Stat 215A Midterm (Fall 2023)

Name:
Student ID:

1 True or False (10 pts, 1 pt each)

1. When we fit OLS to data (x_i, y_i) with $x_i^T \in R^p$ and $y_i \in R^1$ to get $\hat{\beta}$, we could calculate the residual vector

$$e = Y - X\hat{\beta},$$

It follows that $Y = X\hat{\beta} + e$, which is a linear regression model.

2. The K-means algorithm always converges to a global minimum of its loss function.

3. In data cleaning, an unusual or surprising value should be treated as an invalid value.

4. The stability principle can be applied at almost every stage of the data science life cycle, but is not applicable to data visualization.

5. The predictability principle asserts that if a model doesn't generate good predictions then it is doubtful that it captures real phenomena.

6. It is generally recommended that you split your data into training, validation, and test sets at the beginning of the predictive modeling/analysis stage of data science life cycle.

7. Least Squares can be applied only when the Gaussian linear regression assumptions hold.

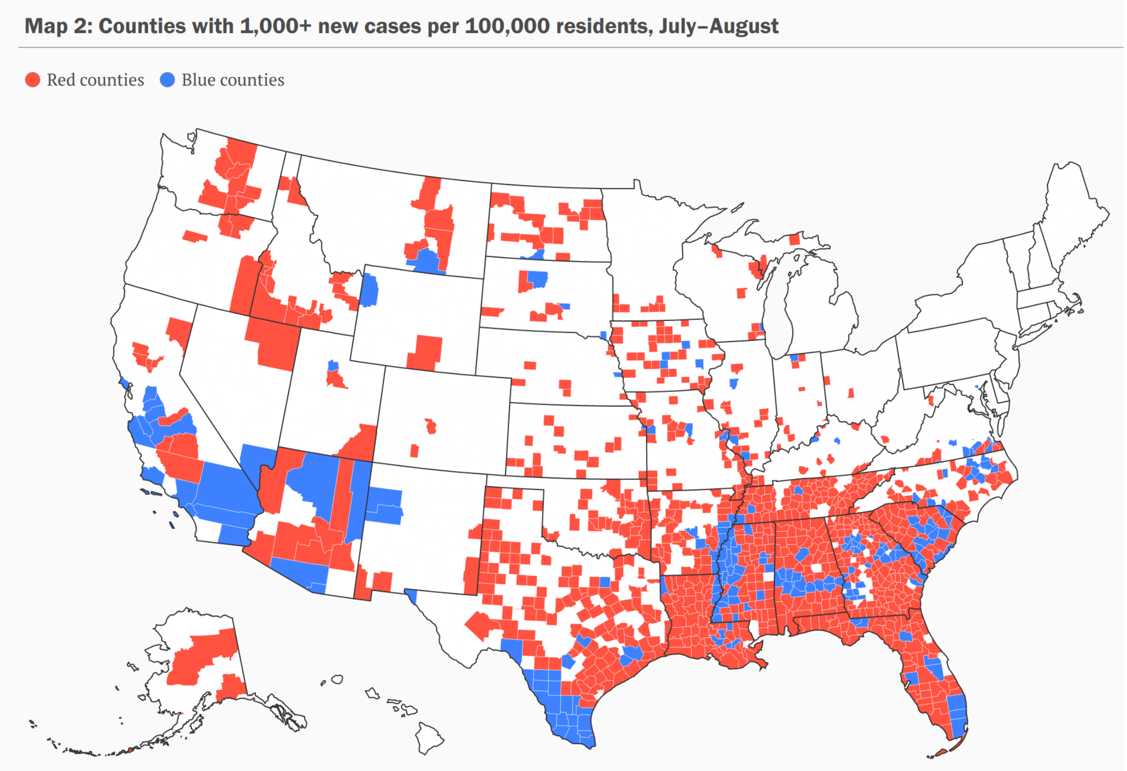
8. For a Gaussian linear regression model, the maximum likelihood estimator for the error variance σ^2 is unbiased.

9. As we decrease the tuning parameter λ to 0, the ridge estimator $\hat{\beta}_{ridge}$ converges to the usual OLS estimator $\hat{\beta}_{OLS}$.

10. Under the Neyman-Rubin causal inference model with a finite population, the randomness in the observed outcome arises from the randomization mechanism.

2 EDA (9 pts)

The figure below comes from a September 2020 article by the Brookings Institution titled *COVID-19's summer surge into red America sets the stage for November's election.* COVID-19 has been an important and controversial subject in the 2020 U.S. presidential election season and was the first topic of discussion during the final presidential debate. In the map below, "Red counties" refer to counties that voted Republican in 2016 and "Blue counties" for those that voted Democrat.



Map 2: Counties with 1,000+ new cases per 100,000 residents, July–August

Figure 1: William H. Frey analysis of New York Times data for confirmed COVID-19 cases, 2019 Census population estimates, and Dave Leip's Atlas of US presidential elections for 2016 voters.

1. (2 points) Comment on the visual quality of the map. Do you think the map is visually appealing? If so, why? If not, why not? Please stick to the aesthetic qualities; we will discuss the information content later.

The author writes

> *The biggest increases in COVID-19 case rates for Republican, rural, and suburban counties occurred in the July-to-August period. Of the 1,084 counties where cases exceeded 1,000 per 100,000 residents, 877 voted Republican in 2016, and all but 32 were classified as non-metropolitan, small metropolitan, or large suburban counties. And even though urban and Democratic counties are larger in size, nearly half of all residents in these 1,084 counties (an aggregation of many small towns and rural areas) voted for Trump in 2016. [...] It remains to be seen whether or not President Trump's attempt to downplay the pandemic's impact on the lives of residents—whether red or blue, rural or urban—will work to his campaign's advantage. However, even if there was some doubt in the late spring of this year, this new data makes it clear that COVID-19 is not the problem of people "somewhere else."*

2. (3 points) Briefly describe another type of visualization that you could create to display the same information detailed in the paragraph above. Would that be a more effective way to display the data, or do you prefer the map? Explain why.

3. (4 points) The author implicitly makes a connection between data, reality, and future data (the distribution of "red" and "blue" counties in the 2020 presidential election). Please address each of the following items:
   - Name one way in which the author's data is a good approximate reflection of reality and one way in which it is lacking.
   - Is their data similar to future data? Why or why not?
   - What additional data might be helpful to evaluate the potential impact of the pandemic on the upcoming presidential election?

3 OLS under logistic regression models (9 pts)

Given n data units (x_i, y_i), $i = 1, \ldots, n$ with $x_i \in R^1$ and $y_i \in \{0, 1\}$.

1. Find the OLS estimator $\hat{\beta}_{OLS}$ of β by minimizing (3 pts)

$$\sum_i (y_i - \beta x_i)^2.$$

Suppose, for given fixed x_i's, Y_i's are generated as independent samples from a logistic regression model with one parameter α: $EY_i = \frac{e^{\alpha x_i}}{1+e^{\alpha x_i}}$. Note that the logistic regression model is a generalization of the standard regression model for binary outcomes. The equivalent formulation for a simple linear regression model can be written as $EY_i = \beta x_i$.

2. Is $\hat{\beta}_{OLS}$ an unbiased estimator of α under the logistic regression model above? Why or why not? (3 pts)

3. Find the variance of $\hat{\beta}_{OLS}$ under the same logistic regression model (3 pts).

4 EM algorithm (12 pts)

Consider the following statistical model:
for $j = 0, 1$ and parameter $\theta = (\mu_0, \mu_1, \pi)$ where $\mu_0, \mu_1 \in R^1, \pi \in (0, 1)$,

$$x_i | z_i = j \sim \text{Laplace}(\mu_j, b)^1$$
$$z_i \sim \text{Ber}(\pi)$$

where $(x_1, z_1), \ldots, (x_n, z_n)$ pairs are i.i.d, $x_1, \ldots, x_n$ are observed while $z_1, \ldots, z_n$ are not. We denote $\boldsymbol{z} = (z_1, \ldots, z_n)$ and $\boldsymbol{x} = (x_1, \ldots, x_n)$

1. Write down the log-likelihood function of θ given $(x_1, z_1), \ldots, (x_n, z_n)$ (3 pts) .

2. Given a fixed value for θ, calculate $\mathbb{P}(z_i = 1 | x_i = x, \theta)$ (3 pts) .

[1] A random variable has a Laplace(μ, b) distribution if $f(x) = \frac{1}{2b} \exp(-\frac{|x-\mu|}{b})$

Assuming b is known, consider estimating $\theta = (\mu_0, \mu_1, \pi)$ using the EM algorithm.

3. Denote by $\theta^{(t)}$ the value of θ after t EM iterations, we define $Q(\theta|\theta^{(t)}) = \mathbb{E}_{\boldsymbol{z}|\boldsymbol{x},\theta^{(t)}} \log(\mathcal{L}((x_1, z_1), \ldots, (x_n, z_n)|\theta))$. Derive the E-step, i.e simplify the function $Q(\theta|\theta^{(t)})$ (3 pts) .

4. Derive the M-step, i.e find $\arg\max_\theta Q(\theta|\theta^{(t)})$ (3 pts) .