Data 102, Spring 2023 Midterm 2

- You have 110 minutes to complete this exam. There are 7 questions, totaling 41 points.
- You may use two 8.5 × 11 sheets of handwritten notes (front and back), and the provided reference sheet. No other notes or resources are allowed.
- You should write your solutions inside this exam sheet.
- You should write your name and Student ID on every sheet (in the provided blanks).
- Make sure to write clearly. We can't give you credit if we can't read your solutions.
- Even if you are unsure about your answer, it is better to write down partial solutions so we can give you partial credit.
- You may, without proof, use theorems and facts that were given in the discussions or lectures, **but please cite them**.
- We don't answer questions individually. If you believe something is unclear, bring your question to us and if we find your question valid we will make a note to the whole class.
- Unless otherwise stated, no work or explanations will be graded for multiple-choice questions.
- Unless otherwise stated, you must show your work for free-response questions in order to receive credit.

Last name	
First name	
Student ID (SID) number	
Berkeley email	
Name of person to your left	
Name of person to your right	

Honor Code (1 point)

I will respect my classmates and the integrity of this exam by following this honor code.

I affirm:

- All of the work submitted here is my original work.
- I did not collaborate with anyone else on this exam.

Signature: ______________________________

This page intentionally left blank for scratch work. No work on this page will be graded.

1. (5 points) For each of the following, determine whether the statement is true or false. For this question, no work will be graded and no partial credit will be assigned.

   (a) (1 point) The bootstrap is a good choice for estimating uncertainty in the range (maximum value minus minimum value) of a dataset.

   ○ A. TRUE [✓ ○ B. FALSE]

   (b) (1 point) When predicting a new y-value using a GLM, only the Bayesian approach can give us a measure of uncertainty for that prediction.

   ○ A. TRUE [✓ ○ B. FALSE]

   (c) (1 point) In order to apply Hoeffding's inequality to a sequence of bounded random variables $x_1, \ldots, x_n$, we must know the exact moment-generating function (MGF) for each x_i.

   ○ A. TRUE [✓ ○ B. FALSE]

   (d) (1 point) In a Bayesian GLM with a uniform prior, any 95% credible interval for a coefficient has a 95% probability of containing the MAP estimate for that coefficient.

   ○ A. TRUE [✓ ○ B. FALSE]

   (e) (1 point) In a Markov decision process, we assume that the sequence of states is i.i.d. conditioned on the actions taken.

   ○ A. TRUE [✓ ○ B. FALSE]

2. (6 points) A professor teaching a seminar wants to find the best teaching style. Each session, the professor chooses a teaching style and measures student learning with the class's average score on a quiz after class. You should assume the following:

   - Each session, the professor chooses one of three styles: talking, discussion, or visual.
   - Each quiz has 10 questions, so the class's average score will be between 0 and 10.

   For parts (a) - (c), the professor decides to try the multi-armed bandits framework to identify the best style.

   (a) (3 points) For this part only, suppose that the professor chooses a teaching style uniformly at random for every session, regardless of any previous sessions. Will this approach result in linear regret, logarithmic regret, or something else? Justify your answer by computing, approximating, or bounding the regret (expected pseudoregret).

   Solution: Linear regret:

   $$\begin{aligned} R_T &= \mathbb{E}\left[\sum_a T_a(t)\Delta_a\right] \\ &= 0 + \frac{T}{3}\Delta_1 + \frac{T}{3}\Delta_2, \end{aligned}$$

   where Δ_1 and Δ_2 are the optimality gaps for the two suboptimal styles.

   (b) (2 points) The professor has an idea of how the three styles compare to each other, but isn't confident enough to just pick one. Which of the following algorithms, if used by the professor, will produce the lowest regret? Choose the single best answer by filling in the circle next to it. **You must justify your answer to receive credit.**

   ○ A. Explore-then-commit (ETC)

   ○ B. Upper confidence bound (UCB)

   [✓ ○ C. Thomson sampling]

   Justification:

   Solution: The professor can use their idea as a prior, and Thomson sampling achieves lower regret than UCB for a well-chosen prior.

   (c) (1 point) In two sentences or less, explain one assumption of multi-armed bandits that might not hold in this example.

   Hint: Even if there are multiple reasons, you need only describe one to receive full credit.

Solution: Some possible answers: the reward distributions might not be stationary or independent if the class is cumulative, if some topics are harder/easier, or if the professor teaches better/worse early on.

3. (6 points) **Staying Calm**

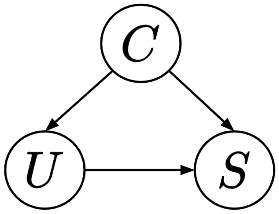
Kit wants to evaluate the effectiveness of a popular meditation app on student stress levels. She collects a dataset from randomly sampled students about their behavior over the last month with the following columns:

- usage (U): whether the student uses the app (1) or not (0)
- stress (S): average stress level over the past month (0-100)
- courseload (C): how many units the student is taking

You should assume that the app came out after the semester started: in other words, students decided on their courseload before they started using or not using the app.

(a) (2 points) Draw a causal DAG for the three variables described above. In two sentences or less, justify any arrows you drew to or from courseload (C).

Solution:



(b) (3 points) Kit decides to use outcome regression to estimate the causal effect of app usage U on stress S: specifically, she treats the estimated coefficient $\hat{\beta}_1$ from the model $S = \beta_1 U + \beta_2 C + \beta_3$ as an unbiased estimate of the ATE. Which of the following assumptions is she making by doing this? Select all answers that apply.

□ A. Conditioned on courseload, app usage is independent of stress.

□ B. Courseload does not have any nonlinear effects on stress.

□ C. Conditioned on courseload, app usage is independent of the potential outcomes for stress.

□ D. Courseload does not have any direct effect on stress; any effect is indirect through app usage.

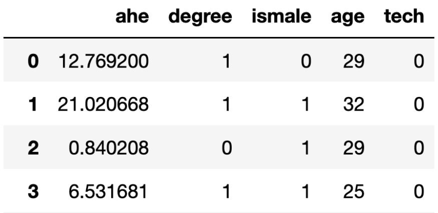
□ E. The students in the sample were randomly assigned to use or not use the app.

□ F. The stable unit treatment value assumption (SUTVA) holds.

(c) (1 point) For this part only, assume that Kit has also collected data on students' heart rate. Give one reason that she shouldn't use this variable as a confounder in her analysis.

Solution: Heart rate is a collider, because app usage (treatment) and stress (outcome) can both have a causal effect on heart rate.

4. (5 points) Ruta looks at a dataset of 583 workers from 2001. Here are the first few rows:

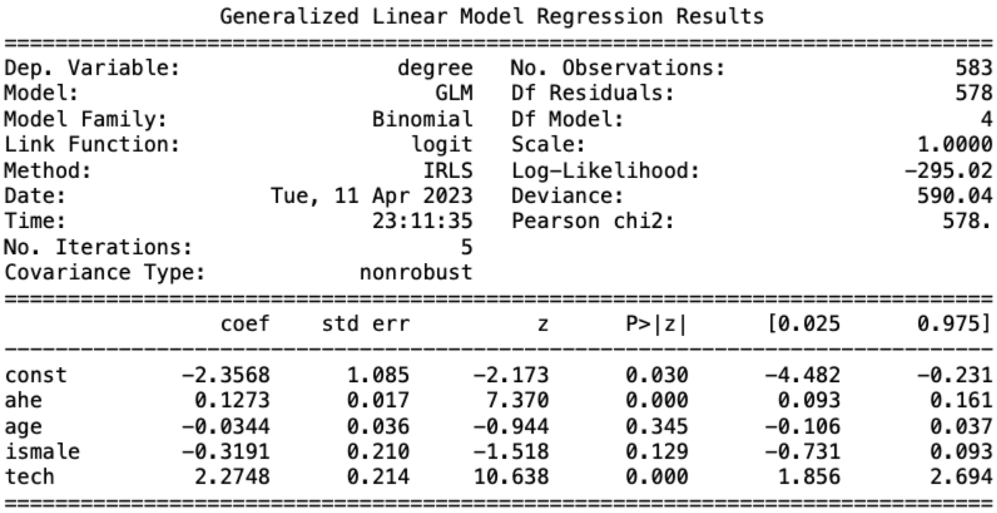


	ahe	degree	ismale	age	tech
0	12.769200	1	0	29	0
1	21.020668	1	1	32	0
2	0.840208	0	1	29	0
3	6.531681	1	1	25	0

The dataset contains the following columns:

- **ahe**: the average hourly earnings (in dollars) of that worker
- **degree**: whether or not the worker has a bachelor's degree
- **ismale**: binarized gender, whether the worker is male (1) or female/other (0)
- **age**: the worker's age in years
- **tech**: whether the worker is in the technology sector (1) or not (0)

Ruta wants to predict whether someone has a degree using the other columns. She takes a frequentist approach and correctly implements the model, obtaining the following results:



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                 degree   No. Observations:                  583
Model:                            GLM   Df Residuals:                      578
Model Family:                Binomial   Df Model:                            4
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -295.02
Date:                Tue, 11 Apr 2023   Deviance:                       590.04
Time:                        23:11:35   Pearson chi2:                     578.
No. Iterations:                     5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.3568      1.085     -2.173      0.030      -4.482      -0.231
ahe            0.1273      0.017      7.370      0.000       0.093       0.161
age           -0.0344      0.036     -0.944      0.345      -0.106       0.037
ismale        -0.3191      0.210     -1.518      0.129      -0.731       0.093
tech           2.2748      0.214     10.638      0.000       1.856       2.694
==============================================================================
```

(a) (3 points) Which of the following are valid interpretations from the results above? Select all answers that apply.

□ A. Negative binomial regression would be a better choice than logistic.

□ B. According to the model, being one year older is associated with about a 0.037 lower probability of having a degree.

□ C. According to the model, there is a strong positive association between working in technology and having a bachelor's degree.

□ D. The results show that the frequentist approach was a better choice than a Bayesian approach for this problem.

□ E. This model is a good fit for the training dataset.

□ F. This model will be a good fit for a similar dataset from 2021.

(b) (2 points) Consider a 25-year-old male worker from 2001 who did not work in tech and whose average hourly earnings were \$18. Based on the model above, write a mathematical expression for the probability that this worker has a degree.

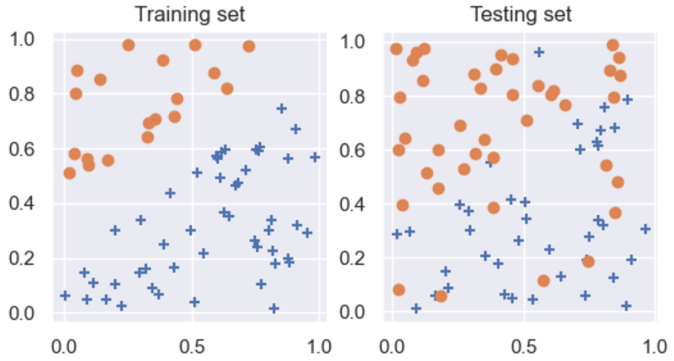
Hint: You do not need to simplify your answer or do any arithmetic, but your answer should not contain any variables or unknowns.

Solution:

$$\text{sigmoid}(-2.3568 + 0.1273 \times 18 - 0.0344 \times 25 - 0.3191)$$

The remainder of this page is for scratch work: no work below this line will be graded.

5. (6 points) For parts (a) and (b), consider the following training and test datasets:



Bea trains different models on the training set (left), and then computes the accuracy (fraction of examples classified correctly) on both the training set and the test set. The models used are logistic regression (with no regularization and an intercept), a decision tree (trained to arbitrary depth), and a random forest.

(a) (1 point) Which of the following model(s) are guaranteed to achieve perfect accuracy on the **training** set? Select all answers that apply.

□ A. Logistic regression

□ B. Decision tree

□ C. Random forest

□ D. None of the above

(b) (2 points) Which of the following model(s) are guaranteed to achieve perfect accuracy on the **test** set? Select all answers that apply.

□ A. Logistic regression

□ B. Decision tree

□ C. Random forest

□ D. None of the above

(c) (3 points) Consider two loss functions:

$$L_a(\theta_1, \theta_2, \theta_3) = \sin(\theta_1\theta_2 + \theta_3) \qquad\qquad L_b(\theta_1, \theta_2, \theta_3) = \theta_1 + \theta_2 + \theta_3$$

For each loss function, Bea does the following:

- Compute the gradient with respect to $(\theta_1, \theta_2, \theta_3)$, once by differentiating naively, and once using backpropagation
- Compute the improvement from backpropagation: how many fewer operations it requires than naive differentiation (in other words, the reduction in the number of operations for backpropagation compared to naive differentiation).

Which loss function will have a bigger improvement? Choose the single best answer by filling in the circle next to it. **You must explain your answer to receive credit.**

Hint: you may find it helpful to draw computation graphs for your explanation.

○ A. L_a

○ B. L_b

○ C. Neither; the improvement is the same for both

Explanation:

Solution:

Computing the gradient of L_a involves using the chain rule to differentiate inside the sin function, and therefore involves redundant computation for naive differentiation. This is eliminated in backpropagation.

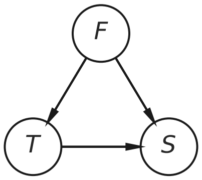
For L_b, there is no redundant computation, and so backpropagation does not offer any improvement.

Name: ______________________________ SID: ______________

6. (8 points) **TutorGPT**

An AI company has developed a new system called TutorGPT to improve student learning, and charges $1000 per semester to each student who uses it. They hire Ada, a data scientist, to determine whether TutorGPT causes students to win (merit-based) academic scholarships.

Ada takes a random sample of 5000 students and measures their family income F (low=0 and high=1), whether they won a merit-based scholarship S, and whether they used TutorGPT T. For this question, assume that **all three variables are binary**, and that the following DAG accurately represents the causal relationships between these variables:



(a) (2 points) Ada computes the difference in scholarship award rates between TutorGPT users and non-users. In two sentences or less, explain why this is not a good estimate for the causal effect the company is interested in.

Solution: This estimate does not take the confounder (family income) into account, and will introduce omitted variable bias.

(b) (2 points) Assuming that there are no other confounders other than family income F, which of the following are true statements about propensity scores for this causal question? Select all answers that apply.

□ A. The propensity score is the conditional probability of using TutorGPT given family income.

□ B. The propensity score is the conditional probability of having high family income given TutorGPT usage.

□ C. The IPW estimator (as described in lecture and lab) will provide an unbiased estimate of the ATE.

For the remainder of the question, assume that Ada used randomized experiments instead of an observational sample. She recruits two random samples:

- One has 200 students with **high** family income:
  - 50 of them are randomly chosen to receive free access to TutorGPT: 10 of these students won a scholarship.
  - The remainder are blocked from using it: 15 of these students won a scholarship.
- The other has 800 students with **low** family income:
  - 160 of them are randomly chosen to receive free access to TutorGPT: 40 of these students won a scholarship.
  - The remainder are blocked from using it: 32 of these students won a scholarship.

You should assume that all students who were given free access used TutorGPT, and all students who were blocked did not use it.

(c) (2 points) Compute the conditional average treatment effect (CATE) of TutorGPT among low family income students.

Solution:

$$\begin{aligned} CATE &= \frac{1}{n}\left(\frac{\sum_T Y_i}{e(T)} - \frac{\sum_C Y_i}{1-e(T)}\right) \\ &= \frac{1}{800}\left(\frac{40}{0.2} - \frac{32}{0.8}\right) = 0.2 \end{aligned}$$

(d) (2 points) Let α be the CATE for low family income students, and β be the CATE for high family income students. Write an expression in terms of α and β for an unbiased estimate of the ATE of TutorGPT on winning the scholarship.

Hint: Do not plug in your answer from part (c) for α; instead, leave your answer in terms of α and β.

Solution:

$$ATE = \frac{800}{1000} * CATE_{low} + \frac{200}{1000} * CATE_{high}$$

7. (4 points) Let $\lambda > 1$, and consider the two random variables $x \sim \text{Poisson}(\lambda)$ and $y \sim \text{Exponential}\left(\frac{1}{\lambda}\right)$.

Hint: You can find some information about the Poisson and Exponential distributions on the reference sheet.

(a) (1 point) Use Markov's inequality to find bounds on $P(X \geq t)$ and $P(Y \geq t)$, and show that the bounds are the same.

Solution:

$$P(X \geq t) \leq \frac{\mathbb{E}X}{t} = \frac{\lambda}{t}$$
$$P(Y \geq t) \leq \frac{\mathbb{E}Y}{t} = \frac{\lambda}{t}$$

(b) (1 point) Use Chebyshev's inequality to find bounds on $\mathbb{P}(|X-\lambda| \geq t)$ and $\mathbb{P}(|Y-\lambda| \geq t)$, and show that the second is equal to λ times the first.

Solution:

$$\mathbb{P}(|X-\lambda| \geq t) \leq \frac{\text{Var}(X)}{t^2} = \frac{\lambda}{t^2}$$
$$\mathbb{P}(|Y-\lambda| \geq t) \leq \frac{\text{Var}(Y)}{t^2} = \frac{\lambda^2}{t^2}$$

(c) (2 points) Why does Markov's inequality produce the same bound while Chebyshev's inequality produces different bounds? Select all answers that apply.

Hint: this question can be answered using only the information from the question statements of parts (a) and (b).

□ A. The exponential distribution is overdispersed.

□ B. The two random variables have the same mean, but one has a higher variance.

□ C. We are comparing a continuous random variable (exponential) to a discrete one (Poisson).

□ D. Chebyshev's inequality provides better (tighter) bounds by using more information about the random variables.