toggle
made for free at coggle.it

Статистическая мощность. Чем выше мощность теста, тем меньше вер-ть совершить ошибку 2 рода
Уровень значимости - вероятность ошибки 1 рода

Часть 1

Основы статистики

Часть 2

Основные ограничения, которые накладываются на методы регрессионного анализа

Нормальное распределение остатков

При нарушении данные идут вверх, сначало много наблюдений и потом, вплоть к разбросу

Гомоскедастичность

Допустим мы построили регрессионную модель с независимой переменной X и зависимой переменной Y. Однако, условно Вы использовали более надежные данные, которые не содержат выбросов. Давайте подумаем, что в этом примере позволит нам ответить на вопрос, есть ли в наших данных нарушение предположения о гомоскедастичности (или гетероскедастичности). Если мы построим регрессию, где зависимой переменной (Y) будут квадраты остатков, а независимой переменной будет предиктор x, и в этой модели окажется статистически значимая связь, это означает, что в данных есть гетероскедастичность.

Независимость предикторов (проблема мультиколлинеарности)

Если в данных есть мультиколлинеарность, то модель будет нормально предсказывать, НО на новых (вне) всех строчках значения (например отрез, так. где этого не может быть)

Способы не решить

Построить корреляционную матрицу

Можно узнать несколько хорошо выбранный предиктор объясняется другим(ТОТ) борись но VIF показывает во сколько раз увеличивается дисперсия коэффициента по сравнению с ситуацией, если бы данный предиктор не был связан с другими

Формула дисперсии коэффициента

второй множитель и есть VIF

Отсутствие автокорреляции остатков

Нарушения, которые мы встречаем в исследовании

Повторные измерения

Проблемы

Сочетание чрезвычайности теста

Искусственное увеличение вводного пострандомизации

Повторные пробы

Искажение россов

Кластеризация данных

Искажение россов

Независимость наблюдений

Гетероскедастичность

Нелинейность

Нормальность распределения остатков

Гистограмма остатков

Расстояние Кук-квадрат

$F^2 \geq [2.25 - 52 / 51.]$

Критерий Дарбина

Распределение Фишера

Распределение суммы квадратов в независимых стандартных (со средним 40 + 6 и дисперсией 50 + 2)

случайных величин, распределенных нормальным образом.

Распределение F с k степенями свободы

чем больше число степеней свободы у распределения хи-квадрат, тем более симметричным становится такое распределение, и тем ближе распределение по квадрат стремится к нормальному.

Критерий Уилкоксона

Анализ номинальных данных

Анализ таблиц сопряженности

Условия применимости критерий χ^2 при анализе таблиц сопряженности

Все наблюдения должны быть независимы

Минимальное значение в каждой ячейке равно 5

Критерий Пирсона

Критерий Фишера

Для маленьких выборок

Используется для анализа таблицы сопряженности две на две, где значения находились в наблюдаемых частот в каждой ячейке больше 5, но меньше 10

$F^2 \leq [3.25 - 35 / 33.6^2 + 35]$

Для маленьких и больших выборок

Используется для анализа таблицы сопряженности две на две, где значения находились в наблюдаемых частот в каждой ячейке больше 5, но меньше 10

$F^2 \leq [3.25 - 35 / 33.6^2 + 35]$

Для маленьких выборок

Используется для анализа таблицы сопряженности две на две, где значения находились в наблюдаемых частот в каждой ячейке больше 5, но меньше 10

$F^2 \leq [3.25 - 35 / 33.6^2 + 35]$

Для больших выборок

Используется для анализа таблицы сопряженности две на две, где значения находились в наблюдаемых частот в каждой ячейке больше 5, но меньше 10

$F^2 \leq [3.25 - 35 / 33.6^2 + 35]$

Интерпретация остатков

Если значение стандартизированных остатков больше 3х, можно считать, что в этой ячейке зафиксирована значимое отклонение.

Анализ остатков (разница между ожидаемыми и эмпирическими значениями) позволяет выявить, какие именно частоты значимо отклоняются от ожидаемых значений

Логистическая регрессия

Используется для исследования взаимосвязи между номинальной зависимой переменной, имеющей всего 2 градации, и различными независимыми переменными (и качественными предикторами, и количественными).

Номинальная переменная с двумя градациями можно представить как две зависимых двоичных переменных, которые равны p и (1 - p)

Логарифм шанса (log odds) = log(p/(1 − p)) = ln(p/(1 − p)) если log odds = 0 , то p = 1 - p

Odds (шансы) = p / (1 - p) - это отношение вероятности успеха (p) к вероятности неудачи (1 - p) = p/(1 − p)

Probability (вероятность) = p = odds / (1 + odds) = 2 / (1 + 2) = 2/3 ≈ 0.67

Для модели: $P = (e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k}) / (1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k})$, мы хотим узнать вероятность успеха, используя только одно значение x_1 (возраст), тогда $P = (e^{\beta_0 + \beta_1 x_1}) / (1 + e^{\beta_0 + \beta_1 x_1})$

Модель без предикторов описывается уравнением: $y = \hat{}$ (частота) (оценка числа, без предикторов)

Интерсепт - натуральный логарифм шансов положительного исхода. И наоборот, если интерсепт отрицательный, то шансы исхода ниже 50%

Интерсепт для нулевой таблицы

Estimate - значение интерсепт (натурафон шансов)

Пояснение к расчетам стандартная ошибка (se) - это стандартное отклонение, которое имело бы нормальное распределение коэффициента интерсепт (в среднем равным 0

Тогда если разделить значение коэффициента интерсепт на стандартную ошибку, мы получим z-value - расстояние до 0 в стандартных отклонениях.

Link Title

Таблица зависимостей шансов выжить с мужчин и женщин

Female Male
No 64 368
Yes 197 93

Модель с одним номинальным предиктором

odds_male = 93 / 368 = 0.25

odds_female = 197 / 64 = 3.08

Интерсепт - натуральный логарифм шансов положительного исхода для женщин

log(odds_female) = 1.13

коэф-т при X - натуральный логарифм отношения шансов положительного исхода для мужчин к шансу для женщин (со зн. разность логарифмов шансов)

log(odds_male) = log(odds_female) + log(odds_ratio)

log(odds_ratio) = -2.48

По свойству логарифмов: log(odds_ratio) = log(odds_male) - log(odds_female)

Link Title

Т.к. переменные Female и Male зависимы друг от друга (мультиколлинеарны), решение в построении модели заключается в том, что мы должны выбрать только один из методов, чаще всего это alphabetical, в нашем примере, значением интерсепт. Поэтому, смотрим, какой градации предиктора нет среди коэффициентов, это и будет интерсепт

log(odds_female) = 1.13 ; 3.48 (ln_Male = 1.13 ; 2.48 0 = 1.13

log(odds_male) = 1.13 ; 3.48 ln(_Male = 1.13 ; 2.48 1 = 1.35

В отличие от теста у Пирсона, логистическая регрессия не может использовать непрерывные переменные, только режимы шансы для разных градаций независимой переменной.

Модель с двумя номинальными предикторами

Уравнение регрессии

log(odds) = 1.13 − 2.77isMale − 0.88*PclassSecond − 1.89*PclassThird - 0.42*SexMale*PclassSecond + 2.07*SexMale*PclassThird

Таблицы шансов/исход шансов выжить у мужчин и женщин в зависимости от класса

По графику

Интерактивные таблицы

Непараметрические методы

Статистический Манна-Уитни

Непараметрический аналог критерия Стьюдента. Используется для оценки различий между двумя независимыми выборками по уровню какого-либо количественного или ранговой шкалы.

Самым популярным непараметрическим критерием для оценки различий между двумя выборками по уровню какого-либо признака является U-тест Манна-Уитни. Он позволяет выявлять различия между малыми выборками, когда уже не срабатывает критерий Стьюдента, но различия между ними уже довольно выражены.

Если при сравнении трех и более групп наблюдается тенденция к статистически достоверной и закономерности распределений, лучше применять непараметрический аналог дисперсионного анализа — критерий Краскела-Уоллиса. Когда данных немного

Критерий Краскела-Уоллиса

Если для каждой из выборок рассчитывается показатель $\chi^2_r = \frac{12}{N \cdot k \cdot (k+1)} \cdot (n_1 \cdot R_1^2 + n_2 \cdot R_2^2 + \ldots + n_k \cdot R_k^2) - 3 \cdot N \cdot (k+1)$, где $n_1, n_2, \ldots, n_k$ — объемы выборок, $R_1, R_2, \ldots$ — суммы рангов в двух группах.

Неизменные из полученных значений и выступает в качестве статистики теста. Легко показать, что при условии верности нулевой гипотезы распределение этой статистики практически не отличается от распределения с параметром степеней свободы $k = 2$: $\chi^2_r \approx \chi^2_{k-1}$

Если $p \leq \chi^2_{кр}$, то H_0 отвергается, различия между группами признаются статистически значимыми.

Критерий Уилкоксона

Критерий Вилкоксона

Критерий Манна-Уитни

Критерий Стьюдента

Критерий Фишера

Критерий Пирсона

Критерий Фридмана

Критерий Кохрена

Критерий Мак-Немара

Критерий Розенбаума

Критерий Ансари-Брэдли

Критерий Муда

Критерий Вилкоксона-Уилкоксона

Критерий Ван дер Вардена

Критерий Ван дер Вардена

Критерий Ван дер Вардена