Домашнее задание №2.

Домашнее задание сдается в электронном виде в SmartLMS.

Решение задания 1 должно быть оформлено в виде pdf-файла. В каждом пункте необходимо прописывать используемые формулы. К заданию 1 необходимо приложить код. Решение задания 2 должно быть оформлено в Jupiter Notebook. Все комментарии и описания функций, если того требует пункт задания, должны быть оформлены в текстовых ячейках. Pdf-файл для данного задания не требуется.

Дедлайн: 30 апреля 2025 года, 23:59.

1. **(40 баллов)** Файл $hw\_2.xlsx$ содержит данные по арестам в течение 1986 года по мужчинам, родившимся в Калифорнии в 1960 или 1961 гг. Каждый из мужчин в выборке был арестован по крайней мере однажды до 1986 г. Вопрос: что объясняет, что мужчина снова был арестован в течение 1986 г. (сколько раз, и т.д.)? Имеются следующие данные:

   $narr86$ — количество арестов в 1986 г.
   $pcnv$ — доля предыдущих арестов, приведших к осуждению (прокси для неотвратимости наказания)
   $avgsen$ — среднее время срока заключения по предыдущим случаям осуждения (в месяцах) (прокси для суровости наказания)
   $tottime$ — общее время, проведенное в тюрьме после достижения возраста 18 лет (в месяцах)
   $ptime86$ — число месяцев в тюрьме во время 1986 г. (не может быть арестован, пока в тюрьме)
   $qemp86$ — количество кварталов, в которых имел работу, в течение 1986 г. (возможности на рынке труда)
   $inc86$ — легальный доход в 1986 г., $100
   $durat$ — длительность последнего периода безработицы (в месяцах)
   $black$ =1, если афроамериканец
   $hispan$ =1, если латиноамериканец
   $born60$ =1, если родился в 1960 г.
   $crime86$ =1, если был арестован хотя бы однажды в 1986 г. (genr crime 86 = narr86>0)

   Указание: во всех пунктах используйте 5%-й уровень значимости.

   (a) (2 балла) Оцените $logit$–модель бинарного выбора для переменной $crime86$. В качестве объясняющих факторов используйте следующие показатели: $constant$, $pcnv$, $avsgen$, $tottime$, $ptime86$, $qemp86$, $inc86$, $durat$, $black$, $hispan$, $born60$. Выпишите оценку модели.

   (b) (1 балл) С помощью оцененной в пункте (a) модели для каждого индивида постройте прогноз вероятности того, что он будет ни разу арестован в 1986 г.

(c) (5 баллов) Используя модель из пункта (a), оцените средние предельные эффекты для вероятности быть арестованным по всем факторам и проинтерпретируйте их. Предельные эффекты для бинарных переменных рассчитайте вручную как разность условных вероятностей. Выпишите используемые формулы предельных эффектов для непрерывных и бинарных переменных.

(d) (3 балла) Используя модель из пункта (a), оцените отношение шансов для всех факторов и проверьте их значимость. Проинтерпретируйте полученные результаты.

(e) (4 балла) Оцените среднее различие вероятности ареста между афроамериканцем и латиноамериканцем для мужчины, родившегося в 1961 г., со средними показателями остальных характеристик ($pcnv$, $avgsen$ и др.), используя модель из пункта (a). Проверьте значимость этой разницы, указав используемую статистику.

(f) (2 балла) Оцените модель упорядоченного выбора $ordered\ logit$ для переменной $narr86$. В качестве объясняющих показателей используйте следующие: $constant$, $pcnv$, $avsgen$, $tottime$, $ptime86$, $qemp86$, $inc86$, $durat$, $black$, $hispan$, $born60$. Выпишите оценку модели.

(g) (1 балл) С помощью оцененной в пункте (f) модели для каждого индивида постройте прогноз вероятности того, что он не будет ни разу арестован в 1986 г.

(h) (3 балла) Для модели из пункта (f) проверьте гипотезу о параллельности (parallel regression assumption). Укажите проверяемую гипотезу и используемую статистику.

(i) (5 баллов) Запишите формулу предельного эффекта j-го фактора на вероятность того, что совершено k преступлений. Используя модель из пункта (f), постройте график зависимости этих средних предельных эффектов от числа арестов в 1986 г. Добавьте на график границы доверительных интервалов для оценнных средних предельных эффектов. Проинтерпретируйте полученные результаты.

(j) (2 балла) Оцените модель множественного выбора $multinomial\ logit$ для переменной $narr86$. В качестве объясняющих показателей используйте следующие: $constant$, $pcnv$, $avsgen$, $tottime$, $ptime86$, $qemp86$, $inc86$, $durat$, $black$, $hispan$, $born60$. Выпишите оценку модели.

(k) (1 балл) С помощью оцененной в пункте (j) модели для каждого индивида постройте прогноз вероятности того, что он не будет ни разу арестован в 1986 г.

(l) (3 балла) Для модели из пункта (j) проверьте гипотезу о независимости от

посторонних альтернатив. Укажите проверяемую гипотезу, используемую статистику и сделайте выводы.

(m) (3 балла) Для модели из пункта (j) рассчитайте предельные эффекты и проверьте их значимость. Сделайте выводы.

(n) (2 балла) Для модели из пункта (j) рассчитайте отношение шансов и проверьте их значимость. Прокомментируйте.

(o) (3 балла) Сравните точность предсказания не быть ни разу арестованным по трём оценённым моделям. Опишите подробно, какие критерии использованы при сравнении, а также в чём заключаются преимущества и недостатки каждой из модели.

2. **(60 баллов)** В данном задании вам предстоит работать с симулированными данными. В качестве инструмента для выполнения рекомендуется использовать Python. К работе нужно будет приложить Jupyter Notebook, в котором помимо расчётов будут комментарии к коду, используемые формулы и развёрнутые ответы на вопросы (в pdf-файл это задание оформлять не нужно).

   (a) (5 баллов) Задайте объём выборки n (от 500 до 1000 наблюдений). Сгенерируйте:

   - коррелированные признаки x_2 и x_3, указав использованные распределения;
   - латентную переменную y^* для $logit$-модели, самостоятельно задав значения свободного коэффициента β_1 и коэффициентов β_2 и β_3 при признаках x_2 и x_3 соответственно;
   - зависимую бинарную переменную y.

   (b) (10 баллов) Напишите функцию, реализующую численное нахождение ММП-оценок для модели из пункта (a). Рассчитайте оценки численно, используя симулированные данные. В Python это можно сделать с помощью optimize из библиотеки scipy. Сравните полученные с помощью вашей функции ММП-оценки параметров с готовой функцией из Python.

   Для выполнения дальнейших пунктов задания вам будет полезна теоретическая справка по матричным статистикам LR, LM и Wald из Семинара 22.

   (c) (10 баллов) Напишите функцию, рассчитывающую предельные эффекты

$$\frac{\partial}{\partial x_{ij}}\mathbb{P}(y_i = 1|x_{i2}, x_{i3}),$$

$$\frac{\partial}{\partial x_{ij}}\mathbb{P}(y_i = 0|x_{i2}, x_{i3}),$$

а также их стандартные ошибки, z-статистики для проверки гипотезы о незначимости предельных эффектов и $p - value$. Оформите вывод, схожий с выводами результатов оценивания $logit$-регрессии из реализованных в Python функций. Рассчитайте их для симулированных данных.

(d) (30 баллов) Напишите **три** функции, позволяющие тестировать гипотезу о параметрах $logit$-модели с помощью теста Вальда (W), отношения правдоподобия (LR) и множителей Лагранжа (LM). Используйте матричные записи из Семинара 22. Подробно опишите каждую из функций (аргументы функции, возвращаемые ею значения и тд).

(e) (3 балла) Используя написанные вами функции, с помощью тестов Вальда, отношения правдоподобия и множителей Лагранжа на уровне значимости 5% на симулированных данных проверьте гипотезу вида $\beta_j = k$, где j и k можно выбрать произвольным, указав выбранное значение. k должно быть отлично от истинного значения параметра.

Укажите расчетное и критическое значения статистик, $p - value$ каждого теста. Сделайте выводы на 5%-ом уровне значимости. Сравните с результатами работы готовых функций в Python.

(f) (3 балла) Используя написанные вами функции, с помощью тестов Вальда, отношения правдоподобия и множителей Лагранжа на уровне значимости 5% на симулированных данных проверьте гипотезу вида $\beta_j + \beta_m = k$, где j, m, k можно выбрать произвольным, указав выбранное значение. k должно быть отличным от истинного значения параметра. Укажите расчетное и критическое значения статистик, $p-value$ каждого теста. Сделайте выводы на 5%-ом уровне значимости. Сравните с результатами работы готовых функций в Python.

(g) (4 балла) Запишите основную гипотезу из предыдущего пункта в любом альтернативном виде. Протестируйте её заново с помощью теста Вальда. Сравните полученные результаты с предыдущим пунктом. Прокомментируйте полученный результат.