ORF 525: Statistical Foundations of Data Science

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

Problem Set #1 Spring 2021

Due Friday, February 12, 2021.

1. Consider the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{W})$ with known positive definite matrix $\mathbf{W}$, and $\mathbf{X}$ is of full rank.

   (a) Show that the general least-squares estimator, which minimizes $(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{W}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$, is the best linear unbiased estimator.

   (b) Give explicitly the least-squares estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$.

   (c) If $\mathbf{W} = (1-\rho)\mathbf{I} + \rho\mathbf{1}\mathbf{1}^T$ is a equi-correlation matrix, what is the least-square estimate $\widehat{\boldsymbol{\beta}}$?

   Hint: Use Sherman-Morrison formula if necessary.

Answers:

(a) Let's denote by $\widehat{\boldsymbol{\beta}}$ the general least-squares estimator of $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{W}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Multiplying $\mathbf{W}^{-\frac{1}{2}}$ on both sides of the linear model, and by denoting

$$\tilde{\mathbf{X}} = \mathbf{W}^{-\frac{1}{2}}\mathbf{X}, \quad \tilde{\mathbf{y}} = \mathbf{W}^{-\frac{1}{2}}\mathbf{y}, \text{ and } \quad \tilde{\boldsymbol{\varepsilon}} = \mathbf{W}^{-\frac{1}{2}}\boldsymbol{\varepsilon},$$

we can have $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}$ where $\tilde{\boldsymbol{\varepsilon}} \sim N(0, \sigma^2\mathbf{I})$ and $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2$. The result then follows from the course.

(b) We clearly have

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2 = \left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^\top\tilde{\mathbf{y}} = \left(\mathbf{X}^\top\mathbf{W}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{W}^{-1}\mathbf{y}$$

and

$$\widehat{\sigma}^2 = \frac{\|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\widehat{\boldsymbol{\beta}}\|^2}{n-p} = \frac{\|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^\top\tilde{\mathbf{y}}\|^2}{n-p} = \frac{\|\tilde{\mathbf{y}} - \mathbf{P}\tilde{\mathbf{y}}\|^2}{n-p},$$

where $\mathbf{P} = \tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^\top = \mathbf{W}^{-1/2}\mathbf{X}\left(\mathbf{X}^\top\mathbf{W}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{W}^{-1/2}$ is a projector of rank p. Then, $\widehat{\sigma}^2$ can be rewritten as

$$\widehat{\sigma}^2 = \frac{\tilde{\mathbf{y}}^\top(\mathbf{I}-\mathbf{P})^2\tilde{\mathbf{y}}}{n-p} = \frac{\tilde{\mathbf{y}}^\top(\mathbf{I}-\mathbf{P})\tilde{\mathbf{y}}}{n-p} = \frac{\mathbf{y}^\top\mathbf{W}^{-1/2}(\mathbf{I}-\mathbf{P})\mathbf{W}^{-1/2}\mathbf{y}}{n-p}.$$

(c) Suppose that $\mathbf{W} \in \mathbb{R}^{n\times n}$. If $\mathbf{W}$ is the equi-correlation matrix with correlation ρ, namely

$$\mathbf{W} = (1-\rho)\mathbf{I} + \rho\mathbb{1}\mathbb{1}^T$$

where $\mathbb{1}$ stands for a vector in $\mathbb{R}^n$ with all elements equal to 1. Then, since $\rho \neq 1$ (as $\mathbf{W}$ is positive definite), by Sherman-Morrison formula, we have

$$\mathbf{W}^{-1} = a\mathbf{I} + b\mathbb{1}\mathbb{1}^T, \quad a = \frac{1}{1-\rho}, b = \frac{1}{1-\rho}\frac{-\rho}{1+(n-1)\rho}.$$

We denote by

$$\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T, \text{ and } \quad \alpha = \mathbb{1}^T\mathbf{P}\mathbb{1},$$

then

$$\begin{aligned}
& (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1} \\
= \; & \left(\mathbf{X}^T(a\mathbf{I}+b\mathbb{1}\mathbb{1}^T)\mathbf{X}\right)^{-1}\mathbf{X}^T(a\mathbf{I}+b\mathbb{1}\mathbb{1}^T) \\
= \; & \left(a\mathbf{X}^T\mathbf{X}+b(\mathbf{X}^T\mathbb{1})(\mathbf{X}^T\mathbb{1})^T\right)^{-1}(a\mathbf{X}^T+b\mathbf{X}^T\mathbb{1}\mathbb{1}^T) \\
= \; & \left(\frac{1}{a}(\mathbf{X}^T\mathbf{X})^{-1}-\frac{b}{a^2}\frac{\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{1}\right]\cdot\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{1}\right]^T}{1+\frac{b}{a}\mathbb{1}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{1}}\right)\cdot(a\mathbf{X}^T+b\mathbf{X}^T\mathbb{1}\mathbb{1}^T) \\
= \; & (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T+\frac{b}{a}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{1}\right]\cdot\mathbb{1}^T-\frac{\frac{b}{a}}{1+\frac{b}{a}\alpha}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{1}\right]\cdot\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{1}\right]^T\cdot\mathbf{X}^T \\
& -\frac{b}{a}\cdot\frac{\frac{b}{a}}{1+\frac{b}{a}\alpha}\cdot\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{1}\right]\cdot\alpha\cdot\mathbb{1}^T \\
= \; & (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T+\frac{b}{a}\left(1-\frac{\frac{b}{a}\alpha}{1+\frac{b}{a}\alpha}\right)\cdot\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{1}\right]\cdot\mathbb{1}^T-\frac{\frac{b}{a}}{1+\frac{b}{a}\alpha}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{1}\right]\cdot\mathbb{1}^T\cdot\mathbf{P} \\
= \; & (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\left(\mathbf{I}-\gamma\mathbb{1}\mathbb{1}^T(\mathbf{I}-\mathbf{P})\right)
\end{aligned}$$

where $\gamma = \dfrac{-b}{a+b\mathbb{1}^T\mathbf{P}\mathbb{1}}$. Hence, the least-squares estimator for $\boldsymbol{\beta}$ with an equi-correlation variance matrix is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\left(\mathbf{I}-\gamma\mathbb{1}\mathbb{1}^T(\mathbf{I}-\mathbf{P})\right)\mathbf{y}.$$

2. Consider the multiple regression model $Y_i = \mathbf{X}_i^T\boldsymbol{\beta}+\varepsilon_i$, where $\varepsilon_i \sim N(0,\sigma^2)$ for $i = 1,\cdots,n$.

   (a) Show that the maximum likelihood estimator is equivalent to the least-squares estimator, which finds $\boldsymbol{\beta}$ to minimize

   $$\sum_{i=1}^{n}(Y_i-\mathbf{X}_i^T\boldsymbol{\beta})^2$$

   and

   $$\widehat{\sigma} = \sqrt{\frac{\text{RSS}}{n}},$$

   where $\text{RSS} = \sum_{i=1}^{n}(Y_i-\mathbf{X}_i^T\widehat{\boldsymbol{\beta}})^2$ and $\widehat{\boldsymbol{\beta}}$ is the least-squares solution.

   (b) Show that $\text{RSS} \sim \sigma^2\chi^2_{n-p}$, where p is the rank of $\mathbf{X}$ (full rank for simplicity).

   (c) Prove that $1-\alpha$ CI for β_j is $\widehat{\beta}_j \pm t_{n-p}(1-\alpha/2)\sqrt{v_j\text{RSS}/(n-p)}$, where v_j is the j^{th} diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$.

   (d) Dropping the normality assumption, if $\{\mathbf{X}_i\}$ are i.i.d. from a population with $E\mathbf{X}\mathbf{X}^T = \boldsymbol{\Sigma}$ and independent of $\{\varepsilon_i\}_{i=1}^n$, which are i.i.d. from a population with $E\varepsilon = 0$ and $\text{var}(\varepsilon) = \sigma^2$, show that

   $$\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{d} N(0,\sigma^2\boldsymbol{\Sigma}^{-1}).$$

Answers:

(a) A straight forward calculation yields

$$\log(P(\mathbf{Y}|\boldsymbol{\beta},\sigma,\mathbf{X})) \propto -n\cdot\log(\sigma) - \frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2,$$

Thus, the MLE becomes (maximization and minimization are unique here):

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} P(\mathbf{Y}|\boldsymbol{\beta},\sigma,\mathbf{X}) = \arg\min_{\boldsymbol{\beta}} \|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2 = \widehat{\boldsymbol{\beta}}^{OLS}.$$

which is identical to the OLS estimator. And by taking the derivative w.r.t. σ, we can also have

$$\widehat{\sigma} = \sqrt{\frac{\|\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|^2}{n}}$$

(b) Let us denote $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}$, then $\mathbf{P}$ is an idempotent matrix so that $\mathbf{I}-\mathbf{P}$ and

$$Tr(\mathbf{I}-\mathbf{P}) = n - Tr((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}) = n - tr(\mathbf{I}_p) = n-p.$$

Since an idempotent matrix can only have eigenvalues 0 or 1, the spectral decomposition gives us

$$\mathbf{I}-\mathbf{P} = \mathbf{U}\mathbf{D}\mathbf{U}^T, \quad \mathbf{D} = \begin{bmatrix} \mathbf{I}_{n-p} & 0 \\ 0 & 0 \end{bmatrix}$$

for some orthogonal matrix $\mathbf{U}\in\mathbb{R}^{n\times n}$ (i.e. $\mathbf{U}^T\mathbf{U}=\mathbf{I}_n$). We then have

$$\|\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 = \mathbf{Y}^T(\mathbf{I}-\mathbf{P})\mathbf{Y} = (\mathbf{U}^T\mathbf{Y})^T\mathbf{D}(\mathbf{U}^T\mathbf{Y}) = \sum_{i=1}^{n-p}(\mathbf{U}^T\mathbf{Y})_i^2$$

We know that $\mathbf{Y}\sim N(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}_n)$, and thus $\mathbf{U}^T\mathbf{Y}\sim N(\mathbf{U}^T\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}_n)$ which implies that the components $\{(\mathbf{U}^T\mathbf{Y})_i\}_{i=1}^n$ are independent with each other. Hence, the quantity $\|\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|^2$ is a sum of $(n-p)$ iid normal distributed random variable $N(0,\sigma^2)$. By definition

$$RSS = n\widehat{\sigma}^2 = \|\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 \sim \sigma^2\chi^2_{n-p}.$$

(c) It is easy to show that $\widehat{\boldsymbol{\beta}}\sim N(\boldsymbol{\beta},\sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$. So, for any vector $u\in\mathbb{R}^d$ where d is the dimension of vector X_i, we must have

$$\frac{u^T(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\sigma\cdot\sqrt{u^T(\mathbf{X}^T\mathbf{X})^{-1}u}} \sim N(0,1)$$

and thus by independence of $\widehat{\beta}$ and $\widehat{\sigma}^2$, we derive

$$\frac{u^T(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\sigma\cdot\sqrt{u^T(\mathbf{X}^T\mathbf{X})^{-1}u}}\cdot\left(\frac{n\widehat{\sigma}^2}{\sigma^2}\frac{1}{n-p}\right)^{-\frac{1}{2}} \sim t_{n-p}.$$

If we choose u that has all zeros but 1 on the jth coordinate, then we get

$$\boldsymbol{\beta}_j \sim \widehat{\boldsymbol{\beta}}_j - \sqrt{\frac{n\widehat{\sigma}^2}{n-p}\cdot(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}\cdot t_{n-p}.$$

Then the result for CI follows.

(d) We can write

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = (\frac{1}{n}\mathbf{X}^T\mathbf{X})^{-1} \cdot \frac{1}{\sqrt{n}}\mathbf{X}^T\boldsymbol{\varepsilon}.$$

We also observe by LLN that

$$\frac{1}{n}\mathbf{X}^T\mathbf{X} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T \longrightarrow \mathbb{E}[\mathbf{X}_1\mathbf{X}_1^T] = \boldsymbol{\Sigma} \quad \text{in probability,}$$

Moreover, by CLT, we have a convergence in distribution for

$$\frac{1}{\sqrt{n}}\mathbf{X}^T\boldsymbol{\varepsilon} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i\epsilon_i \longrightarrow N(0, Var(\mathbf{X}_1\epsilon_1)) = N(0, \mathbb{E}[\epsilon_1^2\mathbf{X}_1\mathbf{X}_1^T]) = N(0, \sigma^2\boldsymbol{\Sigma})$$

where the last equality comes from $\mathbf{X}_1$ and ϵ_1 are independent and $\epsilon_1 \in \mathbb{R}$. By Slutsky's theorem and the continuous mapping theorem, we can drive

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \boldsymbol{\Sigma}^{-1}N(0, \sigma^2\boldsymbol{\Sigma}) = N(0, \sigma^2\boldsymbol{\Sigma}^{-1}).$$

3. Let us consider the 129 macroecnomic time series again as described in the lecture notes http://orfe.princeton.edu/%7Ejqfan/fan/classes/245/chap11.pdf
Let $Y_t = \Delta\log(\text{PCE}_t)$ be the changes in the personal consumption expenditure. Let us take

$$X_{t,1} = \text{Unrate}_{t-1},\ X_{t,2} = \Delta\log(\text{IndPro}_{t-1}),\ X_{t,3} = \Delta\log(\text{M2Real}_{t-1}),$$
$$X_{t,4} = \Delta\log(\text{CPI}_{t-1}),\quad X_{t,5} = \Delta\log(\text{SPY}_{t-1}),\quad X_{t,6} = \text{HouSta}_{t-1},\quad X_{t,7} = \text{FedFund}_{t-1}$$

Let us again take the last 10 years data as testing set and remaining as trainng set. Conduct a similar analysis as those in the lecture notes. Answer in particular the following questions.

(a) What are $\widehat{\sigma}^2$, adjust R^2 and insignificant variables?

(b) Now perform the stepwise deletion, eliminating one least significant variable at a time (by iteratively looking at the smallest $|t|$-statistic) until all variables are statistically significant. Let us call this model as model $\widehat{\mathcal{M}}$. (For the stepwise deletion by AIC, the function `step` can do the job automatically)

(c) Using model $\widehat{\mathcal{M}}$, what are root mean-square prediction error and mean absolute deviation prediction error for the test sample?

Note: It is possible that the resulting model is not as good as the vanilla least squares; It's also possible that the resulting out of sample R^2 is negative, since predicting the difference of log PCE is not an easy task in general. If we take the predictors at time t, then it is the association study and the prediction is somewhat better.

(d) Compute the standardized residuals based on model $\widehat{\mathcal{M}}$. Present the time series plot of the residuals, fitted values versus the standardized residuals, and QQ plot for the standardized residuals.

Answers: cf R codes.

4. Zillow is an online real estate database company that was founded in 2006. The most important task for Zillow is to predict the house price. However, their accuracy has been criticized a lot. According to Fortune, "Zillow has Zestimated the value of 57 percent of U.S. housing stock, but only 65 percent of that could be considered 'accurate' – by its definition, within 10 percent of the actual selling price. And even that accuracy isn't equally distributed". Therefore, Zillow needs your help to build a housing pricing model to improve their accuracy. Download and read the data (traing data: 15129 cases, testing data: 6484 cases)

```
train.data <- read.csv('train.data.csv', header=TRUE)
test.data <- read.csv('test.data.csv', header=TRUE)
train.data$zipcode <- as.factor(train.data$zipcode)
test.data$zipcode <- as.factor(test.data$zipcode)
```

where the last two lines make sure that zip code is treated as factor. Letting $\mathcal{T}$ as a test set, define out-of-sample R^2 as of a prediction method $\{\widehat{y}_i^{pred}\}$ as

$$R^2 = 1 - \frac{\sum_{i\in\mathcal{T}}(y_i - \widehat{y}_i^{pred})^2}{\sum_{i\in\mathcal{T}}(y_i - \bar{y}^{pred})^2},$$

where $\bar{y}^{pred} = \text{ave}(\{y_i\}_{i\in\mathcal{T}_0})$ and $\mathcal{T}_0$ is the training set.

(a) Calculate out-of-sample R^2 using variables "bathrooms", "bedrooms", "sqft_living", and "sqft_lot".

(b) Calculate out-of-sample R^2 using the 4 variables above along with interaction terms.

(c) Compare the result with the nonparametric model using Gaussian kernel (standardize predictors first) with $\gamma = 0.1^2/2$ and $\lambda = 0.1$.

(d) Add the factor zipcode to (b) and compute out-of-sample R^2.

(e) Add the following additional variables to (d): $X_{12} = I(view == 0)$, $X_{13} = L^2$, $X_{13+i} = (L - \tau_i)_+^2$, $i = 1, \cdots, 9$, where τ_i is $10 * i^{th}$ percentile and L is the size of living area ("sqft_living"). Compute out-of-sample R^2.

Answers: cf R codes.

5. Prove the representer theorem in the lecture note (Theorem 1.4). You are allowed to consult the book, but not allow to have verbatim copy. You need to write the solution of your own with at least some changes of notation.

Answer: cf Theorem 2.6 in the book.