18.650. Fundamentals of Statistics
Fall 2022. Problem Set 4

Due Monday, Dec. 4 by 11:59pm

Problem 1

Let (X, Y) be a pair of random variable 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?
   Solution: The regression function is $f(x) = \beta_0^* + \beta_1^* x + \beta_2^*(x^2 - 1)$
2. Sketch the curve of f in the following three cases: (i) $\beta_2 = -1$, (ii) $\beta_2 = 1$, and (iii) $\beta_2 = 0$.
   Solution: The curves are respectively a parabola that opens downwards, a parabola that opens upwards, and a line.
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.
   Solution: Let:

   $$\mathbb{X} = \begin{pmatrix} 1 & X_1 & X_1^2 - 1 \\ \vdots & \vdots & \vdots \\ 1 & X_n & X_n^2 - 1 \end{pmatrix}$$

   Additionally, let $\vec{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ and $\beta^* = (\beta_0^*, \beta_1^*, \beta_2^*)$. Then we have:

   $$\vec{Y} = \mathbb{X}\beta^* + \vec{\epsilon}$$

   Conditioned on $X_1, \ldots, X_n$, the random variable $\mathbb{X}\beta^*$ becomes constant while $\vec{\epsilon}$ remains a $\mathcal{N}_n(0, I_n)$ random variable as the $\vec{\epsilon}$ are i.i.d. Hence, we get the desired.
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]`.

Solution: Expanding out the matrix products we get:

$$\frac{1}{n}(\mathbb{X}^\top \mathbb{X})_{i,j} = \begin{cases} \frac{1}{n}\sum_{i=1}^n X_i & i,j = 1,2 \\ \frac{1}{n}\sum_{i=1}^n (X_i^2 - 1) & i,j = 1,3 \\ \frac{1}{n}\sum_{i=1}^n (X_i^3 - X_i) & i,j = 2,3 \\ \frac{1}{n}\sum_{i=1}^n 1 & i,j = 1,1 \\ \frac{1}{n}\sum_{i=1}^n X_i^2 & i,j = 2,2 \\ \frac{1}{n}\sum_{i=1}^n (X_i^2 - 1)^2 & i,j = 3,3 \end{cases}$$

By the LLN, each summation will converge to the expectation of the random variable the sum is over. These expectations are easily computed using $\mathbb{E}[X_1] = \mathbb{E}[X_1^3] = 0$, $\mathbb{E}[X_1^2] = 1$, and

$$\mathbb{E}[X_1^4] = 3$$

. For example $\mathbb{E}[X_1] = 0$ so $\frac{1}{n}(\mathbb{X}^\top \mathbb{X})_{1,2} \to 0$. Doing this for each element above gives :

$$\frac{1}{n}(\mathbb{X}^\top \mathbb{X}) \to \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

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.
Solution: Using the preceeding part along with AoS theorem 13.13:

$$\begin{aligned} \hat{\beta} &= (\mathbb{X}^\top \mathbb{X})^{-1}\mathbb{X}^\top Y \\ &= \frac{1}{n}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix}\begin{pmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \\ X_1^2 - 1 & \cdots & X_n^2 - 1 \end{pmatrix}\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n Y_i \\ \frac{1}{n}\sum_{i=1}^n X_i Y_i \\ \frac{1}{2n}\sum_{i=1}^n (X_i^2 - 1)Y_i \end{pmatrix} \end{aligned}$$

6. What is the asymptotic distribution of $\hat{\beta}$?
Solution: Again using AoS theorem 13.13,

$$\hat{\beta} \sim \mathcal{N}(\beta^*, (\mathbb{X}^\top \mathbb{X})^{-1}) = \mathcal{N}\left(\beta^*, \frac{1}{n}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix}\right)$$

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.

Solution: By the preceding parts, asymptotically we have $\hat{\beta}_0^* \sim \mathcal{N}(0, \frac{1}{n})$ and $\hat{\beta}_0^* = \overline{Y}_n$ so we can just perform a Wald test. The test statistic is $\sqrt{n}\overline{Y}_n \approx 2.026$, which yields a p-value of approximatley .043.

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. Compute the average $\bar{Y}_n$. **Solution:** Since we're summing over both positive and negative i/n's, the β_1 term vanishes when we take the average, so that $\bar{Y}_n = \beta_0 + \bar{\epsilon}_n$, where β_0 is the ground truth value and $\bar{\epsilon}_n$ is the average of the noise.
2. Compute the MLE $\hat{\beta}_0$ and $\hat{\beta}_1$ (simplify the formula as much as possible). **Solution:** The log likelihood is

$$\ell_n(\beta_0, \beta_1) = \frac{1}{2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 i/n)^2.$$

up to sign and constants. Setting the first partial to zero we get

$$0 = \partial_{\beta_0} \ell_n(\beta_0, \beta_1) = \sum_i (\beta_0 + \beta_i \frac{i}{n} - Y_i) = (2n+1)\beta_0 - \sum_i Y_i.$$

Hence $\hat{\beta}_0 = \bar{Y}_n$. Setting the second partial to zero we get

$$0 = \partial_{\beta_1} \ell_n(\beta_0, \beta_1) = \sum_i (\beta_0 + \beta_1 \frac{i}{n} - Y_i)\frac{i}{n} = \beta_1 \sum_{i=-n}^{n} \left(\frac{i}{n}\right)^2 - \sum_i Y_i \frac{i}{n}.$$

Hence

$$\frac{2n(n+1)(2n+1)}{6n^2} \beta_1 = \sum_{i=-n}^{n} \left(\frac{i}{n}\right)^2 \beta_1 = \sum_i \frac{i}{n} Y_i$$

which gives

$$\hat{\beta}_1 = \frac{3n}{(n+1)(2n+1)} \sum_{i=-n}^{n} \frac{i}{n} Y_i.$$

3. Consider now a noninformative prior $f(\beta_0, \beta_1, \sigma^2) \propto 1$. Show that the posterior distribution of (β_1, β_2) given $Y_1, \ldots, Y_n$ is a a multivariate Gaussian distribution with mean and covariance matrix to be computed explicitly. **Solution:** We have

$$f(\beta_0, \beta_1 | Y_1, \ldots, Y_n) \propto \exp\left(-\frac{1}{2} \sum_{i=-n}^{n} \left(Y_i - \beta_0 - \beta_1 \frac{i}{n} \right)^2 \right).$$

The exponent is quadratic in β_0 and β_1 so we automatically get that it's a Gaussian posterior. Note that

$$\sum_{i=-n}^{n} \left(Y_i - \beta_0 - \beta_1 \frac{i}{n} \right)^2 = \sum_{i=-n}^{n} \left[Y_i^2 + \beta_0^2 + \frac{i^2}{n^2} \beta_1^2 - 2Y_i\beta_0 - 2Y_i\beta_1 \frac{i}{n} + 2\frac{i}{n}\beta_0\beta_1 \right] \tag{1}$$

We see that the $\beta_0\beta_1$ cross term will vanish, so the Gaussian we are getting has independent coordinates. Furthermore we know the mean of the Gaussian has to be the MLE since we used a flat prior. For the variances we get $1/(2n+1)$ for β_0 and $(\sum_i i^2/n^2)^{-1} = \frac{3n}{(n+1)(2n+1)}$ for β_1. To summarize, the posterior is

$$\mathcal{N}\left(\begin{pmatrix} \bar{Y}_n \\ \frac{3n}{(n+1)(2n+1)} \sum_{i=-n}^{n} \frac{i}{n} Y_i \end{pmatrix}, \begin{pmatrix} \frac{1}{2n+1} & 0 \\ 0 & \frac{3n}{(n+1)(2n+1)} \end{pmatrix} \right)$$

4. Compute the Bayes estimator $(\tilde{\beta}_0, \tilde{\beta}_1)$ for (β_0, β_1)? **Solution:** The Bayes estimator is just the mean: $\tilde{\beta}_0 = \bar{Y}_n$ and $\tilde{\beta}_1 = \hat{\beta}_1 = \frac{3n}{(n+1)(2n+1)} \sum_{i=-n}^{n} \frac{i}{n} Y_i$.
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 $n \to \infty$?
Solution: We have $Z_1 \stackrel{d}{=} \mathcal{N}(\hat{\beta}_1, \frac{3n}{(n+1)(2n+1)})$, so $\sqrt{n}(Z_1 - \hat{\beta}_1)$ has distribution $\mathcal{N}(0, \frac{3n^2}{(n+1)(2n+1)})$ which converges to $\mathcal{N}(0, 3/2)$

Problem 3

Let $X \in \{0, 1\}$ be a binary treatment and let (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.

Solution: We have:

$$\mathsf{E}[C_1] = \frac{1}{2}\int_{-1}^{1} u\mathbb{1}(u < 0.5)du = \frac{1}{2}\int_{-1}^{.5} udu = -\frac{3}{16}$$

and

$$\mathsf{E}[C_0] = \frac{1}{2}\int_{-1}^{1} u\mathbb{1}(u > 0)du = \frac{1}{2}\int_{0}^{1} udu = \frac{1}{4}$$

giving:

$$\theta = \mathsf{E}[C_1] - \mathsf{E}[C_0] = -\frac{7}{16}.$$

For the association:

$$\begin{aligned}
\alpha &= \mathsf{E}[Y|X = 1] - \mathsf{E}[Y|X = 0] \\
&= \mathsf{E}[C_1|\mathbb{1}(U > 0) = 1] - \mathsf{E}[C_0|\mathbb{1}(U > 0) = 0] \\
&= \mathsf{E}[U\mathbb{1}(U < .5)|U > 0] - \mathsf{E}[U\mathbb{1}(U > 0)|U < 0] \\
&= \int_0^1 u\mathbb{1}(u < .5)du - \int_{-1}^0 0du \\
&= \frac{1}{8}
\end{aligned}$$

Despite the fact that the treatments were selected randomly, we see that association does not equal average causal effect. This is because the distribution of the outcome was not independent from how the treatments were randomly selected.

Problem 4

Let n, m and K be three positive integers such that $n = Km$. Let $(x_1, Y_1), \ldots, (x_n, Y_n)$ be i.i.d 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. Recall the definition of the regressogram $\hat{f}$ with m bins.
Solution: Set $B_i = [\frac{i}{m}, \frac{i+1}{m})$. For $x \in B_l$, the regressogram is given by:

$$\hat{f}(x) = \frac{\sum_{i=1}^n Y_i \mathbb{1}(X_i \in B_l)}{\sum_{i=1}^n \mathbb{1}(X_i \in B_l)}$$

2. How many x_is fall into the first bin $B_1 = [0, \frac{1}{m})$?
Solution: We have $x_1, \ldots, x_K \in B_1$ so the answer is K. This result also holds for each bin.
3. For $x \in B_1$, what is the distribution of $\hat{f}(x)$? Compute the bias $b(x)$ and the variance $v(x)$ of $\hat{f}(x)$.
Solution: By the previous part, the regressogram simplifies to:

$$\hat{f}(x) = \frac{1}{K}\sum_{i=1}^K Y_i = \frac{1}{K}\sum_{i=1}^K f(x_i) + \frac{1}{K}\sum_{i=1}^K \epsilon_i$$

The first summation is a constant, and the second is an average of K standard normals. Thus $v(x) = \frac{1}{K}$,

$$b(x) = \left| f(x) - \frac{1}{K}\sum_{i=1}^K f(x_i) \right|$$

and

$$\hat{f}(x) \sim \mathcal{N}\left(\frac{1}{K}\sum_{i=1}^K f(x_i), \frac{1}{K}\right).$$

Assume now f is 1-Lipschitz and $K \geq 3$.
4. Show that the bias is upper bounded as

$$|b(x)| \leq \frac{K}{n}$$

`[Hint: if` x `and` x_i `are in the same bin then` $|x - x_i| \leq 1/m$`.]`
Solution: By the 1-Lipshitz property:

$$|f(x) - f(x_i)| \leq |x - x_i| \leq \frac{1}{m}$$

whenever $x, x_i \in B_1$. By the triangle inequality then:

$$b(x) \leq \frac{1}{K}\sum_{i=1}^K |f(x) - f(x_i)| \leq \frac{1}{m} = \frac{K}{n}.$$

5. Give a choice of K such that

$$\mathsf{MISE}(\hat{f}) \leq \frac{2}{n^{2/3}}$$

`[Tip: Don't bother with rounding; when optimizing over` K`, simply assume that the optimizer is an integer. ]`

Solution: Note that the logic from the previous parts applies to all bins not just B_1. Taking $K = n^{-2/3}$, we get:

$$\mathsf{MISE}(\hat{f}) = \int_0^1 (b(x)^2 + v(x))dx \leq \int_0^1 \left(\frac{K^2}{n^2} + \frac{1}{K}\right) dx = \frac{2}{n^{2/3}}$$

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

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

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. **Solution:** Note that $\mathsf{P}(Y_i = 1|X_i) = \text{logit}^{-1}(\beta X_i) = \frac{1}{1+e^{-\beta X_i}}$, and therefore

$$\mathsf{P}(Y_i = -1|X_i) = 1 - (1 + e^{-\beta X_i})^{-1} = \frac{e^{-\beta X_i}}{1 + e^{-\beta X_i}} = \frac{1}{1 + e^{\beta X_i}}.$$

For both $Y_i = \pm 1$, we can therefore write $\mathsf{P}(Y_i|X_i) = (1 + e^{-\beta X_i Y_i})^{-1}$. Hence

$$\ell_n(\beta) = -\sum_{i=1}^n \log(1 + e^{-\beta X_i Y_i})$$

2. Compute the least-squares estimator $\hat{\beta}^{\mathsf{LS}}$ that solves

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

Solution: Setting the derivative to zero we get

$$0 = -2\sum_i X_i(Y_i - \beta X_i),$$

so that

$$\sum_i X_i Y_i = \beta \sum_i X_i^2 = \beta,$$

since $\sum_i X_i^2 = 1$. Thus $\hat{\beta}^{\mathsf{LS}} = \sum_i Y_i X_i$.

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

$$\ell_n'(\beta) = \sum_{i=1}^n X_i Y_i \frac{e^{-\beta X_i Y_i}}{1 + e^{-\beta X_i Y_i}} = \sum_{i=1}^n X_i Y_i \left(1 - \frac{1}{1 + e^{-\beta X_i Y_i}}\right)$$

and therefore

$$\ell_n''(\beta) = -\sum_{i=1}^n (X_i Y_i)^2 \frac{e^{-\beta X_i Y_i}}{(1 + e^{-\beta X_i Y_i})^2}$$

Note that

$$\ell_n'(0) = \frac{1}{2}\sum_i X_i Y_i, \qquad \ell_n''(0) = -\frac{1}{4}\sum_i X_i^2 Y_i^2 = -\frac{1}{4}$$

using that $\sum_i X_i^2 = 1$ and $Y_i^2 = 1$. Therefore near zero we have

$$\ell_n(\beta) \approx \ell_n(0) + \left(\frac{1}{2}\sum_i X_i Y_i\right)\beta - \frac{1}{8}\beta^2$$

This quadratic is minimized at $\sum_i X_i Y_i$. Arguing that the MLE $\hat{\beta}$ is approximately the minimizer of this quadratic, we get that $\hat{\beta} \approx 2\sum_i X_i Y_i = \hat{\beta}^{\mathrm{LS}}$, so $w = 2$.

Problem 6 Let $X_1, \ldots, X_n$ 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 form $Z_i(t) = \mathbb{1}(X_i > t)$, which tells us which iPhone is still working at time t.

Assume first that we observe

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

for all $t > 0$.

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

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

   Note that we do not observe the T_i's

3. What is the distribution of Y_i? **Solution:** Still Bernoulli. We have

$$\begin{aligned}\mathsf{P}(Y_1 = 1) &= \mathsf{P}(Z_1(T_1) = 1) = \mathsf{P}(X_1 > T_1) = \mathsf{E}[\mathsf{P}(X_1 > T_1 \mid T_1)] \\ &= \mathsf{E}[1 - F(T_1)] = \int_0^\infty (1 - F(t))e^{-t}dt \\ &= 1 + \int F(t)(-e^{-t})dt \\ &= 1 + F(t)e^{-t}\big|_0^\infty - \int_0^\infty f(t)e^{-t}dt = 1 - \mathsf{E}[e^{-X_1}]\end{aligned} \tag{2}$$

4. Assume now that[1] $X_1, \ldots, X_n \sim \text{Exp}(\beta)$ i.i.d for some unknown $\beta > 0$. Propose an estimator $\hat{\beta}$ for β based on $Y_1, \ldots, Y_n$ using the plugin method.
Solution: Note that for $X_1 \sim \text{Exp}(\beta)$ we have

$$\mathbb{E}[e^{-X_1}] = \int_0^\infty e^{-x}\frac{1}{\beta}e^{-\frac{x}{\beta}}dx = \frac{1}{\beta}\int_0^\infty e^{-(1+1/\beta)x}dx = \frac{1/\beta}{1+1/\beta},$$

which simplifies to $1/(\beta+1)$. Therefore, $\mathbb{E}[Y_1] = 1 - \frac{1}{\beta+1} = \beta/(\beta+1)$ using part 3. Solving for β, we get $\beta = 1/(1-\mathbb{E}[Y_1]) - 1 = \mathbb{E}[Y_1]/(1-\mathbb{E}[Y_1])$. Finally, the plugin rule tells us to replace expectations by averages, so we take

$$\hat{\beta} = \frac{\bar{Y}_n}{1-\bar{Y}_n}.$$

5. Show that $\hat{\beta}$ is asymptotically normal and compute its asymptotic variance. **Solution:** Let $g(x) = x/(1-x)$ and $p = \beta/(\beta+1)$ so that $g(p) = \beta$. By CLT we have $\bar{Y}_n \approx \mathcal{N}(p, p(1-p)/n)$. Note that $g'(p) = 1/(1-p)^2$. Thus by the delta method,

$$\hat{\beta} = g(\bar{Y}_n) \approx \mathcal{N}(g(p), g'(p)^2 p(1-p)/n) = \mathcal{N}\left(\beta, \frac{p}{n(1-p)^3}\right) = \mathcal{N}\left(\beta, n^{-1}\beta(1+\beta)^2\right),$$

using that $p/(1-p) = \beta$ and $1/(1-p) = \beta+1$. Therefore, the asymptotic variance is $\beta(1+\beta)^2$.

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

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

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 perform better for this task? Justify your answer based on the characteristics of the data and the nature of the problem.
   Solution: In the context of predicting sentiment, which is a binary outcome ("Positive" or "Negative"), logistic regression is expected to perform better than linear regression. Here's why: Logistic regression is specifically designed for binary or categorical outcome variables. It models the probability of a particular class (e.g., "Positive") and is thus inherently suited for classification tasks. The output of a logistic regression model is a probability that ranges between 0 and 1, aligning perfectly with the need to classify posts as either "Positive" or "Negative." Linear regression, on the other hand, is intended for continuous outcome variables. Using it for a binary outcome would involve predicting a continuous value and then categorizing it, which can be less efficient and more prone to error.
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.
   Solution: We use a logistic regression model where the given features characterize each data point and the "Positive" or "Negative" assignments are the labels. Each

feature (Word Count, Hashtags, Mentions, Emojis, Exclamation Marks, Question Marks, Sentiment Score) can be used as a predictor in the logistic regression model. You can choose to use them directly, although in practice these features would usually get transformed. For example, you could choose to transform the number of exclamation marks into a binary variable that indicates whether there are exclamation marks or not (consider whether there's really a fundamental difference between a post with 9 exclamation marks in comparison to a post with 12 exclamation marks that our model needs to capture). You could also add interaction terms to capture some dependencies between the features (for example, the sentiment score might be strongly correlated with the emoji, so you might want to count them together instead of letting the model think that they have independent contributions).

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?
   Solution: The sign of each coefficient indicates the direction of its impact. A positive coefficient increases the log-odds of a post being positive, while a negative coefficient decreases it. Larger absolute values indicate a stronger effect. For instance, a high positive coefficient for the sentiment score suggests that a higher sentiment score strongly predicts a positive post.
4. Propose a method to evaluate the performance of your logistic regression model. Which metrics would you use and why?
   Solution: To test the metric we could use a different dataset of labeled posts (that weren't used to construct the model) and check the model's predictions on this set. We could use several metrics such as accuracy, precision, and recall, which are very common statistics to test regression models.
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?
   Solution: Sentiment analysis is a very important task with many ethical considerations. For example, if the training data is biased (e.g., more negative posts), the model may inherit this bias. Diverse and representative training data is crucial. To mitigate this bias, you may choose to construct a new dataset where you over-sample under-represented posts and under-sample over-represented posts. In addition, as a large company, "18.650 Tech" must ensure compliance with data privacy regulations and user consent. Finally, the model should preferably be interpretable to ensure that its predictions are fair and unbiased.

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?

Solution: Since our regression coefficients are calculated based least-squares, the total error will simply be double the previous error, and so our minimizer remains the same. Thus, our regression coefficients are unchanged. With similar logic, we see that our R^2 value must remain constant as well. However, since our standard error for each of the hypothesis tests $H_{i0} : \beta_i = 0$ vs $H_{i1} : \beta_i \neq 0$ will be smaller with more data points, our test statistics will grow larger and thus our p-values will shrink (as well as our confidence intervals).