# 05.09.2017

Демешев Борис Борисович

Эконометрика

8-903-787-34-22

Шаболовка 26, 2112

bdemeshev.github.io/em611

RStudio

Письменная контрольная работа (30%), ДЗ (20%), экзамен (50%)

Учебник: Типширани,

Задачи:

* Прогнозирование (как в машинном обучении)
* Причинно-следственные связи (интерпретировать зависимость, посчитать предельный эффект)

Оглавление

[05.09.2017 1](#_Toc499763325)

[12.09.17 6](#_Toc499763326)

[Вариант со случайной матрицей и вектором 8](#_Toc499763327)

[26.09.17 10](#_Toc499763328)

[Дискретные распределения: 10](#_Toc499763329)

[Биномиальное распределение: 10](#_Toc499763330)

[Геометрическое распределение: 11](#_Toc499763331)

[Отрицательное биномиальное распределение 12](#_Toc499763332)

[Дифференциальные формы 13](#_Toc499763333)

[Упражнение на дифференциальные формы 14](#_Toc499763334)

[Бета-распределение 15](#_Toc499763335)

[Семинар 17.10.17 16](#_Toc499763336)

[Примеры проверки гипотез 17](#_Toc499763337)

[Семинар 07.11.17 19](#_Toc499763338)

[Теорема Гаусса-Маркова 19](#_Toc499763339)

[Доказательство теоремы Гаусса-Маркова 20](#_Toc499763340)

[Упражнение 21](#_Toc499763341)

[Хорошие свойства МНК 23](#_Toc499763342)

[Теорема 1. О распределении 23](#_Toc499763343)

[Теорема 2. О состоятельности 23](#_Toc499763344)

[Гетероскедастичность 23](#_Toc499763345)

[Лекция 14.11.17 28](#_Toc499763346)

[Тесты на гетероскедастичность 28](#_Toc499763347)

[Тест Goldfeldt-Quandt. 28](#_Toc499763348)

[Тест Уайта (Бройша-Лагана) 29](#_Toc499763349)

[Упражнение 30](#_Toc499763350)

[Метод максимального правдоподобия в примерах 31](#_Toc499763351)

[Информация Фишера 32](#_Toc499763352)

[Упражнение 32](#_Toc499763353)

[Семинар 21.11.17 34](#_Toc499763354)

[Тесты 37](#_Toc499763355)

[Решение 38](#_Toc499763356)

[28.11.2017 40](#_Toc499763357)

[Пример 40](#_Toc499763358)

**Обозначения (международные):**

Вектор – маленькая буква, матрица – большая

– зависимая переменная

количество переменных

Объясняющие (зависимые) переменные - штук (с константой):

**Метод наименьших квадратов**

Модель:

– ошибка

неизвестные константы

МНК:

*Пример:*

|  |  |
| --- | --- |
| *задач* | *пирожков* |
| *1* | *1* |
| *2* | *2* |
| *2* | *3* |
|  |  |

– оценка, - истинное значение.

В общем случае:

## Геометрический смысл МНК

Другая визуализация – векторы в трехмерном пространстве

Минимизируем величину:

*–* расстояние

Минимальное расстояние – если проецировать на

*Вариант с несколькими переменными:*

Линейная оболочка

**Вектор ответов регрессии обязательно лежит в линейной оболочке, так как является линейной комбинацией регрессоров.**

**МНК минимизирует квадрат длины вектора ошибок. Если вектор остатков не ортогонален линейной оболочке регрессоров, то эта величина не минимальна.**

Геометрический смысл МНК: - проекция на линейную оболочку векторов регрессоров.

ортогонален линейной оболочке

Через геометрические соображения можно красиво вывести формулу для

Условие существования обратной матрицы:

Утверждение:

плохая ситуация!

Одно наблюдение – бесконечное множество решений

## Термины МНК

вектор остатков (residuals)

error term

dependent variables

– predictions

По теореме Пифагора:

Принятые названия компонентов суммы:

Предельные случаи:

1. RSS=0 – все прогнозы равны прогнозируемым переменным.
2. ESS=0. т. е. все прогнозы одинаковы

Коэффициент детерминации:

# 12.09.17

Предположения про

1. – неизвестная константа
2. – наблюдаемые константы или случайные величины
   1. Например, влияние количества пестицидов на урожай картошки. Если нет рандомизации, то – константы
   2. Например, зависимость посещаемости от номера семинара
   3. Например, выручка от продажи карасей разного веса. Караси могут попасться разные, лучше считать случайной величиной (нет гарантированной воспроизводимости)
   4. При неслучайных – единственное требование:
   5. При случайных :
      1. Простейший случай: наблюдения – случайная выборка из некоторого множества. Можно считать, что – независимые, одинаково распределенные случайные величины. Следовательно,
      2. Временные данные – есть связь близких моментов времени
      3. Пространственные данные – есть связь близких точек пространства

|  |  |
| --- | --- |
| Предпосылки про | Предпосылки про |
| детерминистические |  |
| – одинаково распределенные независимые | – одинаково распределены  Пары независимы друг от друга |

*Пример на условное математическое ожидание*

|  |  |  |
| --- | --- | --- |
|  | -1 | 1 |
| 1 | 0,2 | 0,3 |
| 2 | 0,3 | 0,2 |

При заданной переменной вероятности пропорционально увеличиваются

Следствия из предпосылок:

1. Если существует, то линейная по
2. Если существует,
   1. При детерминистических - является несмещенной оценкой ()
   2. При случайных – оценка несмещенная и условно несмещенная ()
3. Дисперсия
   1. При детерминированных

()

* + 1. Если то – состоятельность оценки

Неравенство Чебышева

## Вариант со случайной матрицей и вектором

Тождества для безусловных величин

y,z, w – случайные вектора, А, b – константная матрица и константный вектор

Тождества для условных величин

(превратили случайное в константу)

**Упражнение:**

Предпосылки:

Все случайные вектора:

Считаем их матожидания и ковариации каждого с каждым:

# 26.09.17

В предыдущей серии:

## Дискретные распределения:

|  |  |  |
| --- | --- | --- |
| X | 0 | 1 |
| P | 1-p | P |

- распределение Бернулли Bern(p)

### Биномиальное распределение:

### Геометрическое распределение:

Ждем до первого успеха.

– число испытаний (или при другом подходе число неуспехов)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| N | 1 | 2 | 3 |  |  |  |
| вероятность |  |  |  |  |  |  |

Как распределения порождают друг друга:

## Распределение Пуассона и экспоненциальное распределение

Формулы можно получить предельным переходом для геометрического и биномиального распределения.

Увеличиваем число экспериментов в минуту и снижаем вероятность успеха

число экспериментов в минуту

Распределение Пуассона показывает количество успехов за единицу времени

Получим формулу для экспоненциального распределения из общих соображений:

Суммируем прогрессию из таблицы:

## Отрицательное биномиальное распределение и гамма-распределение

Является обобщением геометрического. Непрерывным аналогом является гамма-распределение

– номер испытания, когда накопилось k успехов (или количество неудач)

Гамма-распределение – время до k-го успеха, если мы проводим опытов в минуту,

- среднее число успехов в минуту

Теорема: если и независимы, то

## Дифференциальные формы

Легко интегрировать

Форма некоммутативна:

(так как по свойству выше она равна самой себе со знаком минус).

Упражнение:

Найти:

* Совместную функцию плотности
* f(r)
* f(s)

Члены с произведением одинаковых дифференциалов зануляются

Большими буквами обозначаем случайную величину, маленькими – ее экземпляр.

Пример:

Вероятность попадания в интервал шириной :

– новая функция плотности распределения

## Упражнение на дифференциальные формы

(и независимы друг от друга)

Найти совместную плотность

Второе упражнение:

Переход в полярные координаты

## Бета-распределение

Случайная величина – доля от времени на первые k1 успехов от времени на k1+k2 успехов

Теорема:

Ищем дифференциальную форму:

Следствие:

(сумма k одинаковых распределений)

# Семинар 03.10.17

**Переписано с рукописного конспекта**

Необходимо найти доверительный интервал для

## t-распределение

При условии, что независимы,

Опишем геометрический смысл t-распределения. Пусть

- вектор из

– угол между и осью .

## F-распределение

Величина F имеет F-распределение с степенями свободы, если:

Опишем геометрический смысл. Пусть – вектор в и независимы друг от друга. **Нужно выбрать пространства и , уже не очень понятно, как.**

В линейной регрессии в качестве служит линейная оболочка регрессоров, а в качестве - плоскость с .

Так как независимы друг от друга, то и RSS независима от .

## Проверка линейных гипотез о коэффициентах

Нужно посчитать две регрессии: unrestricted и restricted. Вторую – с учетом ограничений. При справедливости нулевой гипотезы:

Статистика выглядит следующим образом

# Семинар 10.10.2017

# Семинар 17.10.17

Утверждение: если

Оцениваем регрессию

Тогда

Доказательство:

– проекция на линейную оболочку столбцов

– проекция на ортогональное дополнение столбцов

*–* проекция на линейную оболочку единичного вектора

проекция на пересечение ортогонального дополнения линейной оболочки единичного вектора и линейной оболочки столбцов . Размерность пространства -

Ортогональное дополнение линейной оболочки единичного вектора – множество всех векторов, ортогональных этому вектору.

RSS – квадрат длины проекции y на пространство размерности n-k.

(теорема Хершелла – Максвелла)

Прибавление независимого распределения к тому же самому и приводит к бета-распределению

Как преобразовать в :

1. Делить на в числителе и в знаменателе

– результат проецирования вектора с компонентами с распределением у каждого на -мерное пространство.

## Примеры проверки гипотез

Упражнение 1

– цена

– масса

Стандартные ошибки коэффициентов – 13.06 для первого, 14.07 для второго.

1. Построить 95% доверительный интервал для
2. Проверить гипотезу:

Уровень значимости – 10%

1. Построить 95% интервал, который накроет матожидание цены бриллианта для бриллианта, входящего в исходную выборку, и масса которого равна 1 карат
2. Аналогично – для бриллианта с той же массой, не входящего в исходную выборку
3. Построить 95% интервал, который накроет при
4. Построить 95% доверительный интервал для (симметричный по вероятности)

Confidence interval vs predictive interval. Второй – для случайной величины - ?

Решение:

1. Величина имеет (т. к. степеней свободы очень много).

Нормирование на не нужно (ковариационная матрица после обращения дает то же самое)

1. Пусть нулевая гипотеза верна

Большие значения статистики говорят в пользу альтернативы

Нулевая гипотеза не позволит точно вычислить достигаемый уровень значимости, однако сведется к гипотезе в наихудшем случае.

Гипотеза с равенством – простая, с неравенством – сложная.

Уровню значимости 10% соответствует значение статистики 1.28 – гипотеза отвергается.

в.

д.

Интервал:

# Семинар 07.11.17

## Теорема Гаусса-Маркова

Предпосылки:

1. Оцениваем регрессию с помощью МНК
2. (условие обусловленности ковариационной матрицы)

Если они выполнены, то:

1. линейна по
2. Если линейная по и условно несмещенная оценка , то является неотрицательно определенной, в частности, .

– альтернативная оценка

Если дополнительно потребовать нормальность , то любая альтернативная оценка хуже, чем оценка МНК (линейность по уже не требуется).

*Положительно определенная матрица – такая, что все собственные числа положительные (аналогично – любая квадратичная форма от нее положительна)*

*Положительно полуопределенная – все собственные числа неотрицательные (любая квадратичная форма неотрицательная).*

## Доказательство теоремы Гаусса-Маркова

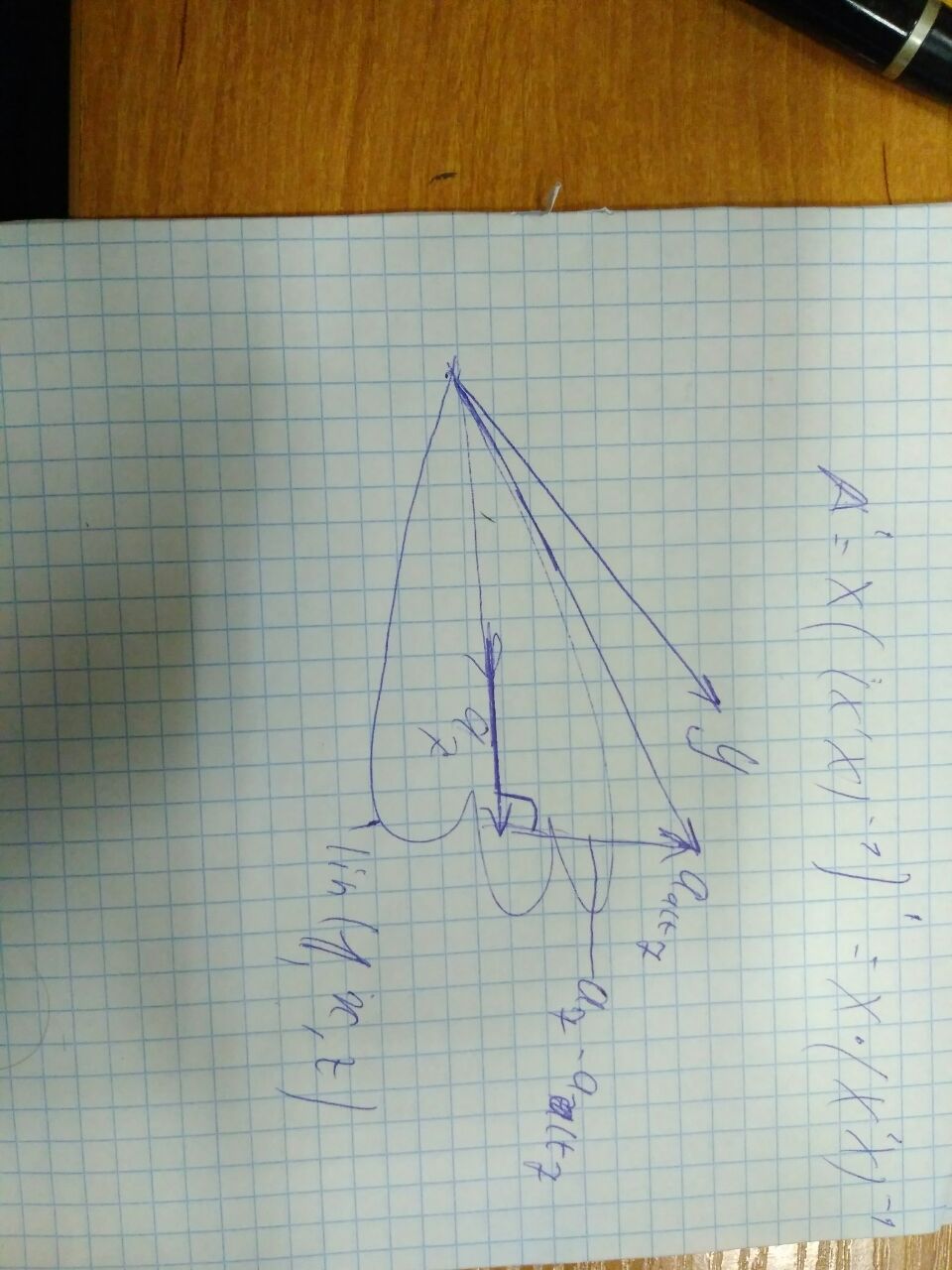
1. Оценка МНК заведомо линейна по (ранее вывели формулу)
2. – было раньше
3. м утверждение про альтернативную оценку

(линейность)

– матрица весов.

– столбцы с весами

– веса в оценке



Требование: . Следовательно:

И в то же время:

Вычтем:

Следовательно: (разница ортогональна любому столбцу )

- гипотенуза, – катет, следовательно, длина больше. Следовательно, и все вариации больше. ЧТД

## Упражнение

*Пусть выборка разделена на две части*

|  |  |  |  |
| --- | --- | --- | --- |
|  | *n* |  |  |
| *Часть А* | *100* | *400* | *200* |
| *Часть В* | *100* | *600* | *400* |

1. *Найти*
2. *Пусть все предпосылки теоремы Гаусса-Маркова выполнены. Является ли оценка:*

*несмещенной и эффективной? Получите оценку лучше, если возможно.*

Проверим несмещенность:

Несмещенность подтверждается.

Проверим эффективность. Оценка МНК:

(суммы разлагаются на суммы по выборкам)

Можно вынести знаменатель и получить линейность по . Для альтернативной оценки:

Для элементов во второй выборке:

Знаменатели у компонентов разные, следовательно, нельзя линейно представить через .

## Хорошие свойства МНК

### Теорема 1. О распределении

Если взять предпосылки теоремы Гаусса-Маркова и

то

### Теорема 2. О состоятельности

Если взять предпосылки теоремы Гаусса-Маркова и (вектор всего того, что относится к -му наблюдению) – независимы и одинаково распределены, то (предел по вероятности: )

Доказательство можно найти в книге Stachursky.

## Гетероскедастичность

**Определение**: гомоскедастичные ошибки – такие, что

*Гетероскедастичные ошибки* – если (разные диагональные члены)

* Где встречается в реальности
  + В реальных задачах почти всегда
* Что будет, если она есть и не учитывается
* Что можно сделать

При случайной выборке

Если ничего не делать, то не выполняются предпосылки теоремы Гаусса-Маркова.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Действие и его свойства** | **Гомоскедастичность** | | | **Гетероскедастичность** | | |
|  | ТГМ | ТГМ + N | ТГМ + незав. и | ТГМ \* | ТГМ\* + N | ТГМ\* + независимы и |
| *Линейность по* | + | + | + | + | + | + |
| *Условная несмещенность* | + | + | + | + | + | + |
| *Состоятельность* | - | - | + | - | - | + |
| *Минимальная дисперсия среди эффективных несмещенных оценок* | + | + | + | - | - | - |
| *Проверка гипотез, построение доверительных интервалов с помощью* | ? |  |  | - | - | - |
| *То же для* | - | - | *+* | - | - | + |

\*без гомоскедастичности

Плюс – свойство выполняется всегда, минус – не всегда

Отсутствие проверки гипотез – самая большая проблема. С эффективностью можно примириться.

*Пример: предпосылки ТГМ выполнены,*

*Оценки совпадут, только если веса при каждом одинаковы*

*Получаем выполнение предпосылок ТГМ. Две формулы для оценок не совпадают, у имеет меньшую дисперсию*

Если известна, то делим каждое наблюдение на . Это взвешенный МНК (WLS, weighted leas squares). Эта оценка линейна по , условно несмещенная и эффективная для всех 6 случаев.

Пояснение – в примере выше

Для этой предпосылки свойства приведены в таблице:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Действие и его свойства | Гомоскедастичность | | | Гетероскедастичность | | |
|  | ТГМ | ТГМ + N | ТГМ + незав. и | ТГМ \* | ТГМ\* + N | ТГМ\* + незав. и |
| Линейность по | + | + | + | + | + | + |
| Условная несмещенность | + | + | + | + | + | + |
| Состоятельность | - | - | + | - | - | + |
| Минимальная дисперсия среди эффективных несмещенных оценок | + | + | + | + | + | + |
| Проверка гипотез, построение доверительных интервалов с помощью | - | *+* | + | - | + | + |

Применяется в реальности редко.

Например, можно считать средний результат ЕГЭ в классе со взвешиванием по количеству учеников в классе

Способ борьбы – оценивать по-другому. Если проверять гипотезы и строить доверительные интервалы с помощью формулы:

– heteroscedascisity consistent

Оценить дисперсий по наблюдениям невозможно. Можно придумать такие оценки:

Есть множество корректировок этой матрицы

# Лекция 14.11.17

Упражнение

Есть разные оценки

1. Уайт:

Разумное требование:

Эксперименты Монте-Карло: генерируем кучу наборов данных и находим наилучшую матрицу. 100-300 наблюдений хватает.

Корректировку стоит использовать всегда, т. к. гетероскедастичность есть почти всегда.

Тесты на гетероскедастичность

1. Простой, но редко используемый:

### Тест Goldfeldt-Quandt.

Нулевая гипотеза:

– монотонно возрастает по переменной . Как правило, переменная – это размер объекта (например, численность персонала).

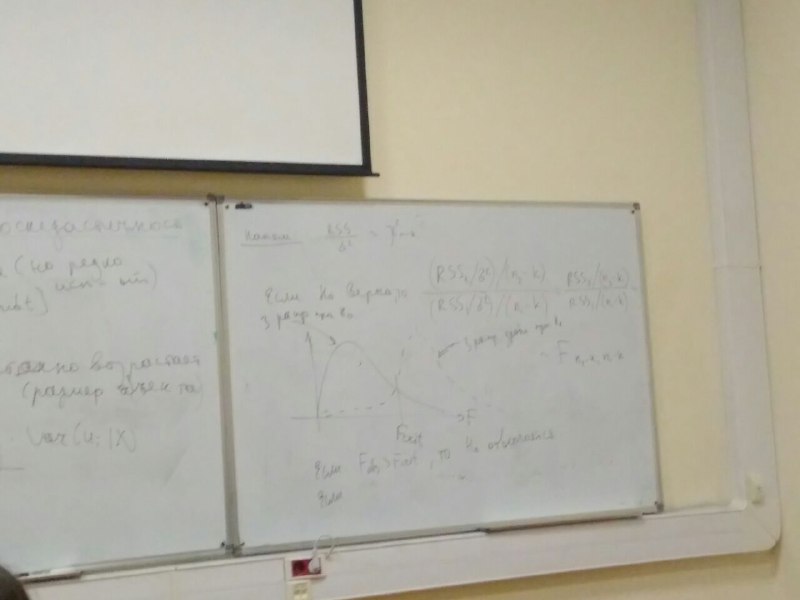
Упорядочим объекты по подозреваемому росту условной дисперсии. Поделим выборку на 3 части:

1. мала
2. 20% выборки в середине выкидываем
3. велика

Считаем сумму квадратов остатков по верхним и по нижним наблюдениям

Поделив, получим по нулевой гипотезе, что сигма сократится и дробь имеет -распределение.

При верна альтернативная гипотеза (делим большее на меньшее)



Распределение при справедливости нулевой гипотезы

Распределение при справедливости альтернативы

### Тест Уайта (Бройша-Лагана)

Два шага:

1. Применяем МНК, получаем остатки
2. Визуальный тест

Построить зависимость . Возрастание означает гетероскедастичность.

Можно проверить и формально: построить вспомогательную регрессию

Для этой регрессии переформулируем гипотезу и альтернативу:

Утверждение: если нулевая гипотеза верна, то

*–* число регрессоров в вспомогательной регрессии (не считая константы)

## Упражнение

Пусть модель:

|  |  |
| --- | --- |
|  |  |
| 1 | 1 |
| 2 | 2 |
| 3 | 2 |

Найти:

Решение

Пояснение к выводу:

.е. – несмещенная оценка

Распределение при условии, что тест Уайта пройден – условное, нужно корректировать формулы. Тест – не сигнал, что дальше нужно считать, что есть гетероскедастичность, нужно основываться на здравом смысле

При гомоскедастичности мы все равно асимптотически сходимся к обычной оценке МНК.

## Метод максимального правдоподобия в примерах

– неизвестный параметр. Нужно получить:

1. – точечная оценка (ML)
2. – интервальная оценка (матрица информации Фишера)
3. низм проверки гипотез
   1. Тест Вальда
   2. Тест Лагранжа
   3. Лемма …

Идея метода максимального правдоподобия (Maximum likelihood, MaxLik): максимизировать вероятность выборки при условии фиксированных параметров

Пример

Score function – производная логарифмированной функции максимального правдоподобия.

ML корректно применять на больших выборках (результаты доказаны только асимптотически)

## Информация Фишера

Теоретическая информация Фишера

Два способа оценить теоретическую информацию:

1. Подставить ML оценку в матожидание
2. Подставить ML оценку в :

Смысл: показывает количество информации о неизвестном параметре , содержащееся в выборке. Позволяет делать только сравнительные оценки (внятной единицы измерения нет).

Используется для оценки дисперсии параметра.

### Упражнение

Представим, что имеют экспоненциальное распределение с

1. двумя способами
2. Двумя способами
3. Доверительный интервал двумя способами

Решение

Точечная оценка

2.

Теоретическая информация:

Вообще здесь могли остаться , но в этом случае нет.

Оценим:

1. Подставляем оценку и честно берем матожидание. Более теоретически корректный способ.
2. Не берем матожидание, сразу подставляем в матрицу Гессе. При наличии результаты были бы другие. Более практичный способ
3. Оцениваем дисперсию

**Условия регулярности**

**При выполнении некоторых условий регулярности на функцию правдоподобия величина**

**(стремится по распределению)**

Границы для 95% интервала (-2; 2)

# Семинар 21.11.17

– исходная функция правдоподобия (до логарифмирования)

– логарифмическая функция правдоподобия.

Функция правдоподобия – случайная величина

Определение: теоретическая информация Фишера

Упражнение: при некоторых условиях регулярности информация Фишера имеет два других представления:

– матрица Гессе от функции



Лемма:

Воспользуемся условиями регулярности. Пусть они позволяют переставить производную и интеграл

Напоминание: ковариационная матрица для вектора

Окончательно получаем:

Распишем для диагональных членов:

Ранее доказано:

Тогда упрощаем первый член:

Проинтегрируем второй член:

Следовательно, подставляем первый член:

Разберемся с недиагональными элементами:

Зануляется аналогично предыдущему.

ЧТД

Утверждение: при выполнении условий регулярности функции правдоподобия:

1. (сходится по вероятности). – состоятельные оценки
2. – асимптотически нормальные. Т. Е. (сходится по распределению)
3. – асимптотически наиболее эффективные среди несмещенных. Т. е. (уже обычная сходимость). Домножение на n делает сходимость нетривиальной (сходятся к нулю с одним темпом)

Неравенство Рао-Крамера: если есть

Для матриц неравенство означает, что матрица разниц неотрицательно определена.

## Тесты

1. Тест отношения правдоподобия
2. Тест Вальда (Wald test)
3. Тест множителей Лагранжа (LM, Lagrange Multipliers)

Тест Вальда: далеко ли от нуля

LR: велика ли разница

LM: велика ли разница

Все тесты асимптотически сходятся к хи-квадрат

– максимум l без ограничений

– максимум с учетом ограничений (restricted)

– вектор параметров, на которые накладываются ограничения.

*Немного напоминаний:*

*Оценка теоретической информации Фишера:*

1. *(без взятия матожидания)*

Пусть – неслучайны

Собственно задание:

1. LR - ?
2. W для первой оценки информации
3. Для второй оценки информации
4. LM – два аналогичных случая

### Решение

Без ограничений:

С ограничениями:

**ДЗ: посчитать оба вариант LM, оба варианта W, обобщить результаты на случай векторной (гипотеза равенства всех коэффициентов нулю).**

# 28.11.2017

Все статистические тесты для правдоподобия являются модификациями LM, LR, W (кроме экзотики)

## Пример

За прошлый месяц поймано:

|  |  |  |
| --- | --- | --- |
| Кого | Сколько |  |
| Бульбазавр | 10 | Y1 |
| Ивизавр | 20 | Y2 |
| Чармандер | 15 | Y3 |
| Сквиртл | 20 | Y4 |

LM, LR, W

Построить 95% доверительный интервал для

Считаем покемонов независимыми.

Вероятностная модель: каждый последующий независимо от прошлых принимает один из четырех типов с заданными вероятностями.

А) ограниченный и неограниченный случай

Б)

Функция Лагранжа:

Ищем информацию

Ищем наблюдаемую информацию Фишера (подставляем либо в матрицу Гессе до подсчета матожидания либо подставить сразу в подсчитанное матожидание).

Оценка асимптотически сходится к

Получим статистику Вальда:

*–* из UR модели

Следовательно:

Получим статистику LM:

Переобозначим:

Если подставлять в матожидания:

Если подставлять в гессиан сразу: