ТЕХНОСФЕРА

Занятие №8

Поиск и определение зависимостей

Сергей Москвин

1. Некоторые факты из прошлой лекции
2. Ковариация и корреляция
3. Меры сходства
4. ANOVA
5. Линейная регрессия

Зависимые и независимые случайные величины

Для независимых событий:

- $P(A \cap B) = P(A) * P(B)$

Для независимых случайных величин:

- $F_{X,Y}(x, y) = F_X(x)F_Y(y)$
- $p_{X,Y}(x, y) = p_X(x)p_Y(y)$, если величины абсолютно непрерывны

Тесты на сопряженность

Тесты

- Критерий Хи-квадрат
- Точный тест Фишера / тест Барнарда

Effect size

- Phi
- Cramer’s V

notebook Chi-squared

Ковариация

$$Cov(X,Y) = E[(X - E[X])(Y - E(Y)] = E[XY] - E[X]E[Y]$$

$$Cov(X,X) = Var(X)$$

Несмещенная выборочная оценка:

$$Q = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})$$

- Не нормализованная величина
- Является мерой линейной связи

Корреляция

Коэффициент корреляции Пирсона:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$$

Выборочная оценка:

$$r = \frac{\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})}{\sqrt{\sum_{n=1}^{N}(x_n - \bar{x})^2}\sqrt{\sum_{n=1}^{N}(y_n - \bar{y})^2}}$$

Выборочная оценка является смещенной

Корреляция

Выборочную оценку корреляции можно проверить на значимость:

- Permutation test
- Bootstrap
- t-аппроксимация: $t = r\sqrt{\frac{n-2}{1-r^2}}$ при $H_0: r = 0$ имеет распределение Стьюдента с n-2 степенями свободы

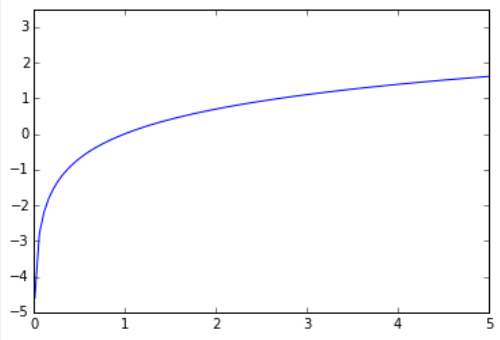
Для тестирования гипотез типа $H_0: r = x$ также применяется трансформация Фишера: $F(r) = \frac{1}{2}\ln\frac{1+r}{1-r}$, $Z = |F(r) - F(P_0)|\sqrt{n-3}$

Mandatory

Correlation ≠ Causation

Альтернативные меры корреляции

- Коэффициент корреляции Пирсона описывает линейную зависимость случайных величин
- Линейная зависимость – не единственно возможная:



$r = 0.85$

- Для определения степени монотонности зависимости используют альтернативные непараметрические методы:
  - Коэффициент ранговой корреляции Спирмена
  - Тау Кенделла

notebook Nonparametric measures

Альтернативные меры корреляции

Выбор между корреляцией Спирмена и корреляцией Пирсона во многом аналогичен другим случаям выбора между параметрическими и непараметрическими методами

Корреляция Пирсона:

- Степень линейной зависимости
- Не является робастной (неустойчива к выбросам)

Ранговая корреляция Спирмена:

- Степень монотонной зависимости
- Устойчива к выбросам
- Выборочная корреляция имеет большее смещение

Mutual information

- Виды зависимости не исчерпываются монотонной зависимостью
- Если мы хотим зафиксировать другие виды зависимости, нам придется использовать более сложные метрики

Для дискретных:

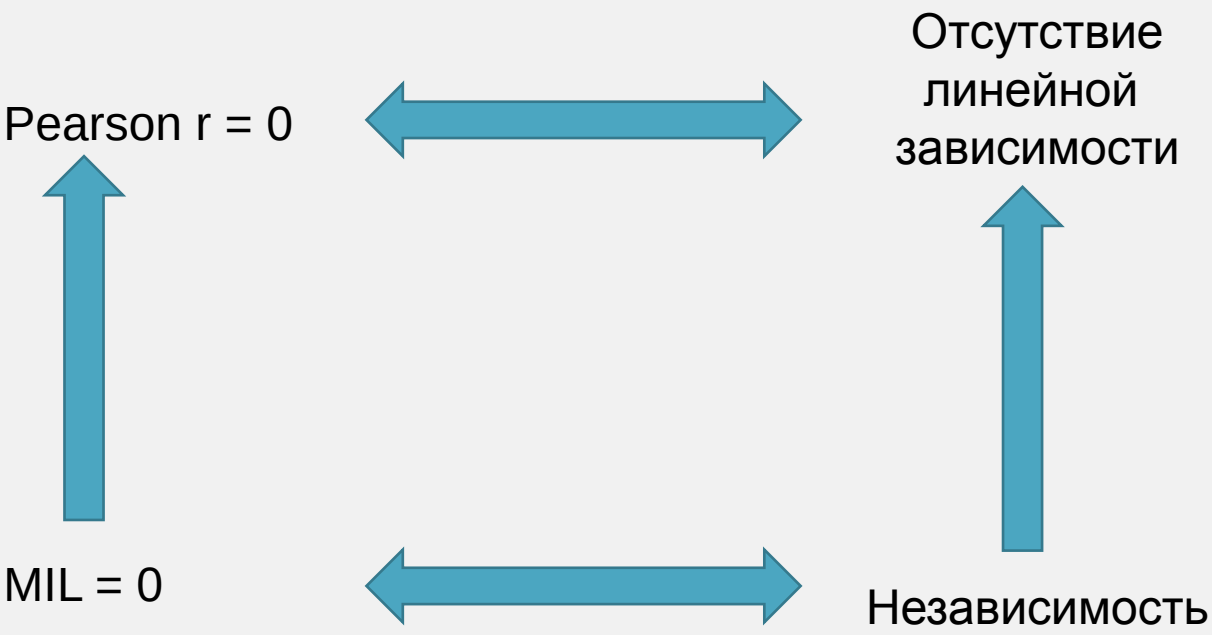
$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

Для непрерывных:

$$I(X;Y) = \int_Y \int_X p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) dx\, dy$$

notebook Dependency Measures

Итоги по метрикам зависимости:



Меры сходства для бинарных переменных

		B	
		0	1
A	0	M_{00}	M_{10}
	1	M_{01}	M_{11}

Мера Жаккара (Jaccard Index)

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} = \frac{M_{11}}{M_{10} + M_{01} + M_{11}}$$

Простая мера схожести (Simple similarity coefficient)

$$SMC = \frac{M_{11} + M_{00}}{M_{10} + M_{01} + M_{11} + M_{00}}$$

notebook Similarity

F-тест и F-распределение

Распределение

$$F \sim \frac{U_1/d_1}{U_2/d_2}$$

Где U_1, U_2 - независимые случайные величины имеющие распределение хи-квадрат с d_1, d_2 степенями свободы

F-test - тест, статистика которого имеет F-распределение

ANOVA

Семейство тестов ANOVA (Дисперсионный анализ) - тесты, проверяющие гипотезы о влиянии категориальных факторов на среднее значение при помощи анализа диспресий по подгруппам

notebook ANOVA

Линейная регрессия

Допустим, что Y имеет линейную зависимость от набора переменных $X_1 \dots X_n$

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_n X_n + \varepsilon$$

Задача: у нас есть выборка из совместного распределения Y, $X_1 \dots X_n$. Как нам по ней оценить параметры $b_0 \dots b_n$ исходной модели?

Метод наименьших квадратов

Ищем значения параметров $b_0 \dots b_n$,
минимизирующие сумму квадратов отклонений:

$$\sum_i (y_i - b_0 - b_1 X_1 \dots - b_n X_n)^2$$

Решение единственно, в матричной форме его можно выразить так:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

notebook OLS

Теорема Гаусса-Маркова

Насколько хороши оценки OLS?

Если выполняются следующие условия:

- Модель правильно специфицирована: зависимость действительно линейная, включены все необходимые объясняющие переменные
- Дисперсия всех ошибок одинакова
- Все ошибки некоррелированны
- Мат.ожидание ошибок равно нулю

, то оценки, полученные путем OLS являются оптимальными (имеют наименьшую дисперсию) в классе линейных несмещенных оценок (Best Linear Unbiased Estimator)

Применение линейной регрессии

Линейная регрессия чаще всего применяется в двух случаях

1) Предсказание (нужна научиться точно предсказывать Y)

2) Объяснение взаимосвязи данных (нужно точно оценить коэффициенты регрессии $b_0 \dots b_n$)

Разновидности регрессии

- Без свободного члена
- С взаимодействием
- С преобразованием переменных

Отклонения от предпосылок

- Модель неправильно специфицирована
- Мультиколлинеаность
- Гетероскедастичность
- Автокорреляция

Неправильная спецификация

Выбраны неправильные объясняющие переменные или неправильное уравнение регрессии (степени, взаимодействие, etc)

В результате:

- Предсказательная способность падает, особенно на данных с другой структурой зависимости переменных
- Коэффициенты регрессии остаются валидными только в случае независимых регрессоров

Что делать?

- Ничего
- Искать правильную спецификацию

Гетероскедастичность

Дисперсия ошибок неодинакова (например, зависит от одного из регрессоров)

В результате:

- Оценки OLS остаются несмещенными, но перестают быть эффективными
- Тесты на значимость в регрессии перестают быть валидными

Что делать?

- Ничего
- Трансформация данных

Мультиколлинеарность

Существует линейная зависимость между переменными модели

В результате:

- Из-за того что изменчивость некоторых переменных частично совпадает, эффективный объем выборки снижается
- Повышается дисперсия оценок
- Понижается мощность

Что делать?

- Ничего
- Поменять выбор переменных

Автокорреляция

Для временных рядов: существует временная корреляция между ошибками модели

В результате:

- Оценки OLS остаются несмещенными, но перестают быть эффективными
- Тесты на значимость в регрессии перестают быть валидными

Что делать?

- Ничего
- Воспользоваться моделями автокорреляции (например, ARIMA)

notebook Time_Series

Выбор модели:

Stepwise regression (не рекомендуется)

- Пробуем поочерёдно добавлять большое количество регрессоров
- Оставляем те, которые дают прирост $R^2{}_{adj}$ и являются значимыми

Cross-Validation

Оставляем все переменные, ничего не убираем

Регуляризация

ТЕХНОСФЕРА

Спасибо за внимание!

Сергей Москвин

smos@list.ru