18.650. Fundamentals of Statistics
Fall 2023. Problem Set 4

Due Monday, Dec. 4

Problem 1

Let (X, Y) be a pair of random variables following the model

$$Y = \beta_0^* + \beta_1^* X + \beta_2^*(X^2 - 1) + \varepsilon\,,$$

where $\varepsilon \sim \mathcal{N}(0, 1)$ is independent of $X \sim \mathcal{N}(0, 1)$.

Assume that we observe n i.i.d copies $(X_1, Y_1), \ldots, (X_n, Y_n)$ of (X, Y).

1. What is the regression function $f(x)$ of Y onto X?
2. Sketch the curve of f in the following three cases: (i) $\beta_2 = -1$, (ii) $\beta_2 = 1$, and (iii) $\beta_2 = 0$.
3. Define $\vec{Y} = (Y_1, \ldots, Y_n)^\top$. Show that

   $$\vec{Y}|\{(X_1, \ldots, X_n)\} \sim \mathcal{N}_n(\mathbb{X}\beta^*, I_n)\,,$$

   for some design matrix $\mathbb{X}$ and some vector β^* to be made explicit.
4. Show that

   $$D = \lim_{n\to\infty} \frac{1}{n}\mathbb{X}^\top\mathbb{X}$$

   is a deterministic diagonal matrix to be made explicit. `[Hint: check the limit of each entry in the matrix]`.

   In the rest of this problem, we assume that n is large enough so that we can take

   $$D = \frac{1}{n}\mathbb{X}^\top\mathbb{X}$$

5. Compute the least squares estimator $\hat{\beta}$ in terms of the X_is and the Y_is.
6. What is the asymptotic distribution of $\hat{\beta}$?
7. We want to test

   $$H_0 : \beta_0^* = 0, \quad vs \quad H_1 : \beta_0^* \neq 0$$

   Assume that $n = 243$ and $\bar{Y}_n = 0.13$ and compute the p-value for this test. Conclude.

Problem 2

Consider the linear regression model with $2n+1$ observations given by

$$Y_i = \beta_0 + \beta_1 \frac{i}{n} + \varepsilon_i, i = -n, -(n-1), \ldots, -1, 0, 1, \ldots, n$$

where ε_i are i.i.d $N(0,1)$.

We recall that

$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}.$$

1. Let $\bar{Y}_n = \frac{1}{2n+1}\sum_{i=-n}^{n} Y_i$. Find the distribution of $\bar{Y}_n$ in terms of the ground truth parameters $\beta_0 = \beta_0^*$ and $\beta_1 = \beta_1^*$.
2. Compute the MLE $\hat{\beta}_0$ and $\hat{\beta}_1$ (simplify the formula as much as possible).
3. Consider now a noninformative prior $f(\beta_0, \beta_1) \propto 1$. Show that the posterior distribution of (β_0, β_1) given $Y_1, \ldots, Y_n$ is a a multivariate Gaussian distribution with mean and covariance matrix to be computed explicitly.
4. Compute the Bayes estimator $(\tilde{\beta}_0, \tilde{\beta}_1)$ for (β_0, β_1)?
5. Let (Z_0, Z_1) be a random variable drawn from the posterior distribution. What is the limiting distribution of $\sqrt{n}(Z_1 - \hat{\beta}_1)$ as $n \to \infty$?

Problem 3

Let $X \in \{0,1\}$ be a binary treatment and (C_0, C_1) denote the corresponding potential outcomes. Let $U \sim \mathsf{Unif}(-1,1)$, and consider an experiment where patients are assigned to the treatment group $X = 1$ if $U > 0$ and to the control group $X = 0$ if $U \leq 0$. The potential outcomes are given by

$$C_1 = U\mathbb{1}(U < 0.5)$$
$$C_0 = U\mathbb{1}(U > 0)$$

Compute the average treatment effect θ and the association α. Comment on your result.

Problem 4 Do not turn in this problem

Let n, m and K be three positive integers such that $n = Km$. Let $(x_1, Y_1), \ldots, (x_n, Y_n)$ be independent such that $x_i = (i-1)/n$ and

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n$$

for some ε_i that are i.i.d $N(0,1)$. Here f is the unknown regression function of interest.

1. ...

Problem 5 Assume that we observe pairs $(Y_1, X_1), \dots, (Y_n, X_n) \in \{-1, 1\} \times \mathbb{R}$, i.i.d such that

$$\operatorname{logit}(\mathbb{P}(Y_i = 1 | X_i = x)) = \beta^* \cdot x$$

for some unknown parameter β^*. Assume further that the X_is have been normalized so that $\sum_i X_i = 0$ and $\sum_i X_i^2 = 1$.

1. Write the log-likelihood $\ell_n(\beta)$ depending on the X_i and Y_i's.
2. Compute the least-squares estimator $\hat{\beta}^{\mathsf{LS}}$ that solves

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - \beta X_i)^2$$

3. Using the second order Taylor expansion of the log-likelihood around $\beta = 0$, find w such that $\hat{\beta}^{\mathsf{MLE}} \approx w \hat{\beta}^{\mathsf{LS}}$.

Problem 6 Let $X_1, \dots, X_n$ be n i.i.d. positive random variables with CDF F here X_i is the lifetime of iPhone i. Rather than observing the X_is, we only have access to current status data of the following form. Let $Z_i(t) = \mathbb{1}(X_i > t)$, that is which iPhone is still working at time t.

Assume first that we observe

$$Z_1(t), \dots, Z_n(t)$$

for all $t > 0$.

1. What is the distribution of $Z_1(t)$?
2. What is the maximum likelihood estimator of $F(t)$?
   Assume now that there are some random times $T_1, \dots, T_n \sim \mathsf{Exp}(1)$ iid, we observe:

$$Y_i := Z_i(T_i), i = 1, \dots, n.$$

   Note that we do not observe the T_i's
3. What is the distribution of Y_i?
4. Assume now that[1] $X_1, \dots, X_n \sim \mathbf{Exp}(\beta)$ i.i.d for some unknown $\beta > 0$. Propose an estimator $\hat{\beta}$ for β based on $Y_1, \dots, Y_n$ using the plugin method.

[1] We use the convention of AoS for exponential distribution.

5. Show that $\hat{\beta}$ is asymptotically normal and compute its asymptotic variance.

Problem 7 Sentiment Analysis with Regression

As a data scientist at "18.650 Tech", a large tech company, your task is to develop a model to analyze customer sentiment on social media. The company is interested in classifying user posts into two categories: "Positive" or "Negative". This classification is based on the text of the posts. In particular, your task is to develop a regression model to predict the sentiment of a post.

For this task, you are given a dataset containing thousands of social media posts. Each post has been manually labeled as "Positive" or "Negative" by human reviewers. Beyond the post itself, each post in the dataset includes various features extracted from the text, such as:

- Word Count: The total number of words in the post.
- Hashtags: The number of hashtags used.
- Mentions: The number of times other users are mentioned.
- Emojis: The number of emojis used.
- Exclamation Marks: The number of exclamation marks.
- Question Marks: The number of question marks.
- Sentiment Score: A pre-computed sentiment score ranging from -1 (very negative) to 1 (very positive), derived using open-source sentiment analysis tools.

1. Given the context of predicting sentiment ("Positive" or "Negative") from social media posts, do you expect a linear regression model or a logistic regression model to be more suitable for this task? Justify your answer based on the characteristics of the data and the nature of the problem.
2. Using your chosen regression model from part (1), develop a model to predict the sentiment of a social media post (discuss what are the features and what are the labels). Describe how you would use the given features in your model. Would you consider transforming any of these features or creating new features? Explain your reasoning.
3. Once your model is developed, interpret the coefficients of the regression model. What do the signs and magnitudes of the coefficients suggest about the importance and impact of each feature on the sentiment of a post?
4. Propose a method to evaluate the performance of your logistic regression model. Which metrics would you use and why?

5. Discuss any ethical considerations you should take into account when developing and deploying a model that analyzes social media posts. How might biases in the data affect your model, and what steps could you take to mitigate these biases?

Hedge fund interview question

In every PSet, we have an additional question taken from a hedge fund interview. This question is not mandatory and does not hold any point but you are welcome to give it a shot.

Problem 8 (Source: Two Sigma)

Suppose we ran a least-squares linear regression on a set of data and record the regression coefficients, R^2 value, and confidence intervals for our regression coefficients. Now, suppose this dataset was accidentally duplicated and someone re-runs the regression. Which values will change and why?