ORF 405: Regression and Time Series

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

Problem Set #1 Fall 2026

Due Saturday, September 12, 2026.

Instructions. Submit one PDF file. Show your work for mathematical problems. For the empirical problem, include concise code, relevant figures or tables, and a short interpretation.

1. Let $(\mathbf{X}, Y)$ denote a generic observation and let $f^*(\mathbf{x}) = \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$.

   (a) Let $\widehat{f}$ be an estimator constructed from an independent training sample, and fix a testing point $\mathbf{x}_0$. Assume $Y_0 = f^*(\mathbf{x}_0) + \varepsilon_0$, where $E\varepsilon_0 = 0$ and $\mathrm{var}(\varepsilon_0) = \sigma^2(\mathbf{x}_0)$. Show the bias-variance decomposition of the mean-square prediction error:

   $$\mathbb{E}\{Y_0 - \widehat{f}(\mathbf{x}_0)\}^2 = \sigma^2(\mathbf{x}_0) + \{\mathbb{E}\widehat{f}(\mathbf{x}_0) - f^*(\mathbf{x}_0)\}^2 + \mathrm{var}\{\widehat{f}(\mathbf{x}_0)\}.$$

   (b) Explain briefly why flexible regression methods may have small bias but large variance.

   (c) Define the population best linear predictor by

   $$\boldsymbol{\beta}^* = \mathrm{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} E\{Y - \mathbf{X}^T\boldsymbol{\beta}\}^2.$$

   Assuming $\boldsymbol{\Sigma} = \mathbb{E}(\mathbf{X}\mathbf{X}^T)$ is nonsingular, show that $\boldsymbol{\beta}^* = \boldsymbol{\Sigma}^{-1}\mathbb{E}(\mathbf{X}Y)$ and that $\boldsymbol{\beta}^* = \mathrm{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} E\{f^*(\mathbf{X}) - \mathbf{X}^T\boldsymbol{\beta}\}^2$. Thus, linear regression is the best linear approximation in mean squared prediction error and in approximating the true regression function f^*.

   (d) Let $(\mathbf{X}_i, Y_i)$, $i = 1, \cdots, n$ be i.i.d. observations from the general regression problem. Derive directly, without using matrix notation, the least square estimator that minimizes $n^{-1}\sum_{i=1}^n (Y_i - \mathbf{X}_i^T\boldsymbol{\beta})^2$ is given by

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

   Note that this least-squares estimator converges to $\boldsymbol{\beta}^*$ defined in problem 1(c), as shown in 1(e), even if the true regression is nonlinear.

   (e) By the law of large numbers, it holds under the mild conditions

   $$n^{-1}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^T \xrightarrow{P} \boldsymbol{\Sigma}, \qquad n^{-1}\sum_{i=1}^n \mathbf{X}_i Y_i \xrightarrow{P} \mathbb{E}\mathbf{X}Y,$$

   where $\xrightarrow{P}$ means convergence in probability as $n \to \infty$. Use this to show that the bias of the linear predictor $\widehat{f}(\mathbf{x}_0) = \widehat{\boldsymbol{\beta}}^T\mathbf{x}_0$ is approximately $f^*(\mathbf{x}_0) - (\boldsymbol{\beta}^*)^T\mathbf{x}_0$.

2. Consider the linear model

   $$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad E\boldsymbol{\varepsilon} = \mathbf{0}, \qquad \mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n,$$

   where $\mathbf{X}$ is an $n \times p$ full-column-rank matrix.

(a) Show that the residual vector $\widehat{\boldsymbol{\varepsilon}} = \mathbf{y} - \widehat{\mathbf{y}}$ is orthogonal to every column of $\mathbf{X}$, namely the residual is uncorrelated with each covariate.

(b) Prove the Gauss-Markov Theorem.

(c) If $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, show that $\mathbb{E}\widehat{\sigma}^2 = \sigma^2$ and $\mathrm{var}(\widehat{\sigma}^2) = 2\sigma^4/(n-p)$.

(d) Consider the ANOVA problem. Let $\mathbf{P}_i = \mathbf{X}_i(\mathbf{X}_i^T\mathbf{X}_i)^{-1}\mathbf{X}_i^T$ be the projection matrices ($i = 1, 2$) with $\mathbf{X}_2$ has more columns than $\mathbf{X}_1$ (smaller model). Shows that the RSS reduction

$$\mathrm{RSS}_0 - \mathrm{RSS}_1 = \mathbf{Y}^T(\mathbf{P}_2 - \mathbf{P}_1)\mathbf{Y}^T \overset{under\ H_0}{=} \boldsymbol{\varepsilon}^T(\mathbf{P}_2 - \mathbf{P}_1)\boldsymbol{\varepsilon}^T$$

where $H_0 : \mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and that $(\mathbf{I}_n - \mathbf{P}_2)(\mathbf{P}_2 - \mathbf{P}_1) = 0$.

(e) If $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, show that $\mathrm{RSS}_0 - \mathrm{RSS}_1$ and RSS_1 are independent.

3. For observations (x_i, y_i), $i = 1, \cdots, n$, consider the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

(a) Show that

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \qquad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}.$$

(b) Show that $\mathrm{var}(\widehat{\beta}_1 | x_1, \cdots, x_n) = \sigma^2 / \sum_{i=1}^n (x_i - \bar{x})^2$ if $\mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$.

(c) Derive the above results in part (a) and (b) using the formulas given in the class with the matrix notation.

(d) Define $\mathrm{TSS} = \sum_i (y_i - \bar{y})^2$, $\mathrm{RSS} = \sum_i (y_i - \widehat{y}_i)^2$, and $\mathrm{ESS} = \sum_i (\widehat{y}_i - \bar{y})^2$. Prove that $\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}$. This is indeed a special case of problem 2(d).

4. **(Polynomial regression and cross-validation).** Generate data from

$$X_i \sim \mathrm{Uniform}[-2, 2], \qquad Y_i = 1 + 2X_i - X_i^2 + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2),$$

with $n = 200$.

(a) Fit polynomial regressions of degrees $1, 2, \cdots, 8$ with $\sigma = 1$.

(b) Use 5-fold cross-validation to choose the polynomial degree.

(c) Repeat the experiment for two noise levels: $\sigma = 0.5$ and $\sigma = 2$.

(d) Report the cross-validation errors in a table and plot the fitted curve for the selected model.

(e) Discuss how the selected degree changes with the noise level.

5. (**Prediction of Housing Prices**) 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 $\{\widehat{y}_i^{method}\}$ as

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

where $\bar{y}^{train} = \text{ave}(\{y_i\}_{i\in\mathcal{T}_0})$ 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) Add the factor zipcode to (b) and compute out-of-sample R^2.

(d) 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.

(e) Now without assessing to the test data, using the 5-fold cross-validation and the training data alone, compute the out-of-sample R^2 for the last model.