Classical Estimation Topics

Namrata Vaswani, Iowa State University

February 25, 2014

This note fills in the gaps in the notes already provided (l0.pdf, l1.pdf, l2.pdf, l3.pdf, LeastSquares.pdf).

1 Min classical Mean Squared Error (MSE) and Minimum Variance Unbiased Estimation (MVUE)

1. First, assume a scalar unknown parameter θ
   (a) Min classical MSE estimator:
   $$\hat{\theta}(X) = \arg\min_{\hat{\theta}} \mathbb{E}_X[(\theta - \hat{\theta}(X))_2^2]$$
   (b) Often the resulting estimator is not realizable, i.e. it depends on θ.
   (c) MVUE estimator:
   $$\hat{\theta}_{MVUE}(X) = \arg\min_{\hat{\theta}:\mathbb{E}[\hat{\theta}(X)]=\theta} \mathbb{E}_X[(\theta - \hat{\theta}(X))_2^2] = \arg\min_{\hat{\theta}:\mathbb{E}[\hat{\theta}(X)]=\theta} var[\hat{\theta}(X)]$$
2. Vector parameter θ
   (a) MVUE: $\hat{\theta}_{MVUE,i}(X) = \arg\min_{\hat{\theta}_i:\mathbb{E}[\hat{\theta}_i(X)]=\theta_i} var[\hat{\theta}_i(X)]$
3. Sufficient statistic (ss)
   (a) A stat $Z := T(X)$ is a ss for θ if $p_{X|Z}(x|T(x);\theta) = p_{X|Z}(x|T(x))$, i.e. it *does not* depend on θ
4. Minimal ss
   (a) A stat $T(X)$ is a minimal ss if it is a ss and it is a function of every other ss.
5. Complete ss:

(a) $T(X)$ is a complete ss for θ iff it is a ss and if $\mathbb{E}[v(T(X))] = 0$ (expectation taken w.r.t. pdf/pmf $p(x;\theta)$) for all θ, implies that $Pr_\theta(v(T(X)) = 0) = 1$ (if we can show that $v(t) = 0$ for all t, this gets satisfied).

(b) $T(X)$ is a complete ss for θ iff it is a ss and there is *at most one* function $g(t)$ such that $g(T(X))$ is an unbiased estimate of θ, i.e. $\mathbb{E}[g(T(X))] = \theta$.

6. Factorization Theorem (Neyman) to find a ss

   (a) A stat $T(X)$ is a ss for θ iff the pmf/pdf $p_X(x;\theta)$ can be factorized as

   $$p_X(x;\theta) = g(T(x), \theta)h(x)$$

   for all x and for all $\theta \in \Theta$ (Θ: parameter space).

   (b) Proof: proof for the discrete rv case is easy and illustrates the main point. Idea: just use definition of ss.

7. Rao-Blackwell-Lehmann-Scheffe (RBLS) theorem:

   (a) RB theorem: given a ss, and some unbiased estimator for θ, find another unbiased estimator with equal or lower variance. Statement: Let $\check{\theta}(x)$ be an unbiased estimator for θ. Let $T(X)$ be a ss for θ. Let $Z := T(X)$. Define a function

   $$\hat{\theta}(z) := \mathbb{E}_{X|Z}[\check{\theta}(X)|z].$$

   Then

   i. $\hat{\theta}(T(x))$ is a realizable estimator
   ii. $\hat{\theta}(T(X))$ is unbiased, i.e. $\mathbb{E}_X[\hat{\theta}(T(X))] = \theta$
   iii. $var[\hat{\theta}(T(X))] \leq var[\check{\theta}(X)]$

   (b) Proof:

   i. follows from the definition of a ss
   ii. follows by using iterated expectation: $\mathbb{E}_X[\hat{\theta}(T(X))] = \mathbb{E}_X[\mathbb{E}_{X|Z}[\check{\theta}(X)|T(X)]] = \mathbb{E}_X[\check{\theta}(X)] = \theta$
   iii. follows by using conditional variance identity.

   (c) LS theorem: if $T(X)$ is a complete ss for θ and if there is a function $g(t)$ s.t. $\mathbb{E}[g(T(X))] = \theta$, then $g(T(X))$ is the MVUE for θ.

   (d) LS theorem (equivalent statement): if $T(X)$ is a complete ss for θ, then $\hat{\theta}(T(X))$ defined above is MVUE for θ and in fact $g(T(X)) = \hat{\theta}(T(X))$

   (e) Proof: follows from the fact that $\hat{\theta}(T(X)) := \mathbb{E}_{X|T(X)}[\check{\theta}(X)|T(X)]$ is a function of $T(X)$

(f) Thus, LS theorem implies that if I can find a function, $g(t)$, of a complete ss, $T(X)$ that is an unbiased estimate of θ, then $g(T(X))$ is the MVUE. Or if take any unbiased function and compute its conditional expectation conditioned on the complete ss, then also I will get the MVUE.

8. Completeness Theorem for Exponential Families: see later

9. Examples

   (a) Example proving completeness of a ss: Kay's book

   (b) MVUE computation

2 Information Inequality and Cramer Rao Lower Bound

1. l2.pdf is quite complete.

2. Poor's book also talks about the scalar case CRLB.

3. score function: derivative of log likelihood w.r.t. θ, score $= \frac{\partial \log p(X;\theta)}{\partial\theta}$

4. Under "regularity",

   (a) expected value of score function is zero

   (b) Fisher Information Matrix (or Number for a scalar) is defined as $\mathbb{E}[\text{score score}^T]$

   (c) under more "regularity" FIM is the negative expected value of the derivative of score (Hessian of log likelihood w.r.t. θ)

5. Info inequality and CRLB (scalar case): Assume "regularity".

   (a) Consider a pdf/pmf family $p(x;\theta)$. Assume that $0 < I(\theta) < \infty$. The variance of any statistic $T(X)$ with finite variance and with $\mathbb{E}[T(X)] = \psi(\theta)$ is lower bounded as follows
   $$var[T(X)] \geq \frac{|\psi'(\theta)|^2}{I(\theta)}$$
   with equality occurring if and only if score is an affine function of $T(X)$, i.e.
   $$\text{score} = k(\theta)(T(X) - b(\theta))$$

   (b) Under "regularity", this is achieved if and only if $p(x;\theta)$ is a one parameter exponential family.

   (c) Proof idea:

      i. Write out expression for $\psi'(\theta)$ and re-arrange it as $\mathbb{E}[T(X)\text{ score}]$.

ii. Use Cauchy-Schwartz and the fact that the score function is zero mean

(d) details: see page 6 of l2.pdf or see Kay's book (Appendix of Chap 3) or see Poor's book

6. Info inequality and CRLB vector case: Assume "regularity".

   (a) Consider a pdf/pmf family $p(x;\theta)$. Assume that the FIM $I(\theta)$ is non-singular. Consider any vector statistic $T(X)$ with finite variance in all directions and with $\mathbb{E}[T(X)] = \psi(\theta)$. Then,

   $$cov[T(X)] \succeq \psi'(\theta) I(\theta)^{-1} \psi'(\theta)^T$$

   with equality occurring if and only if score is an affine function of $T(X)$, i.e.

   $$\text{score} = K(\theta)(T(X) - b(\theta))$$

   (for matrices $M_1 \succeq M_2$ means $a'(M_1 - M_2)a \geq 0$ for any vector a).

   (b) CRLB: Special case where $\psi(\theta) = \theta$. In this case, $cov[T(X)] \succeq I(\theta)^{-1}$ with equality if and only if

   $$\text{score} = I(\theta)(T(X) - \theta)$$

   (c) Here $\psi'(\theta) := \frac{\partial \psi(\theta)}{\partial \theta^T}$, i.e. $(\psi'(\theta))_{i,j} = \frac{\partial \psi_i(\theta)}{\partial \theta_j}$. From this notation notice that $\frac{\partial \psi(\theta)}{\partial \theta^T} = (\frac{\partial \psi^T(\theta)}{\partial \theta})^T$.

   (d) Proof idea:

      i. Write out expression for $\psi'(\theta)$ and re-arrange it as $\psi'(\theta) = \mathbb{E}[T(X) \text{ score}^T]$

      ii. $\mathbb{E}[T(X) \frac{\partial \log p(X;\theta)}{\partial \theta}^T]$ is now a matrix

      iii. Apply C-S to $a'\psi'(\theta)b = a'\mathbb{E}[T(X) \text{ score}^T]b$ and use the fact that the score function is zero mean to get $(a'\psi'(\theta)b)^2 \leq var(a'T(X)) var(\text{score}^T b)$

      iv. Notice that $var(a'T(X)) = a' cov(T(X)) a$ and $var(\text{score}^T b) = var(b^T \text{score}) = b^T I(\theta) b$

      v. Set $b = I(\theta)^{-1} \psi'(\theta)^T a$ to get the final result.

   (e) Details: see (Appendix of Chap 3 of Kay's book)

7. We say an estimator is efficient if it is unbiased and its variance is equal to the Cramer Rao lower bound.

8. *If more parameters are unknown, the CRLB is larger (or equal if the FIM is diagonal).* Consider a pdf/pmf with 2 parameters. First suppose that only one parameter is unknown and suppose its CRB is c. Now for the same pdf/pmf if both the parameters are unknown, the CRB will be greater than or equal to c. It is equal to c only if the FIM for the 2-parameter case is diagonal. Same concept extends to multiple parameters.

(a) Denote the FIM for the 2-parameter case by $I(\theta)$. Recall that $[I_{11}(\theta)]^{-1}$ is the CRLB for θ_1 (when θ_2 is known). When both are unknown then, $[I(\theta)^{-1}]_{11}$ is the CRLB. We claim that

$$[I(\theta)^{-1}]_{11} \geq [I_{11}(\theta)]^{-1}$$

with equality if and only if $I(\theta)$ is a diagonal matrix.

(b) This, in turn, follows by using C-S for vectors on $\sqrt{I(\theta)}e_1$, $\sqrt{I(\theta)^{-1}}e_1$

9. Gaussian CRB: see l2.pdf, Theorem 5, page 32.

10. Examples

3 Exponential Family

1. Multi-parameter expo family:

$$p(x;\theta) = h(x)C(\theta)\exp\left[\sum_{i=1}^{k} \eta_i(\theta)T_i(x)\right]$$

   (a) single-parameter expo family: special case where $k = 1$.

   (b) Examples: Gaussian, Poisson, Laplacian, binomial, geometric

2. By factorization theorem, easy to see that $T(X) = [T_1(X), \dots T_k(X)]'$ is the ss for the vector parameter θ.

3. Completeness Theorem: if the parameter space for the parameters $\eta_i(\theta)$'s, contains a k-dimensional hyper-rectangle, then $T(X)$ is a complete ss for $\eta_i(\theta)$'s.

   (a) See proposition IV.C.3 of Poor's book (that is stated by first re-parameterizing $p(x;\theta)$ as $p(x;\phi) = h(x)C(\phi)\exp(\sum_{i=1}^{k} \phi_i T_i(x))$ to make things easier).

4. Example IV.C.3 of Poor's book

5. Easy to see that the support of $p(x;\theta)$, i.e. the set $\{x : p(x;\theta) > 0\}$, does not depend on θ.

6. One parameter expo family and EE/MVUE: Under "regularity" (the partial derivative w.r.t. θ can be moved inside or outside the integral sign when computing the expectation of any statistic, and if $E[|T_1(X)|] < \infty$), $T_1(X)$ is the efficient estimator (EE), and hence MVUE, for its expected value.

   (a) In fact, under regularity, an estimator $T(X)$ achieves the CRLB for $\eta(\theta) = \mathbb{E}_\theta[T(X)]$ if and only if $p(x;\theta) = h(x)C(\theta)\exp(\eta(\theta)T(X))$ (one parameter expo family of the form).

(b) See Example IV.C.4 of Poor for a proof.

7. Multi-parameter expo family: The vector $T(X)$ is an EE and hence MVUE for its expected value.

   (a) Proof:

      i. Rewrite expo family distribution as $p(x;\theta) = h(x)C(\eta(\theta))\exp\left[\sum_{i=1}^{k}\eta_i(\theta)T_i(x)\right]$. This is always possible to do because $C(\theta)$ is given by
      $$\frac{1}{C(\theta)} = \int h(x)\exp\left[\sum_{i=1}^{k}\eta_i(\theta)T_i(x)\right]dx$$
      and hence it is actually a function of $\eta(\theta)$.

      ii. Thus expo family distribution can always be re-parameterized in terms of η as
      $$p(x;\eta) = h(x)\exp\{[\sum_{i=1}^{k}\eta_i T_i(x)] - A(\eta)\}, \quad A(\eta) := -\log C(\eta)$$

      iii. With this, clearly,
      $$score = T(X) - \frac{\partial A(\eta)}{\partial \eta}$$

      iv. Since $\mathbb{E}[\text{score}] = 0$, thus, $\mathbb{E}[T(X)] = \frac{\partial A(\eta)}{\partial \eta}$

      v. Thus, $cov(T(X)) = \mathbb{E}[\text{score score}^T] = I(\eta)$

      vi. Also $\psi(\eta) = \mathbb{E}[T(X)] = \frac{\partial A(\eta)}{\partial \eta}$ implies that $\psi'(\eta) = \frac{\partial^2 A(\eta)}{\partial \eta \eta^T}$

      vii. But $I(\eta)$ also satisfies $I(\eta) = \mathbb{E}[-\frac{\partial \text{score}}{\partial \eta^T}] = \frac{\partial^2 A(\eta)}{\partial \eta \eta^T} = \psi'(\eta)$

      viii. Thus, $\psi'(\eta)I(\eta)^{-1}\psi'(\eta)^T = I(\eta)I(\eta)^{-1}I(\eta)^T = I(\eta) = cov(T(X))$

      ix. Thus, $T(X)$ is EE and hence MVUE of its expected value.

   (b) Example on page 30 of l2.pdf of applying this theorem.

   (c) But to my knowledge there is no "if and only if" result, i.e. one cannot say that an estimator achieves CRLB for its expected value only for multi-parameter expo families. ??check

8. Some more properties and FIM expression for single parameter expo families in l2.pdf, page 13-17.

4 Linear Models

1. Linear model means the data $\underline{X}$ satisfies
$$X = H\theta + W$$

where X is an $N \times 1$ data vector, θ is a $p \times 1$ vector of unknown parameters and W is the zero mean noise, i.e. $\mathbb{E}[W] = 0$.

2. The above model is identifiable iff H has rank p.

3. If W is Gaussian noise, then the MVUE exists. In fact the MVUE is also the efficient estimator (EE).

4. If $W \sim \mathcal{N}(0, C)$ then,

$$\hat{\theta}_{MVUE}(X) = (H'C^{-1}H)^{-1}H'C^{-1}X$$

Proof:

(a) Show unbiasedness

(b) Compute $cov[\hat{\theta}_{MVUE}(X)]$ and show that it is equal to the CRLB.

5 Best Linear Unbiased Estimation (BLUE)

1. For any given pdf/pmf $p(X;\theta)$, find the "best" estimator among the class of linear and unbiased estimators. Here "best" means minimum variance.

2. Scalar parameter θ: $\hat{\theta}_{BLUE}(X) = a'_B X$ where a_B is a vector obtained by solving

$$a_B = \arg \min_{a : a'\mathbb{E}[X] = \theta} a' cov(X) a$$

(recall that the expectation of any linear estimator, $a'X$, is $a'\mathbb{E}[X]$ and its variance is $a' cov(X) a$).

3. Vector parameter θ: $\hat{\theta}_{BLUE}(X) = A'_B X$ where A_B is a $n \times p$ matrix obtained by solving

$$(A_B)_i = \arg \min_{A : A'_i \mathbb{E}[X] = \theta_i} A'_i cov(X) A_i$$

here A_i refers to the i^{th} column of the matrix A.

4. To prove that a given matrix A_B is the minimizer: typical approach is as follows. Try to show that

(a) $A'_B \mathbb{E}[X] = \theta$

(b) For all matrices A satisfying $A'\mathbb{E}[X] = \theta$, the following holds

$$A' cov(X) A - A'_B cov(X) A_B \succeq 0$$

(here $M \succeq 0$ means that the matrix M is positive semi-definite, i.e. it satisfies $z'Mz \geq 0$ for any vector z).

(c) By letting $z = e_i$ where e_i is a vector with 1 at the i^{th} coordinate and zero everywhere else, we can see that the above implies that A_B indeed is the BLUE.

5. Example of finding a BLUE: l2.pdf

6. Example situation where cannot find even one linear estimator that is unbiased, and so BLUE does not exist: l2.pdf

   (a) In the above case, if transform the data in some way, it may be possible to find a BLUE.

6 Maximum Likelihood Estimation

1. Define MLE.

2. We assume Identifiability: $p(x, \theta_1) = p(x, \theta_2)$ if and only if $\theta_1 = \theta_2$

   - Example: in case of a linear model of the form $X = H\theta + W, W \sim \mathcal{N}(0, \sigma^2 I)$, so that $p(x; \theta) = \mathcal{N}(x; H\theta, \sigma^2 I)$, θ is identifiable if and only if H has full column rank. This, in turn, means that H has to be a full rank square or tall matrix.

3. Given $X_1, \ldots X_N$ iid with pdf or pmf $p(x; \theta)$, i.e. for discrete case, $Pr(X_i = x) = p(x; \theta)$, for continuous case, $Pr(X_i \in [x, x + dx]) \approx p(x; \theta)dx$ (or more precisely $Pr(X_i \leq x) = \int_{t=\infty}^{x} p(t; \theta)$).

4. Then define $\hat{\theta}_N(\underline{\mathrm{X}}) := \arg\max_\theta \prod_{i=1}^{N} p(X_i; \theta)$.

5. Consistency and Asymptotic Normality of MLE: If $X_1, X_2, \ldots X_n$ iid $p(x; \theta)$, then under certain "regularity conditions", $\hat{\theta}_N(\underline{\mathrm{X}})$ is consistent and asymptotically normal, i.e.

$$\text{for any } \epsilon > 0, \ \lim_{N \to \infty} Pr(|\hat{\theta}_N(\underline{\mathrm{X}}) - \theta| > \epsilon) = 0, \ \text{and}$$

$$\sqrt{N}(\hat{\theta}_N((\underline{X})) - \theta) \to Z \sim \mathcal{N}(0, i_1(\theta)^{-1}), \ \text{in distribution as } N \to \infty$$

   where

$$i_1(\theta) := \mathbb{E}[(\frac{\partial}{\partial \theta} \log p(X_1; \theta))^2] = \mathbb{E}[-\frac{\partial^2}{\partial \theta^2} \log p(X_1; \theta)]$$

   is the Fisher information number for X_1.

   Proof approach:

   (a) Show consistency of $\hat{\theta}_N$

      i. See Poor's book for a correct proof. See Appendix of Kay's book for this rough idea:

ii. Jensen's inequality (or non-negativity of Kullback-Leibler divergence), followed by taking a derivative w.r.t. θ on both sides, tells us that $\int p(x;\theta)\frac{\partial}{\partial\theta}\log p(x;\theta)dx \geq \int p(x;\theta)\frac{\partial}{\partial\theta}\log p(x;\tilde{\theta})dx$ or in other words $\arg\max_{\tilde{\theta}} \int p(x;\theta)\frac{\partial}{\partial\theta}\log p(x;\tilde{\theta})dx = \theta$.

iii. The MLE, $\hat{\theta}_N(\underline{\mathrm{X}}) = \arg\max_{\tilde{\theta}} \frac{1}{N}\sum_i \frac{\partial}{\partial\theta}\log p(X_i;\tilde{\theta})$.

iv. By WLLN, $\frac{1}{N}\sum_i \frac{\partial}{\partial\theta}\log p(X_i;\tilde{\theta})$ converges to its expected value, $\int p(x;\theta)\frac{\partial}{\partial\theta}\log p(x;\tilde{\theta})dx$, in probability.

v. By using an appropriate "continuity argument", its maximizer, $\hat{\theta}_N$, also converges in probability to the maximizer of the RHS, which is θ.

(b) $\frac{1}{N}\sum_i \frac{\partial}{\partial\theta}\log p(X_i;\hat{\theta}_N) = 0$ by definition of the MLE (first derivative equal to zero for maximizer): for differentiable functions with maximizer inside an open region.

(c) Using the above fact and Mean Value Theorem on $\frac{1}{N}\sum_i \frac{\partial}{\partial\theta}\log p(X_i;\hat{\theta}_N)$, we can rewrite

$$\sqrt{N}(\hat{\theta}_N - \theta) = \frac{\sqrt{N}R_N(\hat{\theta}_N)}{T_N(\tilde{\theta}_N)}$$

for some $\tilde{\theta}_N$ lying in between $\hat{\theta}_N$ and θ

(d) Use consistency of $\hat{\theta}_N$ to show that $T_N(\tilde{\theta}_N)$ converges to $T_N(\theta)$ in probability and hence in distribution.

(e) Use WLLN to show that $T_N(\theta)$ converges in probability to $i_1(\theta)$.

(f) Use Central Limit Theorem on $\sqrt{N}R_N(\hat{\theta}_N)$ to show that it converges in distribution to $Z \sim \mathcal{N}(0, i_1(\theta))$.

(g) Use Slutsky's theorem to get the final result

(h) For details: see class notes or see Appendix of Chapter 7 of Kay's book or see Poor's book.

6. Define Asymptotic Efficiency: two ways that different books define it.

(a) First definition (used in Lehmann's book and mentioned in Poor's book): $\sqrt{N}(\hat{\theta}_N - \theta)$ converges to a random variable Z in distribution and $E[Z] = 0$ and $Var[Z] = i_1(\theta)^{-1}$. The asymptotic normality proof directly implies this.

(b) Second definition: $\hat{\theta}_N$ is asymptotically unbiased, i.e. $\lim_{N\to\infty} \mathbb{E}[\hat{\theta}_N(\underline{\mathrm{X}})] = \theta$ and its asymptotic variance is equal to the CRLB, i.e. $\lim_{N\to\infty} Var[\sqrt{N}\hat{\theta}_N(\underline{\mathrm{X}})] = i_1(\theta)^{-1}$.

i. Under more regularity conditions, I believe that proofs of the above statement do exist.

ii. **Note though: I did not prove the above in class. Notice: neither convergence in probability nor convergence in distribution imply**

convergence of the moments, i.e. neither implies that $\lim_{N\to\infty} \mathbb{E}[\hat{\theta}_N(\underline{\mathbf{X}})] = \theta$ **or that the variance converges.**

7. Need for MLE: consider the example $X_n \sim \mathcal{N}(A, A)$, iid. In this case, we showed that we do not know how to compute either an efficient estimator (EE) or a minimum variance unbiased estimator (MVUE). But MLE is always computable, either analytically or numerically. In this case, it is analytically computable.

8. If an efficient estimator (EE) exists, then MLE is equal to it (Theorem 5 of l3.pdf)

   - Proof: easy. If EE exists, score$(\theta) = I(\theta)(\hat{\theta}_{EE}(X) - \theta)$. If MLE lies inside an open interval of parameter space, it satisfies, score$(\hat{\theta}_{ML}(X)) = 0$. Thus, if EE exists $\hat{\theta}_{EE} = \hat{\theta}_{ML}$.
   - Vice versa is not true for finite N, but is true asymptotically under certain "regularity": discussed above.

9. ML Invariance Principle: Theorem ? of l3.pdf

10. Examples showing the use of ML invariance principle: amplitude and phase estimation of a sinusoid from a sequence of noisy measurements: l3.pdf

11. Example application in digital communications: ML bit decoding: l3.pdf

12. Newton Raphson method and its variants: l3.pdf

7 Least Squares Estimation

1. No probability model at all. Just a linear algebra technique that finds an estimate of θ that minimizes the 2-norm of the error, $\|X - f(\theta)\|_2^2$.

2. Closed form solutions exist for the linear model case, i.e. the case where $X = H\theta + E$ and we want to find the θ that minimizes $\|E\|_2^2$.

3. Assume that X is an $n \times 1$ vector and θ is a $p \times 1$ vector. So H is an $n \times p$ matrix.

4. LS:
$$\hat{\theta} = \arg\min_{\theta} \|X - H\theta\|_2^2, \quad \|X - H\theta\|_2^2 := (X - H\theta)'(X - H\theta)$$

5. Weighted LS
$$\hat{\theta} = \arg\min_{\theta} \|X - H\theta\|_W^2, \quad \|X - H\theta\|_W^2 := (X - H\theta)'W(X - H\theta)$$

6. Regularized LS

$$\hat{\theta} = \arg\min_{\theta} \|\theta - \theta_0\|_R^2 + \|X - H\theta\|_W^2$$

7. Notice that both weighted LS and LS are special cases of regularized LS with $R = 0$ (weighted LS), $R = 0, W = I$ (LS).

8. Recursive LS algorithm: recursive algorithm to compute regularized LS estimate. Derived in LeastSquares.pdf

9. Consider basic LS. Recall that H is an $n \times p$ matrix.

$$\hat{\theta} = \arg\min_{\theta} \|X - H\theta\|_2^2$$

   Two cases: $rank(H) = p$ and $rank(H) < p$

   (a) If $rank(H) = p$, the minimizer is unique and given by

$$\hat{\theta} = (H'H)^{-1}H'x$$

   (b) If $rank(H) < p$, there are infinitely many solutions.
   $rank(H) < p$ can happen in two ways

      i. If $n < p$ (fat matrix), then definitely $rank(H) < p$
      ii. Even when $n \geq p$, (square or tall matrix), it could be that the columns of H are linearly dependent, e.g. suppose $p = 3$ and $H = [H_1, H_2, (H_1 + H_2)]$, then $rank(H) \leq 2 < p$.

10. Nonlinear LS: $\hat{\theta} = \arg\min_\theta \|X - f(\theta)\|_2^2$.

   (a) In general: no closed form solution, use Newton Raphson, or any numerical optimization algorithm.
   (b) If partly linear model, i.e. if $\theta = [\alpha, \beta]$ and $f(\theta) = H(\alpha)\beta$, then

      i. first compute closed form solution for β in terms of α, i.e. $\hat{\beta}(\alpha) = [H(\alpha)'H(\alpha)]^{-1}H(\alpha)'X$
      ii. solve for α numerically by solving $\min_\alpha \|X - H(\alpha)\hat{\beta}(\alpha)\|_2^2$