ORF350 (Spring 2017) Analysis of Big Data **Assignment: 2**

Maximum Likelihood and Regression

Lecturer: Han Liu *Email: hanliu@princeton.edu*

Note that the due date for this assignment is Wednesday March 8, 5:00pm. Please submit your hard copy at the ORF 350 dropbox in the Sherrerd Hall student lounge. Electronic submissions will not be accepted unless there is an extreme and compelling case.

Q1. Maximum Likelihood Estimator (MLE) and Asymptotic Normality (30 pts)

Maximum likelihood is one of the most fundamental principals in parameter estimation. Suppose we have n i.i.d. random samples $\{X_i\}_{i=1}^n$ that have probability density function $p_\theta(x)$. We are interested in estimating the parameter θ. Denote the correspondent MLE by $\widehat{\theta}_n$. In the lecture, we have known that under some regularity conditions, the MLE enjoys the asymptotic normality

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{D} N(0, \frac{1}{I(\theta)}), \tag{0.1}$$

where

$$I(\theta) := \mathbb{E}_\theta\Big(-\frac{\partial^2}{\partial\theta^2}\log p_\theta(X)\Big) = -\int_{\mathcal{X}}\Big(\frac{\partial^2}{\partial\theta^2}\log p_\theta(x)\Big)p_\theta(x)dx$$

is the Fisher information and $\mathcal{X}$ is the range of X_i.

Part I. Let z_α denote the α upper quantile of standard normal distribution. Prove that

$$C_n = \left[\widehat{\theta}_n - \frac{z_{\alpha/2}}{\sqrt{nI(\widehat{\theta}_n)}}, \widehat{\theta}_n + \frac{z_{\alpha/2}}{\sqrt{nI(\widehat{\theta}_n)}}\right]$$

is a $(1-\alpha)$ asymptotic confidence interval for θ, i.e., $\lim_{n\to\infty} P(\theta \in C_n) = 1-\alpha$. We assume $I(\theta)$ is a continuous function.
(Hint 1: According to Slutsky's Theorem, if $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} c$, where c is some constant, then $X_nY_n \xrightarrow{D} cX$. Use this result to prove that $\sqrt{nI(\widehat{\theta}_n)}(\widehat{\theta}_n - \theta) \xrightarrow{D} N(0,1)$.)

Part II. Suppose $\{X_i\}_{i=1}^n$ have the following probability density function:

$$p_\theta(x) = (\theta-1)x^{-\theta}\cdot \mathbb{1}\{x \geq 1\},$$

where $\theta > 1$ is the parameter of interest. Denote this distribution by P_θ and MLE of θ by $\widehat{\theta}_n$.

(a) Derive the MLE $\widehat{\theta}_n$.

(b) From (0.1), we know that the MLE $\widehat{\theta}_n$ is asymptotically normal. Calculate the asymptotic variance in terms of θ.

(c) Derive a 95% asymptotic confidence interval (CI) for the parameter θ.

Inverse transform sampling
inverse of cdf takes probabilities (sort of) and outputs x.
So you input the samples from runif(n, 0, 1)

(d) Use simulations to verify the effectiveness of the CI in (c) when $n = 100$ and $\theta = 2$.
(Hint 1: For the simulation part, we need to figure out how to generate random variables that follow P_θ. Suppose we have obtained the cumulative distribution function $F(x)$ of P_θ, and we generate a random variable U that is uniformly distributed over $[0, 1]$. It is true that $F^{-1}(U) \sim P_\theta$.)
(Hint 2: By effectiveness of the CI, we mean whether the constructed CI will cover the true θ with probability around 95%. To verify the effectiveness of the CI, we can independently generate a large number of datasets (e.g., 10,000 datasets) with each having the sample size n. Calculate CI's for all generated datasets respectively and then summarize the frequency of CI's covering the true θ. If the frequency is around 95%, then we can claim that the constructed CI is effective.)

Q2. Non-Existence of MLE (15 pts)

Suppose we have n i.i.d. data $X_i \sim Bernoulli(1/(1 + \log(\theta)))$, where $\theta > 1$. Show that if our observations are all ones or zeros, then the MLE does not exist.
(Hint 1: Try to derive the MLE. What happens?)

Q3. Exploratory Data Analysis (15 pts)

Read in the `housingprice.csv` file using `read.csv()` function.

(a) Rank the zipcodes by their average housing prices. What are the top 3 zipcodes whose average housing prices are most expensive? Create three boxplots of housing prices for these 3 zipcodes respectively.

(Hint: First convert the `zipcode` column into factors. Then use `tapply()` and `sort()` functionals to compute the result.)

(b) Visualize the relationship between `sqft_living` and `housing price` by creating a scatter plot.

Q4. A Simple Linear Model (20 pts)

This question continues from Question 3. Load the training data `train.data.csv` and testing data `test.data.csv`. We'll build our regression model on the training data and evaluate the model on the testing data.

(a) Build a linear model on the training data using lm() by regessing the housing price on these variables: `bedrooms`, `bathrooms`, `sqft_living`, and `sqft_lot`. What's the R^2 of the model on training data? What's the R^2 on testing data?

(b) Add `zipcode` in your linear model. What's the R^2 of the new model on the training data and testing data respectively?

(c) The image below is Bill Gates' house. Load the file `fancyhouse.csv` to obtain the features of the house. Guess the price of his house using your linear model. Do you think the predicted price is reasonable?

(d) Suppose we have a linear regression problem with n training samples and d covariates. If $n > d + 1$, show that adding another covariate in the model never hurts R^2 over the training data.

(Hint: By definition, the ordinary least squares (OLS) estimator $\widehat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta} \in R^d} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$, where $\boldsymbol{Y}$ is the response and $\boldsymbol{X}$ is the design matrix. Denote the new design matrix with the additional covariate by $\boldsymbol{X}_1$. Then

Figure 1: Image fom Wikipedia Commons

the OLS estimator for the new regression problem $\widehat{\boldsymbol{\beta}}_1 = \text{argmin}_{\boldsymbol{\beta} \in R^{d+1}} \|\boldsymbol{Y} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2$. Compare $\|\boldsymbol{Y} - \boldsymbol{X}_1\widehat{\boldsymbol{\beta}}_1\|_2^2$ and $\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2$).

Q5. Feature Engineering (20 pts)

Let's continue to improve the linear model we have. Instead of throwing only the raw data into the statistical model, we might want to use our intuition and domain expertise to extract more meaningful features from the raw data. This step is called feature engineering. Using meaningful features in the model is often crucial for successful data analysis.

(a) Add another variable by multiplying the number of bedrooms by the number of bathrooms, which describes the combined benefit of having more bedrooms and bathrooms. Add this variable to the linear model we have in Question 4 (b). What's the R^2 of the new model on the training data and testing data respectively?

(Hint: You don't have to create a new column in the data frame. Try this trick in `lm()`: `lm( y ~ x1 + x2 + x1 * x2, data = your.data))`

(b) Polynomial regression is a general technique that allows you to add nonlinear features in your statistical model. Based on the model we have in Question 4 (b), add polynomial terms of the bedrooms variable of degrees 2 and 3, and also bathroom variable of degrees 2 and 3 in your model. Find out the R^2 of the new model on training data and testing data.

(Hint: You don't have to create a new data frame. Try this trick in `lm()`. `lm( y ~ poly(x1, degree) + x2 + x3, data = your.data))`

Acknowledgement

The origin of the dataset `housingprice.csv` is from the Coursera open course Machine Learning Foundations: A Case Study Approach by Prof. Carlos Guestrin and Prof. Emily Fox. The open course also inspired the linear regression part of this assignment.