LECTURE 2
Linear regression the proceduralist approach

2.1. Standard multivariate linear regression

The regression problem is usually stated as

$$Y = f(X) + \varepsilon, \quad \varepsilon \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$$

the multivariate random variable $X \subseteq \mathbb{R}^p$ are called covariates and the univariate random variable $Y \subseteq \mathbb{R}$ is called the response, the function belongs to a class of functions $f \in \mathcal{F}$. The random variables X, Y have associated with them the following distributions

$$\text{joint: } \rho_{X,Y}(x,y), \quad \text{marginals: } \rho_X(x), \rho_Y(y), \quad \text{conditional: } \rho(y \mid x).$$

For now the class of functions $\mathcal{F}$ will be linear functions

$$f(x) = \beta^T x, \quad \beta \in \mathbb{R}^p.$$

The data we are given consists of n observations $D = \{(x_i, y_i)\}_{i=1}^n \overset{iid}{\sim} \rho(x,y)$, we call this a sample and the sample size is n. We will assume the data we observed is consistent with the following model

$$Y_i = \beta^T X_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2).$$

Our goal is given the data D solve the following problems:

(1) parameter inference: what is a reasonable estimate for β, we'll call the estimate $\hat{\beta}$
(2) prediction: given a new x_* what is the corresponding y_*, try $y_* = \hat{\beta}^T x_*$.
(3) estimating the conditional distribution: what is $Y \mid X = x$.

An idea to estimate $\hat{\beta}$ that goes back to Gauss is the minimizing the least squares error

$$\arg\min_{\beta \in \mathbb{R}^p} \left[\frac{1}{n} \sum_i (y_i - \beta^T x_i)^2, \right].$$

One can derive the above estimator from the following probabilistic model

$$\mathrm{Lik}(D; \beta) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\|y_i - \beta^T x_i\|^2}{2\sigma^2} \right),$$

by maximizing the likelihood above with respect to β

$$\arg\max_{\beta} \text{Lik}(D;\beta) \equiv \arg\min_{\beta \in \mathbb{R}^p} \left[\frac{1}{n} \sum_i (y_i - \beta^T x_i)^2, \right].$$

We can state the negative of the log likelihood as

$$L = \sum_{i=1}^n (y_i - x_i^T \beta)^2.$$

We will rewrite the above in matrix notation. In doing this we define a matrix $\mathbf{X}$ which is $n \times p$ and each row of the matrix is a data point x_i. We also define a column vector Y $(p \times 1)$ with y_i as the i-th element of y. Similarly β is a column vector with p rows. We can rewrite the error minimization as

$$\arg\min_{\beta} \left[L = (Y - \mathbf{X}\beta)^T (Y - \mathbf{X}\beta) \right],$$

taking derivatives with respect to θ and setting this equal to zero (taking derivatives with respect to β means taking derivatives with respect to each element in β)

$$\begin{aligned} \frac{dL}{d\beta} &= -2\mathbf{X}^T(Y - \mathbf{X}\beta) = 0 \\ &= \mathbf{X}^T(Y - \mathbf{X}\beta) = 0. \end{aligned}$$

this implies

$$\begin{aligned} \mathbf{X}^T Y &= \mathbf{X}^T\mathbf{X}\beta \\ \hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y. \end{aligned}$$

If we look at the above formula carefully there is a serious numerical problem – $(\mathbf{X}^T\mathbf{X})^{-1}$. The matrix $\mathbf{X}^T\mathbf{X}$ is a $p \times p$ matrix of rank n where $p \gg n$. This means that $\mathbf{X}^T\mathbf{X}$ cannot be inverted so we cannot compute the estimate $\hat{\beta}$ by matrix inversion. There are numerical approaches to address this issue but the solution will not be unique or stable. A general rule in estimation problems is that numerical problems in estimating the parameters usually coincide with with statistical errors or variance of the estimate.

2.2. The Stein estimator

The above numerical problem is related to an amazing result that was first observed in 1956 by Charles Stein. The question that Stein asked is if one is given n observations from a multivariate normal

$$(x_i)_{i=1}^n \overset{iid}{\sim} \text{N}(\theta, \sigma^2\mathbf{I}),$$

what is the best estimator of θ. In statistics if there exists an estimator that is better than the one you are using then your estimator is called inadmissible. What Stein found was that for $p \geq 3$ the sample mean

$$\hat{\theta} = \frac{1}{n} \sum_i x_i,$$

is not admissible. A better estimator called the James-Stein estimator is the following

$$\hat{\theta} = \left(1 - \frac{(p-2)\frac{\sigma^2}{n}}{\|\bar{x}\|^2}\right) \bar{x}.$$

The intuition about the above estimator is take the sample mean $\bar{x}$ and move it a bit towards zero, this is called shrinkage or shrinkage towards zero. Now what is an optimal estimator, it is one which minimizes

$$\arg\min_{\beta \in \mathbb{R}^p} \left[\mathbb{E}_{X,Y}\left[(y - \beta^T x)^2\right] = \int_{Y,X} (y - \beta^T x)^2 \rho(x,y) \, \mathrm{d}x \, \mathrm{d}y\right],$$

the idea above is minimizing the error on unseen data, later in the course we will call the above the generalization error

$$I[\hat{f}] = \mathbb{E}_{X,Y}\left[(y - \hat{f}(x))^2\right], \quad \hat{f}(x) = \hat{\beta}^T x,$$

and provide some theory to justify the estimators we will propose.

2.3. Cross-validation

One cannot estimate the generalization error since one does not have access to the generating distribution $\rho(x,y)$. A common proxy for the generalization error is the leave-one-out cross-validation error

$$I[cv] \equiv \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_{D^{\setminus i}}(x_i))^2,$$

where $D^{\setminus i}$ is the data set with the i-th sample removed and $\hat{f}_{D^{\setminus i}}$ is the function estimated when the i-th sample is removed. The idea is to remove the i-th sample, estimate a function with that sample left out, then test the error made on the i-th sample, and average this over all n observations. Of course one need not leave-out one observation but can leave-out k observations. The leave-one-out estimator is (almost) unbiased.

2.4. Shrinkage models

We can now adapt the idea behind the James-Stein estimator to the linear regression problem. We will minimize the following loss function

$$\arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\|\beta\|^2,$$

where $\lambda > 0$ is a parameter and $\|\beta\|^2 = \sum_{i=1}^{p} \beta_i^2$.

$$\begin{aligned}\frac{dL}{d\beta} &= -2\mathbf{X}^T(Y - \mathbf{X}\beta) + 2\lambda n\beta = 0 \\ &= \mathbf{X}^T(\mathbf{X}\beta - Y) + \lambda n\beta.\end{aligned}$$

this implies

$$\begin{aligned}\mathbf{X}^T Y &= \mathbf{X}^T\mathbf{X}\beta + \lambda n\beta \\ \mathbf{X}^T Y &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})n\beta \\ \hat{\beta} &= (\mathbf{X}^T\mathbf{X} + \lambda n\mathbf{I})^{-1}\mathbf{X}^T Y,\end{aligned}$$

where $\mathbf{I}$ is the $p \times p$ identity matrix. The matrix $(\mathbf{X}^T\mathbf{X} + \lambda n \mathbf{I})$ is invertible and this penalized loss function approach has had great success in problems with more variables than observations.