36700 – Probability and Mathematical Statistics

Spring 2019

Homework 8

Due Friday, April 19th at 12:40 PM

- All homework assignments shall be uploaded using Gradescope through the Canvas portal). Late submissions are not allowed.

1. (Analysis of Variance) For $k = 1, ..., K$, let X_{ki} $(i = 1, ..., n)$ be independent $N(\mu_k, \sigma^2)$ random variables, where $\mu_1, ..., \mu_K, \sigma^2$ are unknown parameters. We want to test

$$H_0 : \mu_1 = \mu_2 = ... = \mu_K \text{ vs } H_1 : \mu_k \neq \mu_l \text{ for some } k, l.$$

Define

$$\bar{X}_k = \frac{1}{n}\sum_{i=1}^{n} X_{ki}\,,$$

$$\bar{X} = \frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n} X_{ki} = \frac{1}{K}\sum_{k=1}^{K}\bar{X}_k\,,$$

$$W_1 = \sum_{k=1}^{K}\sum_{i=1}^{n}(X_{ki} - \bar{X}_k)^2\,,$$

$$W_0 = \sum_{k=1}^{K}\sum_{i=1}^{n}(X_{ki} - \bar{X})^2\,,$$

(a) Show that $\sigma^{-2}W_1 \sim \chi^2_{k(n-1)}$.

(b) Show that under H_0, $\sigma^{-2}(W_0 - W_1) \sim \chi^2_{k-1}$ and is independent with W_1.

(c) How would you construct a rejection rule with type I error controlled at level α? [Hint: if $Y_1, ..., Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$, then $\bar{Y}$ and $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ are independent. Furthermore, $\sigma^{-2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \sim \chi^2_{n-1}$. You may also find this helpful: for any $x_1, ..., x_m$ and any y, $\sum_{i=1}^{m}(x_i - y)^2 = m(\bar{x} - y)^2 + \sum_{i=1}^{m}(x_i - \bar{x})^2$.]

2. Consider regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\mathbb{E}\boldsymbol{\epsilon} = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}_n$. Assume that $\mathbf{X}$ has k columns and has full rank. Let $\hat{\boldsymbol{\beta}}$ be the least square estimate. Let $\tilde{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}$, where $\tilde{\boldsymbol{\epsilon}}$ is an independent draw from the same distribution as $\boldsymbol{\epsilon}$. Show that

$$\mathbb{E}\left\{\|\tilde{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2\right\} = (n + k)\sigma^2\,,$$

where the expectation is taken over the randomness of both $\tilde{\mathbf{Y}}$ (as a function of $\tilde{\boldsymbol{\epsilon}}$) and $\hat{\beta}$ (as a function of $\boldsymbol{\epsilon}$).

Comparing this with $\mathbb{E}\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 = (n - k)\sigma^2$, we see that Mallows' C_p and GCV provide unbiased estimates of the predictive risk in the overfitting case.

3. In linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\mathbf{X}$ is non-random (fixed design) and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ with **unknown** σ^2. Derive the AIC formula for this model.

4. In a linear regression problem, let $\hat{\boldsymbol{\beta}}$ be the least square estimate, and $\hat{\boldsymbol{\beta}}^{(i)}$ be the least square estimate without data point i. Let x_i^T be the ith row of $\mathbf{X}$, then it is straightforward to verify that

$$\hat{\boldsymbol{\beta}}^{(i)} = \left[\mathbf{X}^T\mathbf{X} - x_i x_i^T\right]^{-1} (\mathbf{X}^T\mathbf{Y} - y_i x_i)\,.$$

Let $\hat{\epsilon}_i = y_i - x_i^T \hat{\boldsymbol{\beta}}$ and $\tilde{\epsilon}_i = y_i - x_i^T \hat{\boldsymbol{\beta}}^{(i)}$. Prove that

$$\tilde{\epsilon}_i = \frac{\hat{\epsilon}_i}{1 - \mathbf{H}_{ii}}$$

where $\mathbf{H}_{ii} = x_i^T (\mathbf{X}^T\mathbf{X})^{-1} x_i$ is the ith diagonal entry of $\mathbf{H}$.
Hint: you will find the Sherman-Morrison identity useful: For a square d by d invertible matrix A and d by 1 vector u

$$(A - uu^T)^{-1} = A^{-1} + \frac{A^{-1}uu^T A^{-1}}{1 - u^T A^{-1} u}$$

5. Assume $(X_i, Y_i)_{i=1}^n$ with $X_i \in \mathbb{R}^d$ and $Y_i \in \{0, 1\}$ follows a logistic regression model with unknown parameter $\boldsymbol{\beta}$. Find the Fisher information matrix $I_n(\boldsymbol{\beta})$ (assuming fixed design: X is non-random, Y is random) and construct a χ^2 test for $H_0 : \boldsymbol{\beta} = 0$ vs $H_1 : \boldsymbol{\beta} \neq 0$ (You can assume the regularity conditions hold so that the MLE is asymptotically normal).

6. Download the data file "`hw8q6.csv`" on Canvas. The data set contains $n = 100$ pairs $(X_i, Y_i)_{i=1}^{100}$ generated by $Y = r(X) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$ is independent noise, with σ^2 unknown. We know that $r(x) = \mathbb{E}(Y|X = x)$ is a polynomial of order $d \leq 6$ and would like to perform model selection to determine d.

   (a) Find d using AIC, and report AIC_k for each k.

   (b) Find d using BIC, and report BIC_k for each k.

   (c) Find d using Mallows' C_p, and report $\mathrm{C}_p(k)$ for each k.

   (d) Find d using loocv, and report CV_k for each k.

   (e) Find d using 5-fold cross-validation and report CV_k for each k.

   There is no need to submit code, because the answers are deterministic, except part (e).

Optional problem. Prove the hint in Q1 through the more general result: If $Y_1, ..., Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$ then $\bar{Y}$ and $(Y_1 - \bar{Y}, Y_2 - \bar{Y}, ..., Y_n - \bar{Y})$ are independent. Moreover $\sigma^{-2} \sum_{i=1}^n (Y_i - \bar{Y})^2 \sim \chi^2_{n-1}$.