MIDTERM–2020

ORFE/FIN 525: Statistical Foundations of Data Science

April 1, 2020

Instructions: The exam is open book and open notes, and lasts 80 minutes exactly. Work independently in accordance with the honor codes. You may use any results in the notes, homework problems, and our textbook. You can also use any results in the subproblems before the subproblems that you intend to solve.

1. (40%) **Understanding Statistical Results**
   For the Zillow data, we use log-prices as the Y-variable so that the prediction errors reflect the percentages of errors. We fit the model

$$\log(\text{price}_i) = \mu + \alpha_{\text{zipcode}_i} + \beta_1 \text{bathrooms}_i + \beta_2 \text{bedrooms}_i + \beta_3 \text{sqft\_living}_i + \beta_4 \text{sqft\_lot}_i + \varepsilon_i$$

   and got the following results.

```
lm(formula = price ~ bathrooms + bedrooms + sqft_living + sqft_lot +  zipcode,
         data = train_data)

               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.186e+01  1.544e-02 768.226  < 2e-16 ***
bathrooms     3.980e-02  3.637e-03  10.945  < 2e-16 ***
bedrooms     -2.501e-02  2.379e-03 -10.516  < 2e-16 ***
sqft_living   3.286e-04  3.352e-06  98.048  < 2e-16 ***
sqft_lot      7.011e-07  5.031e-08  13.936  < 2e-16 ***
zipcode98002 -5.227e-02  2.260e-02  -2.313 0.020746 *
zipcode98003  4.303e-02  2.083e-02   2.065 0.038894 *
zipcode98004  1.198e+00  2.044e-02  58.622  < 2e-16 ***
zipcode98005  8.142e-01  2.432e-02  33.478  < 2e-16 ***
zipcode98006  7.377e-01  1.824e-02  40.437  < 2e-16 ***
zipcode98007  7.034e-01  2.702e-02  26.031  < 2e-16 ***
zipcode98008  7.148e-01  2.110e-02  33.875  < 2e-16 ***
zipcode98010  2.712e-01  2.954e-02   9.182  < 2e-16 ***
zipcode98011  4.581e-01  2.322e-02  19.731  < 2e-16 ***
zipcode98014  2.364e-01  2.824e-02   8.370  < 2e-16 ***
zipcode98019  3.129e-01  2.347e-02  13.335  < 2e-16 ***
zipcode98022  9.305e-02  2.257e-02   4.122 3.78e-05 ***
zipcode98023 -1.011e-02  1.803e-02  -0.561 0.575139
.........................
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2181 on 15055 degrees of freedom
```

```
Multiple R-squared:  0.8287,Adjusted R-squared:  0.8279
F-statistic:   998 on 73 and 15055 DF,  p-value: < 2.2e-16
```

In the above fitting, the baseline zip code is 98001 whose effect is setting as zero so that the reported coefficient in front of each zipcode reflects the difference (contrast) between their intercepts. For example, the coefficient -0.05227 can be understood as that zipcode 98002 house prices are on average 5.227% lower than those in zipcode 98001, after adjusting regression effect.

(a) Among the zip codes shown above, which areas have highest prices and which areas have the lowest prices?

(b) On average, how much zipcode 98005 sells higher than zipcode 98001? What is the 95% confidence interval for the estimate?

(c) What is the sample size used? How many zip codes (including those omitted in the above presentation) are used?

(d) Given a house at zipcode 98010, what is the formula for the expected house price? Write down the formula.

(e) We then run another regression model with 8 additional variables and obtain the residual standard error 0.2179. What are the RSS for the original model (null hypothesis) and the new model (alternative hypothesis)? Compute the F-test statistic for testing whether these 8 additional variables are statistically significant.

(f) If we would like to examine the possible nonlinear effect of the variable `sqft_living` with a quadratic spline basis and 4 knots, how would you create new variables.

2. (30%) **Understanding statistical methods**

(a) Suppose that we have n data points $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$. We wish to fit a sparse linear model $y_i = \mathbf{x}_i^T\boldsymbol{\beta}^* + \varepsilon_i$ robustly using the adaptive Huber loss and a folded concave penalty $p_\lambda(\cdot)$. Write down the criterion to be minimized and describe briefly how the tuning parameters to be chosen.

(b) Suppose that we would like to use the Winsorization (truncation) method to estimate robustly each element σ_{ij} of the covariance matrix of $\mathbf{X}$ with exponential concentration. Show your estimator. For simplicity, let us assume that $EX_i = EX_j = 0$ so that you do not need to estimate the mean.

(c) Suppose that we wish to estimate the bivariate nonparametric interaction models for logistic regression:

$$\log \frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \sum_{j=1}^d f_j(x_j) + \sum_{k<j} f_{j,k}(x_j, x_k)$$

and approximate the $f_j(x_j)$ and $f_{j,k}(x_j, x_k)$ by using the basis $\{\phi_\ell(x_j)\}_{\ell=1}^L$ and their bivariate tensor products. Write down the approximation. Describe briefly how would you use Lasso to fit the model.

3. (30%) **Understanding Statistical Theory**

(a) Let $\mathbf{K}$ be a $p \times p$ positive definite matrix, which admits the eigen-decomposition:

$$\mathbf{K} = \sum_{j=1}^p \lambda_j \boldsymbol{\xi}_j \boldsymbol{\xi}_j^T,$$

where $\lambda_j > 0$ is an eigenvalue of $\mathbf{K}$ and $\boldsymbol{\xi}_j$ is its associated eigenvector. Then any vector $\mathbf{x} \in \mathbb{R}^p$ can be written as $\mathbf{x} = \sum_{j=1}^p \beta_j \boldsymbol{\xi}_j$. Define the inner product $< \mathbf{x}, \mathbf{y} >_{\mathcal{H}_K}$ for the reproducing Hilbert space and verify the reproducibility property $< \mathbf{K}(\cdot, i), \mathbf{x} >_{\mathcal{H}_K} = x_i$, where $\mathbf{K}(\cdot, i)$ is the

i^{th} column of the matrix $\mathbf{K}$ and x_i is the i^{th} component of $\mathbf{x}$. **Hint**: $\mathbf{K}(\cdot, i) = \sum_{j=1}^{p} \lambda_j \boldsymbol{\xi}_j^T \mathbf{e}_i \boldsymbol{\xi}_j$ where $\mathbf{e}_i$ is the unit vector with 1 being at the i^{th} position.

(b) Consider the logistic regression and its penalized likelihood estimator

$$\widehat{\boldsymbol{\beta}} \in \text{argmin}_{\boldsymbol{\beta}} \{\ell_n(\boldsymbol{\beta}) + \lambda_n \|\boldsymbol{\beta}\|_1\},$$

where $\ell_n(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} [b(\mathbf{x}_i^T \boldsymbol{\beta}) - Y_i \mathbf{x}_i^T \boldsymbol{\beta}]$ is the negative log-likelihood. For $\varepsilon_i = b'(\mathbf{x}_i^T \boldsymbol{\beta}^*) - Y_i$ and non-random $\mathbf{x}_i$, show that

$$P\left(\left| n^{-1} \sum_{i=1}^{n} \varepsilon_i x_{ij} \right| > \frac{t}{\sqrt{n}} \right) \leq 2 \exp\left(- \frac{nt^2}{2 \sum_{i=1}^{n} x_{ij}^2} \right).$$

(c) If we assume further that $n^{-1} \sum_{i=1}^{n} x_{ij}^2 = 1$ (standardized), show that for any constant $a > 0$,

$$P\left(\|\nabla \ell_n(\boldsymbol{\beta}^*)\|_\infty > a \sqrt{\frac{\log p}{n}} \right) \leq 2p^{1-a^2/2}.$$