STA 561: Homework 1 (Due Jan 19 at midnight)

Reminder: work together! Share ideas, brainstorm, explain/verify your answers but write up your own work. Your homework should be submitted as pdf file generated using either latex or an python notebook.

1. (Covariate shift) Suppose that $(\boldsymbol{X}, Y) \in \mathbb{R}^p \times \mathbb{R}$ are drawn from a distribution P with density $p(\boldsymbol{x}, y) = p(y|x)p(x)$. Define

$$\begin{aligned} \boldsymbol{\beta}_P^* &= \arg\min_{\boldsymbol{\beta}} P(Y - \boldsymbol{X}^\intercal \boldsymbol{\beta})^2 \\ &= \arg\min_{\boldsymbol{\beta}} \int (y - \boldsymbol{x}^\intercal \boldsymbol{\beta})^2 p(\boldsymbol{x}, y) d\boldsymbol{x} dy. \end{aligned}$$

We are interested in estimating $\boldsymbol{\beta}_P^*$.

Suppose that we observe an i.i.d. sample $\left\{(\widetilde{\boldsymbol{X}}_i, \widetilde{Y}_i)\right\}_{i=1}^n$ that is not drawn from $p(\boldsymbol{x}, y)$ but rather from a distribution Q with density $q(\boldsymbol{x}, y) = p(y|\boldsymbol{x})q(\boldsymbol{x})$, i.e., the distribution of $\widetilde{Y}|\widetilde{\boldsymbol{X}}$ is the same as $Y|\boldsymbol{X}$ but the marginal distribution of $\widetilde{\boldsymbol{X}}$ is different from that of $\boldsymbol{X}$ ($p \neq q$). This can arise if we learn a model using one population (say patients in the Duke Healthcare system) but want to make predictions in another population (say patients in the University of Michigan healthcare system).

   (a) Define $\boldsymbol{\beta}_Q^* = \arg\min_{\boldsymbol{\beta}} Q(Y - \boldsymbol{X}^\intercal \boldsymbol{\beta})^2$ and provide generative models P and Q as described above where $\boldsymbol{\beta}_Q^* \neq \boldsymbol{\beta}_P^*$.

   (b) Show that

$$\boldsymbol{\beta}_P^* = \arg\min_{\boldsymbol{\beta}} Q\left\{\frac{p(\boldsymbol{X})}{q(\boldsymbol{X})}(Y - \boldsymbol{X}^\intercal \boldsymbol{\beta})^2\right\},$$

   provided $q(\boldsymbol{x}) = 0$ whenever $p(\boldsymbol{x}) = 0$.

   (c) Using the previous part, suggest and estimator of $\boldsymbol{\beta}_P^*$ for use with data drawn from Q. You can assume that p and q are known. Show this estimator is consistent and asymptotically normal (provide an explicit formula for the limiting covariance).

   (d) Suppose that p and q were unknown but you had a sample $\{\boldsymbol{X}_j\}_{j=1}^M$ drawn from p (this might happen, for example, if we had the covariates but not the responses from another health system). Suggest an estimator of $\boldsymbol{\beta}_P^*$.

2. (Number detective) In this problem you will be exploring data from a study on gene expression conducted at the Lineberger Cancer Center at UNC-CH. The data are stored in the csv file lineberger_study_data.csv which can be downloaded from the class webpage. In this study, researchers from UNC-CH argued they can predict normalized survival across several cancer types (column 2) accurately from a set of only 17 genes (expression levels recorded in columns 3-19; the first column are row numbers which can be removed). However, several of these genes have no obvious biological connection to survival and so researcher are skeptical. Your job is to dig in to the data to look for signs of fraud. The investigators used a linear predictive model which you will replicate and diagnose here.

   (a) Load the data and make a scatter plot matrix. Do you see any obvious outliers? Which predictors appear most strongly linearly correlated with the outcome?

   (b) Fit a linear model. Report the estimated coefficients and compute a 95% confidence interval for each coefficient. Also compute the (unadjusted) R^2.

   (c) Plot the fitted values against the residuals. Do you see any patterns that might suggest a problem with their data/model? Your answer should consist of two words: ___ _____!