ORF 525: Statistical Foundations of Data Science

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

Problem Set #1 Spring 2024

Due Friday, February 9, 2024.

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.

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

   Note: The proof for this problem is standard. You are welcome to look at classical book for multiple regression and produce the solution in your own language.

3. Let us consider the 129 macroecnomic time series as described in the lecture notes

   https://fan.princeton.edu/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 training set. Conduct a similar analysis as those in the lecture notes. Answer in particular the following questions.

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

(b) Now perform the stepwise deletion, eliminating one least significant variable at a time (by looking at the smallest $|t|$-statistic) until all variables are statistically significant. Let us call this model as model $\widehat{\mathcal{M}}$. What is $\widehat{\mathcal{M}}$? (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 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.

4. Zillow is an online real estate database company that was founded in 2006. An 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 the training and testing data:

https://fan.princeton.edu/fan/classes/525/DataSets/train.data.csv
https://fan.princeton.edu/fan/classes/525/DataSets/test.data.csv

and read the data (training 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 $\bar{y}^{pred} = \mathrm{ave}(\{y_i\}_{i\in\mathcal{T}_0})$ 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.

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