Дорогой храбрый воин или храбрая воительница! Удачи тебе на большом празднике по прикладной статистике! Начни с того, что напиши клятву и подпишись под ней:

Я клянусь честью студента, что буду выполнять эту работу самостоятельно.

А теперь — задачки:

1. Известно, что в среднем за час в Ромашково прибывает λ паровозиков. Дежурный по станции, во всём придерживающийся байесовского подхода, решил оценить параметр λ. Для этого он собрал очень большую выборку $X_1, \ldots, X_n$ моментов прибытия паровозиков. Не будем держать интригу и сразу скажем, что $X_i \sim \text{Pois}(\lambda)$ и все X_i независимы.

   a) Пусть $\lambda \sim \Gamma(\alpha, \beta)$. Покажите, что апостериорное распределение λ также является гамма-распределением.

   Напоминание: плотность гамма-распределения имеет вид

   $$f(x) = \frac{\alpha^\beta x^{\beta-1}}{\Gamma(\beta)} e^{-\alpha x}$$

   при $x \in (0, +\infty)$, $\alpha > 0$, $\beta > 0$.

   b) Постройте 95%-ый байесовский доверительный интервал для λ.

   c) Выведите априорное распределение Джеффриса. Используя его, выведите апостериорное распределение λ.

   d) Для любого из предыдущих пунктов выведите в явном виде какие-нибудь две возможные точечные байесовские оценки для λ.

2. Исследовательница Кларисса считает, что в модели

   $$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

   имеется гетероскедастичность следующего вида: $\text{Var}(\varepsilon_i) = \exp(\alpha_0 + \alpha_1 x_i)$.

   a) Скорректируйте гетероскедастичность и выведите формулу эффективной оценки в явном виде.

   b) Поясните, как построить доверительный интервал, устойчивый к гетероскедастичности, используя стандартные ошибки Уайта.

   c) Сформулируйте гипотезу о гомоскедастичности и найдите оценки неизвестных параметров в предположении о гомоскедастичности методом максимального правдоподобия.

3. Неаккуратный исследователь Иннокентий хочет оценить следующую линейную модель:

$$y_i = \beta_0 + \beta_x X_i + \beta_z Z_i + \beta_m M_i + u_i$$

при помощи МНК. Иннокентий считает, что все регрессоры являются стохастическими с математическим ожиданием μ_j и дисперсией σ_j^2, $j \in \{x, z, m\}$. Внутренний голос говорит Иннокентию, что все регрессоры независимы между собой и со случайной ошибкой, кроме M_i: $\text{Cov}(M_i, u_i) \neq 0$.

К сожалению, при сборе данных Иннокентий часто отвлекался, а потому получилось, что

- Был получен не X_i, а $X_i^* = X_i + \alpha$, α — константа.
- Был получен не Z_i, а $Z_i^* = Z_i + \nu$, ν — случайная величина с математическим ожиданием 0 и дисперсией σ_ν^2.

Иннокентий не заметил ошибок при сборе данных, а потому оценивает регрессию

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_x X_i^* + \hat{\beta}_z Z_i^* + \hat{\beta}_m M_i.$$

a) Найдите предел при вероятности оценок $\hat{\beta}_x$, $\hat{\beta}_z$, $\hat{\beta}_m$. Прокомментируйте, является ли каждая из оценок состоятельной.

b) Для каждой несостоятельной оценки из предыдущего пункта предложите корректировку, которая сделала бы её состоятельной.

c) Иннокентий подозревает, что в модели есть проблема эндогенности. Проведя в поисках четыре дня, Иннокентий нашёл четыре переменные Q_i, коррелирующие со всеми регрессорами в его модели и при этом не зависимые от случайной ошибки. Выведите оценки двухшагового МНК для модели Иннокентия.

4. Исследователь Винни-Пух использует две модели, описывающие вектор $y = (y_1, y_2, \ldots, y_n)$. Одна модель подсказана Совой, вторая — Кроликом. Как известно, у Винни-Пуха опилки в голове, поэтому обе модели содержат $k = 0$ параметров.

Величины y_i в обеих моделях и в реальности независимы и одинаково распределены.

Докажите, что величина $\hat{\Delta} = (AIC_{\text{Кролик}} - AIC_{\text{Сова}})/2$ состоятельно оценивает $\Delta = KL(p||p_{\text{Кролик}}) - KL(p||p_{\text{Сова}})$.

Здесь p — реальное распределение вектора y, а $p_{\text{Кролик}}$ и $p_{\text{Сова}}$ — модельные.

5. Исследовательница Мадлен проводит снижение дисперсии с помощью преобразования CUPED:

$$X_{\text{cuped}} = X_{\text{post}} - \theta(X_{\text{pre}} - \mathrm{E}(X_{\text{pre}})).$$

Здесь X_{post} — метрика после начала эксперимента, а X_{pre} — метрика до начала эксперимента.

а) Явно напишите, какая целевая функция оптимизируется при подборе θ.

б) Выведите формулу для оптимального θ.

в) Постройте график зависимости отношения дисперсий $\mathrm{Var}(X_{\text{cuped}})/\mathrm{Var}(X_{\text{post}})$ от корреляции $\rho = \mathrm{Corr}(X_{\text{pre}}, X_{\text{post}})$.

6. Априорно исследователь Аверкий считает, что вероятность дождя имеет ожидание равное $1/3$ и дисперсию $1/32$.

Затем Аверкий выбирает 10 случайных дней и 5 из них оказываются дождливыми.

а) Выберите подходящее удобное априорное распределение.

б) Постройте 90%-ый апостериорный байесовский интервал для вероятности дождя.

в) Друг Аверкия Аркадий считает, что вместо байесовского подхода можно было получить тот же результат в рамках классического подхода. Аркадий предлагает заменить априорное распределение на дополнительные фиктивные наблюдения и использовать метод максимального правдоподобия. Сколько фиктивных дней наблюдений нужно добавить, и сколько из них должны быть дождливыми, чтобы точечная оценка Аркадия совпала с апостериорным ожиданием Аверкия?