STA5001: High Dimensional Statistics

Gong Wenwu 12031299

April 11, 2021

1. Weighted Least Square Problem

- (a). WTS: for any matrix B, define $P = B(B^TB)^{-1}B^T$, we can infer that $P^2 = P$, i.e., the eigenvalue of P is 0 or 1.
  Proof:

$$Since\ P^2 = B(B^TB)^{-1}B^T \times B(B^TB)^{-1}B^T = B(B^TB)^{-1}B^T = P,$$
$$then\ P(P - 1_n) = 0,\ i.e.,\ the\ eigenvalue\ of\ P\ is\ 0\ or\ 1.$$

  So, $A^T(1_n - B(B^TB)^{-1}B^T)A \succeq 0$, this complete $A^TB(B^TB)^{-1}B^TA \preccurlyeq A^TA$. □

- (b). The performance of estimator $\hat{\beta}$ w.r.t. wrong correlation matrix $W = diag(W_0)$ under WLS method: unbiased and $n^{1/2}$ consistent estimator.
  Proof: Under true covariance matrix W_0, the linear regression model $Y = X\beta + \epsilon$, where $\epsilon \sim N(0, \sigma^2 W_0)$, σ^2 and W_0 are both unknown. We assure that $\tilde{\beta}$ is the BLUE:

$$\tilde{\beta} = (X^TW_0^{-1}X)^{-1}X^TW_0^{-1}X\beta + (X^TW_0^{-1}X)^{-1}X^TW_0^{-1}\epsilon,$$
$$E(\tilde{\beta}) = \beta,\ Var(\tilde{\beta}) = (X^TW_0^{-1}X)^{-1}\sigma^2.$$

  For the wrong correlation matrix $W = diag(W_0)$:

$$\hat{\beta} = (X^TW^{-1}X)^{-1}X^TW^{-1}Y = AY,\ linear,$$
$$E(\hat{\beta}) = E((X^TW^{-1}X)^{-1}X^TW^{-1}X\beta) = \beta,\ unbiased,$$
$$Var(\hat{\beta}) = (X^TW^{-1}X)^{-1}X^TW^{-1}W_0W^{-1}X(X^TW^{-1}X)^{-1}\sigma^2,\ order\ O(n^{-1}).$$

  Then $Var(\hat{\beta}) \succeq Var(\tilde{\beta})$ holds (BLUE of $\tilde{\beta}$). (i.e., there is matrix B that $Var(\hat{\beta}) = A^TA\,Var(\epsilon) \succeq A^TB(B^TB)^{-1}B^TA\,Var(\epsilon) = Var(\tilde{\beta})$ holds.) □

2. Linear model inference under Gaussian-Markov conditions

Linear regression model $Y = X\beta + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$ and σ^2 unknown.

- (a.) Under H_0 (a model constraint): we have $C\beta = h$, i.e., a linear map $\beta \in \Re^p \to h \in \Re^q$ and $r(C) = q \leq p$, then the dimension of the kernel is $p - q$. So, $RSS_0(\beta)$ has the following expression:

$$RSS_0(\beta) = \parallel Y - X\beta \parallel_2{}^2 = (Y - X\beta)^T(Y - X\beta) + \lambda^T(C\beta - h).$$

  We have $\hat{\beta} = (X^TX)^{-1}(X^TY - C^T\lambda)$, $\lambda = (C(X^TX)^{-1}C^T)^{-1}(C(X^TX)^{-1}X^TY - h)$.

$$\hat{\beta}_j = \begin{cases} 0, & j = 1, \cdots, q \\ (X^T{}_jX_j)^{-1}X^T{}_jY, & j = q+1, \cdots, p \end{cases}$$

- (b.) Under H_1, we don't assume that a subset of the covariates have zero regression coefficients, i.e., we have the full model. By the properties: $\hat{Y} = X\hat{\beta} = X(X^TX)^{-1}X^TY = P_XY$ and $P_XX_j = X_j,\ j = 1, \cdots, p$. We have

$$Y - \hat{Y} = (1_n - P_X)Y = (1_n - P_X)(X\beta + \epsilon) = (1_n - P_X)\epsilon,$$

  then,

$$RSS_1 = \parallel Y - \hat{Y} \parallel_2{}^2 = \epsilon^T(1_n - P_X)\epsilon \sim \chi^2(n-p)\sigma^2,$$

  Similarly, under H_0, the constrained model has reduced q degree of projection matrix P_X, i.e.,

$$RSS_0 = \parallel Y - \hat{Y} \parallel_2{}^2 = Y^T(1_n - P_X)Y \sim \chi^2(n - (p-q))\sigma^2,$$

  and these q covariates are unrelated to the remaining variables. Further, we have

$$(RSS_0 - RSS_1)/\sigma^2 \sim \chi^2(n - (p-q) - (n-p)) = \chi^2(q).$$

  which is independent of $RSS_1/\sigma^2 \sim \chi^2(n-p)$.($SS_{full}$ and $SS_{res} - SS_{full}$ are independent R.V.s)

- (c.) By (b.), under H_0, we known that $RSS_1/\sigma^2 \sim \chi^2(n-p)$, $(RSS_0 - RSS_1)/\sigma^2 \sim \chi^2(q)$ and they are independent. So, the null hypothesis asserts a F-statistic:

$$\frac{(RSS_0 - RSS_1)/\sigma^2}{q} / \frac{RSS_1/\sigma^2}{n-p} = \frac{(RSS_0 - RSS_1)/q}{RSS_1/n-p} \sim F(q, n-p).$$

3. Garrote Solution when X is orthonormal and initial $\hat{\beta}_{ols}$

Proof: Set the initial least-square estimate (OLS) of the regression coefficients $\tilde{\beta} \in \Re^p$ and solve the optimization problem:

$$\hat{c} = \arg\min_{c \in \Re^p} \{ \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} c_j x_{ij} \tilde{\beta}_j)^2 \}, s.t., c \succeq 0, \| c \|_1 \leq t.$$

which is equivalent to solve the Lagrangian form:

$$\min_{c \in \Re^p} \frac{1}{2N} \{ \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} c_j x_{ij} \tilde{\beta}_j)^2 \} + \lambda \| c \|_1 = \min_{c \in \Re^p} f(c), \; c \succeq 0 \; and \; \lambda \geq 0. \tag{1}$$

Differentiating $f(c)$ w.r.t. c and setting the gradient vector to zero, we obtain equation:

$$-\frac{1}{N} \{ \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} c_j x_{ij} \tilde{\beta}_j) x_{ij} \tilde{\beta}_j \} + \lambda = 0, \tag{2}$$

For orthonormal case of X, we have

$$x_{ij} \cdot x_{ik} = \begin{cases} {x_{ij}}^2, & j = k \\ 0, & otherwise \end{cases} \tag{3}$$

So, Eq. 2 has explicit form:

$$-\frac{1}{N} \tilde{\beta}_j \langle X_j, Y \rangle + c_j {\tilde{\beta}_j}^2 \frac{1}{N} \sum_{i=1}^{N} x_{ij} + \lambda = 0. \tag{4}$$

Typically, we standardize the sample $(X_i, Y_i)_{i=1}^{N}$, i.e.,

$$\frac{1}{N} \sum_{i=1}^{N} x_{ij} = 0 \;, \quad \frac{1}{N} \sum_{i=1}^{N} y_i = 0, \quad \frac{1}{N} \sum_{i=1}^{N} x_{ij}^2 = 1. \tag{5}$$

And least-square method assures that $\frac{1}{N} \langle X_j, Y \rangle = \tilde{\beta}_j$ holds. Then, Eq. 4 can be replaced and solved by Eq. 6. This complete the proof.

$${\tilde{\beta}_j}^2 - \lambda = c_j {\tilde{\beta}_j}^2 \Rightarrow \; \hat{c}_j = (1 - \frac{\lambda}{{\tilde{\beta}_j}^2})_+. \tag{6}$$

□

4. Uniqueness of LASSO fitted values

- (a.) WTS: Every LASSO solution $\hat{\beta}$ gives the same fitted value $X\hat{\beta}$.
  Proof: Suppose that we have two solutions $\hat{\beta}^1$ and $\hat{\beta}^2$ with $X\hat{\beta}^1 \neq X\hat{\beta}^2$. For any $0 < \alpha < 1$, we know that $\alpha\hat{\beta}^1 + (1-\alpha)\hat{\beta}^2$ is also a solution for LASSO convex minimization problem. Set the common optimal value of LASSO solutions is c^*, then

$$\frac{1}{2N} \| Y - X(\alpha\hat{\beta}^1 + (1-\alpha)\hat{\beta}^2) \|_2{}^2 + \lambda \| \alpha\hat{\beta}^1 + (1-\alpha)\hat{\beta}^2 \|_1 < \alpha c^* + (1-\alpha)c^* = c^*. \quad (7)$$

  where the strict inequality is due to the strict convexity of the function $f(x) = \| y - x \|_2{}^2$ along with the convexity of the l_1-norm. The Eq. 7 means that$\alpha\hat{\beta}^1 + (1-\alpha)\hat{\beta}^2$ attains a lower criterion value than c^*, this contradicts our assumption $\hat{\beta}^1$ and $\hat{\beta}^2$ are LASSO solutions. □

- (b.) WTS: If $\lambda > 0$, every LASSO solution has same l_1-norm, i.e., $\| \hat{\beta}^1 \|_1$ = $\| \hat{\beta}^2 \|_1$.
  Proof: By (a.), any two LASSO solutions must have the same fitted value, i.e., the same squared error loss. Further, the solutions also attain the same value of the lasso criterion, and if $\lambda > 0$, then they must have the same l_1-norm. □

5. Computation of LASSO solution

Consider the LASSO problem:

$$\min_{\beta \in \Re^p} \frac{1}{2N} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \min_{\beta \in \Re^p} \frac{1}{2N} \| Y - X\beta \|_2{}^2 + \lambda \| \beta \|_1 = \min_{\beta \in \Re^p} f(\beta_\lambda). \quad (8)$$

- (a.) By using subgradient equation, we can differentiate $f(\beta_\lambda)$ w.r.t. β_λ and set the gradient vector to zero, then yields Eq. 9 which can be used to solve $\hat{\beta}$.

$$N^{-1}X^T(Y - X\beta) + \lambda S(\beta) = 0 \quad (9)$$

  where $S(\beta) = \begin{cases} sign(\beta), & \beta \neq 0 \\ [-1, 1], & \beta = 0 \end{cases}$. For the j_{th} component of $\hat{\beta}$, such as $\hat{\beta}_j = 0$, Eq. 9 solves that $\lambda > |N^{-1}X_j^T(Y - X\hat{\beta})|$, this shows Eq. 10 holds.

$$\begin{cases} \lambda = \quad N^{-1}X_j^T(Y - X\beta), & \hat{\beta}_j > 0 \\ \lambda = -N^{-1}X_j^T(Y - X\beta), & \hat{\beta}_j < 0 \\ \lambda > \ |N^{-1}X_j^T(Y - X\beta)|, & \hat{\beta}_j = 0 \end{cases} \quad (10)$$

- (b.) If $\lambda > \| N^{-1}X^TY \|_\infty = \max_{j=1,2,\cdots,p} |N^{-1}\langle X_j, Y\rangle|$, the results of (a.) have shown that $\hat{\beta}_j = 0, j = 1, 2, \cdots, p$, and the Uniqueness of LASSO fitted values assure that $\hat{\beta}_\lambda = 0$, i.e., $\lambda_{max} = \max_{j=1,2,\cdots,p} |N^{-1}\langle X_j, Y\rangle|$.

6. Degrees of freedom for LASSO in orthogonal design

The LASSO is a truly adaptive fitting, it is typically that the degrees of freedom is larger than K. However, LASSO not only selects predictors, but also shrinks their coefficients toward zero, this shrinkage turns out to be just the right amount to bring the degrees of freedom down to K. We will give this proof in the special case of an orthogonal design(Just as shown in Eq. 3).

Proof: Typically, the sample have been standardized (see Eq. 5), so the model can be denoted as

$$y_i = f(x_i) + \epsilon_i = \sum_{j=1}^{K} x_{ij}\beta_j + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)\ and\ \sigma^2\ unknown. \tag{11}$$

When design matrix X is orthogonal, LASSO problem (Eq. 8) has the subgradient form solution:

$$\hat{\beta}_j = S_\lambda(\frac{1}{N}\langle X_j, Y\rangle) \tag{12}$$

where $S_\lambda(x) = sign(x)(|x| - \lambda)_+$. Applying the orthonormal case, $\hat{\beta}_j$ can be denoted as:

$$\hat{\beta}_j = \begin{cases} \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} x_{ij}y_i - \lambda, & N^{-1}\langle X_j, Y\rangle\ > \lambda \\ 0, & |N^{-1}\langle X_j, Y\rangle| \leq \lambda \\ \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} x_{ij}y_i + \lambda, & N^{-1}\langle X_j, Y\rangle\ < -\lambda \end{cases} \tag{13}$$

Since $\hat{y}_i = \langle x_i, \beta\rangle = g(y_i)$, then $Cov(\hat{y}_i, y_i) = E(\hat{y}_i \cdot y_i) - E(\hat{y}_i) \cdot E(y_i) = E(g(y_i) \cdot (y_i - E(y_i)))$, the degrees of freedom $df(\hat{y})$ can be denoted as $\sum_{i=1}^{N} \sigma^{-2} E(g(y_i) \cdot (y_i - E(y_i)))$. *Stein's multivariate lemma* states that

$$df(\hat{y}) = \sum_{i=1}^{N} \sigma^{-2} E(g(y_i) \cdot (y_i - E(y_i))) = \sum_{i=1}^{N} E(\nabla g(y_i)).$$

Under the model assumption (see Eq. 11) and $\hat{\beta} = (\hat{\beta}_j)_{j=1}^{p} \neq 0$ (see Eq. 13), then $\frac{\partial \hat{y}_i}{\partial \hat{\beta}_j} = \sum_{j=1}^{K} x_{ij}$ and $\frac{\partial \hat{\beta}_j}{\partial y_i} = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$. So, we can calculate the LASSO degree of freedom $df(\hat{y})$ under orthogonal case:

$$df(\hat{y}) = \sum_{i=1}^{N} E(\nabla \hat{y}_i) = \sum_{i=1}^{N}\sum_{j=1}^{K} x_{ij}\frac{1}{N}\sum_{i=1}^{N} x_{ij} = \sum_{j=1}^{K}(\frac{1}{N}\sum_{i=1}^{N} {x_{ij}}^2) = K.$$

□

7. Robust regression and outliers constrained

Consider model:

$$y_i = f(x_i) + \gamma_i + \epsilon_i = \sum_{j=1}^{p} x_{ij}\beta_j + \gamma_i + \epsilon_i, \tag{14}$$

where $\epsilon_i \sim N(0, \sigma^2)$ and σ^2, γ_i are both unknown constant. Then the penalty term effectively limits the number of outliers has optimization problem:

$$\min_{\beta\in\Re^p, \gamma\in\Re^N} \frac{1}{2}\sum_{i=1}^{N}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j - \gamma_i)^2 + \lambda\sum_{i=1}^{N}|\gamma_i| \tag{15}$$

- (a.) WTP: Eq. 15 is jointly convex in β and γ.
  Proof: For any $0 < \alpha_1, \alpha_2 < 1$,

$$\begin{aligned} f(\alpha_1\beta + (1-\alpha_1)\beta, \alpha_2\gamma + (1-\alpha_2)\gamma) &= \| Y - \alpha_2\gamma + (1-\alpha_2)\gamma - X(\alpha_1\beta + (1-\alpha_1)\beta) \|_2{}^2 \\ +\lambda \| \alpha_2\gamma + (1-\alpha_2)\gamma \|_1 & \\ &\leq \| Y - \alpha_2\gamma - X\alpha_1\beta \|_2{}^2 + \lambda \| \alpha_2\gamma \|_1 \\ &+ \| Y - (1-\alpha_2)\gamma + (1-\alpha_1)\beta \|_2{}^2 + \lambda \| (1-\alpha_2)\gamma \|_1 \\ &= f(\alpha_1\beta, \alpha_2\gamma) + f((1-\alpha_1)\beta, (1-\alpha_2)\gamma) \end{aligned}$$

  This complete the proof. □

- (b.) WTP: Eq. 15 has same β solution with Huber's M-estimation.
  Proof: To solve Eq. 15, we need to define the outlier i which will be penalized under l_1-norm. So, we taken i_0, for each $i \geq i_0$, γ_i allows y_i to be an outlier. Then Eq. 15 can be replaced by:

$$\min_{\beta\in\Re^p, \gamma\in\Re^N} \frac{1}{2}\sum_{i=1}^{N}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j - \gamma_i)^2 + \lambda\sum_{i=i_0}^{N}|\gamma_i| = \min_{\beta\in\Re^p, \gamma\in\Re^N} f(\beta, \gamma) \tag{16}$$

  For a fixed value of β, if $i \geq i_0$, Eq.16 can be solved by subgradient equation(see Eq.12), i.e., the criterion $f(\beta, \gamma)$ is minimum at

$$\hat{\gamma}_i(\beta) = \begin{cases} y_i - X_j{}^T\beta, & if\ i < i_0 \\ sign(y_i - X_j{}^T\beta)(|y_i - X_j{}^T\beta| - \lambda)_+, & if\ i \geq i_0 \end{cases}$$

  Therefore, finding $\hat{\beta}$, solution to Eq. 15, amounts in finding $\hat{\beta}$ minimizing the criterion $f(\beta, \hat{\gamma}(\beta))$. Now, we denotes $I = \{i = i_0, \cdots, n,\ |y_i - \sum_{j=1}^{p} x_{ij}\beta_j| < \lambda\}$ is the outlier index, then $f(\beta, \hat{\gamma}(\beta))$ can be expressed by:

$$f(\beta, \hat{\gamma}(\beta)) = \frac{1}{2}\sum_{I}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \frac{1}{2}\sum_{I^c}\lambda^2 + \lambda\sum_{I^c}(|y_i - \sum_{j=1}^{p} x_{ij}\beta_j| - \lambda) \tag{17}$$

  which is same as Huber's M-estimation problem of β. □

8. Out-of-sample ${R_{os}}^2$

Out-of-sample ${R_{os}}^2$ is used to evaluate the prediction accuracy based on test data. Compared with the usual R^2, computed on residuals and is in-sample quantities, ${R_{os}}^2$ maintain the idea of usual R^2 but replace RSS by the out of sample MSE of the model under analysis (MSE_m). And in place of TSS is used the out of sample MSE of one benchmark model (MSE_{bmk}). The validation data of *CV* procedure is the out-of-sample data excepting for *hyperparameter tuning*.

$${R_{os}}^2 = \frac{MSE_m}{MSE_{bmk}} = 1 - \frac{\sum_{i \in T}(y_i - \hat{y}_i^{pred})^2}{\sum_{i \in T}(y_i - \bar{y}_i^{train})^2}$$

- (a.) 0.5051
  $price \sim\ bedrooms + bathrooms + sqft\ living + sqft\ lot$

- (b.) 0.5328
  $price \sim\ bedrooms + bathrooms + sqft\ living + sqft\ lot + bedrooms * bathrooms + bathrooms * sqft\ living + bathrooms * sqft\ lot + sqft\ living * sqft\ lot$

- (c.) *Kernel ridge regression.* In general, R package CVST and DRR are useful in dealing *kernel ridge regression.* The results are guided by *KRR.r* created by *Mlgruby* (Although I haven't got the results). Define the kernel K? See He!

- (d.) 0.7615
  $price \sim\ bedrooms + bathrooms + sqft\ living + sqft\ lot + zipcode + bedrooms * bathrooms + bathrooms * sqft\ living + bathrooms * sqft\ lot + sqft\ living * sqft\ lot$

- (e.) 0.7801

- (f.) This is due to more information has been included in calculation, such as the interaction terms, factor zipcode and penalty, so it help us to enhance the accuracy of model.