ORF 525: Statistical Foundations of Data Science

Jianqing Fan — Frederick L. Moore'18 Professor of Finance

Midterm Solutions

1. (a) The zipcode with the highest prices: 98004; The zipcode with the lowest prices: 98002.

   (b) Zipcode 98005 prices are 81.42% higher than 98001. The boundaries of the confidence interval should be $0.8142 \pm 0.02432 \cdot t_{15055}(1 - 0.05/2)$; By calculation, we obtain that the confidence interval is $[0767, 0.862]$.

   (c) Sample size $= 15055 + 73 + 1 = 15129$; Number of zipcodes $= 73 - 4 = 69$.

   (d) The formula for expected price is

$$\begin{aligned} \mathrm{E}\log(\text{price}_i) = 11.86 &+ 0.2712 + 0.0398 \times \text{bathrooms}_i \\ &- 0.02501 \times \text{bedrooms}_i + 3.286 \times 10^{-4} \times \text{sqft\_living}_i \\ &+ 7.011 \times 10^{-7} \times \text{sqft\_lot}_i. \end{aligned}$$

   (e) Recall that the residual standard error $= \sqrt{\frac{RSS}{DF}}$, where DF denotes the effective degrees of freedom. Using the above formula, we obtain that

$$\begin{aligned} RSS_{original} &= 15055 \times 0.2181^2 = 716.13, \\ RSS_{new} &= (15055 - 8) \times 0.2179^2 = 714.44. \end{aligned}$$

   Therefore the F-statistic should be

$$F = \frac{\frac{1}{p_{new}-p}\left[\sum_{i=1}^n (y_i - \widehat{y}_i)^2 - \sum_{i=1}^n (y_i - \widehat{y}_{i,new})^2\right]}{\frac{1}{n-p-1}\sum_{i=1}^n (y_i - \widehat{y}_i)^2} = \frac{1/8 \cdot (716.13 - 714.44)}{.2181} = 4.449.$$

   (f) Create 5 additional variables as follows:

$$X^{(k)} = (\text{sqft\_living} - q_i)_+^2, \quad k \in \{0, 1, 2, 3, 4\},$$

   where q_i is the $(20 \cdot i)$-th percentile `sqft_living`.

2. (a) We can solve the following minimization problem

$$\widehat{\boldsymbol{\beta}}_{\tau,\lambda} = \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^d} \mathcal{L}_\tau(\boldsymbol{\beta}) + \sum_{j=1}^p p_\lambda(|\beta_j|),$$

   where $\mathcal{L}_\tau(\boldsymbol{\beta}) := \frac{1}{n}\sum_{i=1}^n l_\tau(y_i - \mathbf{x}_i^T\boldsymbol{\beta})$, and

$$l_\tau(x) = \begin{cases} x^2/2, & \text{if } |x| \le \tau, \\ \tau|x| - \tau^2/2 & \text{if } |x| > \tau. \end{cases}$$

   In practice, the tuning parameters can be chosen by cross validation.

   (b) Note that $\sigma_{ij} = EX_iX_j$. Therefore, its Winsorized mean is given by

$$\tilde{\sigma}_{ij} = n^{-1}\sum_{k=1}^n \operatorname{sgn}(X_{ki}X_{kj})\min(|X_{ki}X_{kj}|, \tau_{ij})$$

   where τ_{ij} is some appropriately chosen threshold.

(c) The approximated function is

$$\log \frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \sum_{j=1}^{d}\sum_{l=1}^{L} \beta_{jl}\phi_l(x_j) + \sum_{k<j}\sum_{l=1}^{L}\sum_{m=1}^{L} \beta_{j,k,l,m}\phi_l(x_j)\phi_m(x_k).$$

Thinking basis and term interaction terms as the newly created $p = dL + \frac{1}{2}d(d-1)L^2$ variables, run penalized logistic regression with Lasso penalty to fit the model.

3. (a) For $\mathbf{x} = \sum_{j=1}^{p} \beta_j \boldsymbol{\xi}_j$ and $\mathbf{y} = \sum_{j=1}^{p} \beta'_j \boldsymbol{\xi}_j$, the inner product is defined as

$$< \mathbf{x}, \mathbf{y} >_{\mathcal{H}_K} = \sum_{j=1}^{p} \beta_j \beta'_j / \lambda_j.$$

Using the hint, for the vector $\mathbf{K}(\cdot, i)$, its j^{th} coefficient in the Hilbert space is $\beta'_j = \lambda_j \boldsymbol{\xi}_j^T \mathbf{e}_i$. Hence,

$$< \mathbf{K}(\cdot, i), \mathbf{x} >_{\mathcal{H}_K} = \sum_{j=1}^{p} \beta_j \lambda_j \boldsymbol{\xi}_j^T \mathbf{e}_i / \lambda_j = \mathbf{x}^T \mathbf{e}_i = x_i.$$

(b) Let $Z_i = x_{ij}\epsilon_i = x_{ij}[b'(\mathbf{x}_i^\top \boldsymbol{\beta}^*) - Y_i]$. Since $b'(\mathbf{x}_i^\top \boldsymbol{\beta}^*) = EY_i \in (0,1)$ and $Y_i \in \{0,1\}$, we have $|Z_i| \le |x_{ij}|$. According to Hoeffding's Inequality,

$$P(|\frac{1}{n}\sum_{i=1}^{n} x_{ij}\epsilon_i| \ge \frac{t}{\sqrt{n}}) = P(|\sum_{i=1}^{n} x_{ij}\epsilon_i| \ge \sqrt{n}t) \le 2e^{-\frac{2nt^2}{\sum_{i=1}^{n}(2x_{ij})^2}} = 2e^{-\frac{nt^2}{2\sum_{i=1}^{n} x_{ij}^2}}.$$

(c) According to the definition of ℓ_n, we can easily obtain that

$$\nabla \ell_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}[b'(\mathbf{x}_i^\top \boldsymbol{\beta}) - Y_i]\mathbf{x}_i$$

Thus, $[\nabla \ell_n(\boldsymbol{\beta})]_j = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i x_{ij}$. Using the results from (c) and given that columns of $\mathbf{x}_j$ are standardized, we obtain that

$$P(|[\nabla \ell_n(\boldsymbol{\beta})]_j| > \frac{t}{\sqrt{n}}) \le 2e^{-\frac{nt^2}{2\sum_{i=1}^{n} x_{ij}^2}} = 2e^{-t^2/2}.$$

Applying a union bound, we further obtain that

$$P(\|\nabla \ell_n(\boldsymbol{\beta})\|_\infty > \frac{t}{\sqrt{n}}) \le \sum_{j=1}^{p} P(|[\nabla \ell_n(\boldsymbol{\beta})]_j| > \frac{t}{\sqrt{n}}) \le 2pe^{-t^2/2}.$$

The proof is now complete by letting $t = a\sqrt{\log p}$.