{min}}(X) \leq ||X (\hat{\beta} - \beta^*)||^2 \] (8.22)

which gives us a meaningful (though not necessarily optimal) bound on \( ||\hat{\beta} - \beta^*||^2 \) if \( \lambda^2_{\text{min}}(X) > 0 \). This does not help us, of course, when \( d > n \) as in that case \( \lambda^2_{\text{min}}(X) = 0 \) 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 \( \hat{\beta}^{\text{PLS}} \) satisfies

\[
\hat{\beta}^{\text{PLS}} \in \arg\min_{\beta} \left\{ \frac{1}{2n} ||Y - X\beta||^2 + \lambda_n f(\beta) \right\} \quad (8.23)
\]

The LASSO estimator \( \hat{\beta}^{\text{LASSO}} \) is the penalized least squares estimator with the \( \ell_1 \) 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 \( \lambda_n \), \( B \) and \( R \) respectively:

\[
\arg\min_{\beta} \frac{1}{2n} ||Y - X\beta||^2 + \lambda_n ||\beta||_1 \quad (8.24)
\]

\[
\arg\min_{\beta} ||\beta||_1 \text{ s.t.} \frac{1}{2n} ||Y - X\beta||^2 \leq B^2 \quad (8.25)
\]

\[
\arg\min_{\beta} \frac{1}{2n} ||Y - X\beta||^2 \text{ s.t.} ||\beta||_1 \leq R \quad (8.26)
\]

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

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

\[
\frac{||X (\hat{\beta} - \beta^*)||^2}{n} \leq 4||\beta^*||_1 \lambda_n \quad (8.27)
\]