Эконометрика-1

Филатов Александр Юрьевич

(Главный научный сотрудник, доцент ШЭМ ДВФУ)

alexander.filatov@gmail.com

http://vk.com/alexander.filatov, http://vk.com/baikalreadings

Лекции 1.1-1.2

Введение.
Корреляционный анализ

Немного о себе

Филатов Александр Юрьевич
Главный научный сотрудник, доцент ШЭМ ДВФУ.

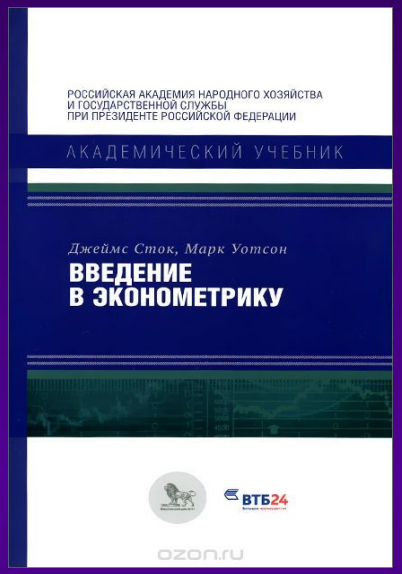
Образование:
ИГУ «Математические методы в экономике» (1998).
Кандидат физико-математических наук (2001), доцент (2005).

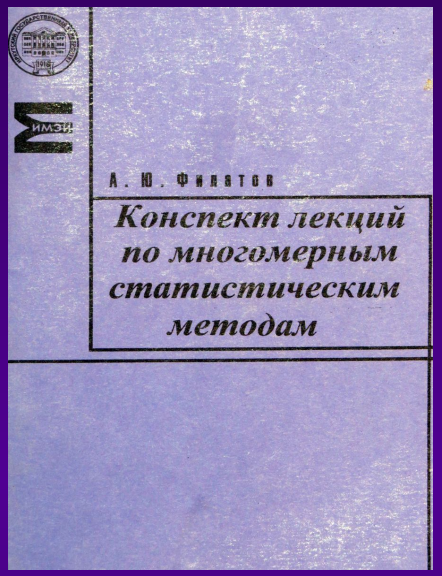
Программы повышения квалификации:
РЭШ, НИУ ВШЭ, МГУ, Европейский университет СПб,
CERGE-EI, IOS, Indiana University.

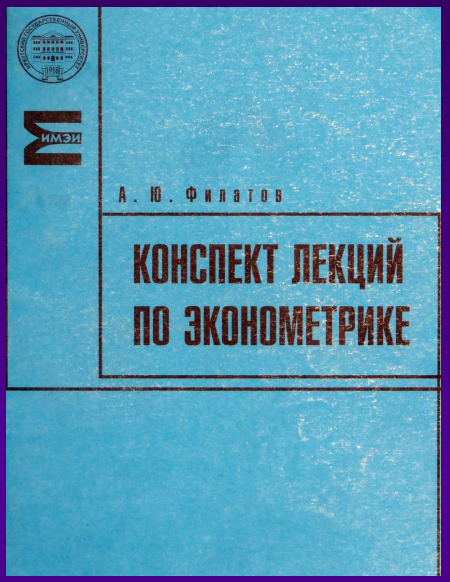
Научные интересы:
Экономика отраслевых рынков, пространственная экономика, олигополия, монополия и монополистическая конкуренция, экономика энергетики, экономика неоднородности, теория игр, прикладная эконометрика

Связь:
alexander.filatov@gmail.com.
http://vk.com/alexander.filatov, http://vk.com/baikalreadings.

Литература







Дополнительные материалы:

1. Материалы курса в системе BlackBoard.
2. Презентации, книги, видеолекции в группе http://vk.com/baikalreadings, в т.ч. курс эконометрики Дмитрия Вихрова (CERGE-EI).
3. «РЭШ. Экономика: просто о сложном»: https://www.nes.ru/ru/events/nes-public-lectures/lectures-in-politech/past.
4. Coursera: курс эконометрики Бориса Демешева (с 27 апреля) https://www.coursera.org/learn/ekonometrika.

Экзамен

1. Посещение и краткие еженедельные тесты = 9*2 = **18**.
2. Домашние контрольные работы = 7*6 = **42** (выполняются на индивидуальных данных).
3. Активность на занятии (ответы на вопросы, дополнительные задания и т.д.) = **10** – «долларовая система».
4. Коллоквиум (2 теоретических вопроса + практическое задание) = **30**.

Ориентировочная шкала оценок:

≥ 50 баллов – удовлетворительно;

≥ 65 баллов – хорошо;

≥ 80 баллов – отлично.

Содержание курса

1. Введение в эконометрику. Данные и их предварительная обработка.
2. Корреляционный анализ количественных переменных. Коэффициент детерминации. Коэффициент корреляции. Корреляционное отношение.
3. Корреляционный анализ количественных переменных. Частные и множественный коэффициенты корреляции.
4. Корреляционный анализ порядковых и категоризованных переменных.
5. Регрессионный анализ. Метод наименьших квадратов. Значимость регрессоров и модели.
6. Проблема мультиколлинеарности. Методы устранения. Метод главных компонент.
7. Гетероскедастичность и автокорреляция остатков. Взвешенный и обобщенный МНК.
8. Модели с переменной структурой. Использование дамми-переменных. Неоднородность данных.
9. Нелинейные модели, поддающиеся непосредственной линеаризации. Процедура Бокса-Кокса.

Содержание курса

10. Бинарные результирующие показатели. Логит- и пробит-модели.
11. Анализ временных рядов. Тренд. Сезонность. Цикл.
12. Аналитические тренды. Скользящее среднее. Экспоненциально взвешенное скользящее среднее.
13. Сезонность и ее устранение.
14. Модели обработки остатков. ARMA-модели и их идентификация.
15. Учет временных лагов. Модели с распределенными лагами. Модель Койка.
16. Панельные данные. Модель с фиксированными эффектами.
17. Системы одновременных уравнений. Проблема эндогенности. Инструментальные переменные.
18. Введение в оценивание с использованием статистического пакета «Stata».

Введение в эконометрику

Эконометрика – «измерения в экономике» (Рагнар Фриш, 1926);
– придает количественное выражение качественным закономерностям, вводимым экономической теорией.

Микро-, макроэкономика → (коэффициенты в моделях заменяются оценками, полученными на основе экономических данных) → **Эконометрика**

Основания эконометрики:

1. Экономические законы (микроэкономика, макроэкономика).
2. Информационное обеспечение (экономическая статистика).
3. Методы (математико-статистический инструментарий).

Уровни иерархии:

1. Макроуровень (страны, мир).
2. Мезоуровень (регионы, отрасли).
3. Микроуровень (домашние хозяйства, фирмы).

Введение в эконометрику

Конечные прикладные цели:

Мониторинг, прогнозирование, управление, устойчивое развитие.

Принципиальная идея – наличие взаимосвязей между переменными.

спрос ← цена, доход, реклама, цены на другие товары;
издержки ← объем производства, цены на факторы производства;
потребление ← доход, активы, предельная норма потребления.

Используемые методы:

1. Корреляционный анализ.
2. Регрессионный анализ.
3. Анализ временных рядов.
4. Системы одновременных уравнений.

} Статистическое исследование структуры и характера взаимосвязи между переменными

5. Методы классификации (всю популяцию из n объектов разбить на не-большое число однородных подгрупп).
6. Методы снижения размерности признакового пространства (перейти от исходных p переменных к меньшему их числу).

Введение в эконометрику

Выборка (реально наблюдаемая) эмпирические св-ва и характ-ки	→	**Популяция** (теор.домысливаемая) теоретические св-ва и характ-ки

Цель: найти как можно точнее

Важны:

1. Правильный выбор модели (линейная / нелинейная; аддитивная / муль-типликативная; учет лагов,…)
2. Правильный выбор статистической обработки данных.

Даже при фиксации объясняющих переменных на едином уровне есть варьирование отклика – имеется случайная составляющая!

При этом в эконометрическом анализе часто нет никаких сведений о вероятностной природе анализируемых данных, есть только соображения конкретно-содержательного плана.

Исходные данные (что на входе)

1. Матрица «объект-свойство»:

$$\begin{pmatrix} x_1^{(1)}(t) & \dots & x_1^{(p)}(t) \\ x_2^{(1)}(t) & \dots & x_2^{(p)}(t) \\ \dots & \dots & \dots \\ x_n^{(1)}(t) & \dots & x_n^{(p)}(t) \end{pmatrix}$$

$i = 1,\dots, n$ – объекты, наблюдения

$j = 1,\dots, p$ – свойства, переменные

$t = \{t_1,\dots, t_T\}$ – моменты времени

Часто равноотстоящие моменты времени: $t_2 - t_1 = \dots = t_T - t_{T-1} = \Delta t$

Частные случаи:

1. $n > 1,\ p > 1,\ T = 1$ – пространственная выборка (cross-section)
   ## Зависимость объемов продаж от цен и рекламных бюджетов.
2. $n = 1,\ p = 1,\ T > 1$ – одномерный временной ряд (time series).
   ## динамика курса доллара.
3. $n = 1,\ p > 1,\ T > 1$ – многомерный временной ряд (time series).
   ## динамика валютных курсов.
4. $n > 1,\ p > 1,\ T > 1$ – панельные данные (panel data).
   ## динамика макроэкономических показателей стран мира.

Исходные данные (что на входе)

2. Матрица парных сравнений:

$$\begin{pmatrix} \gamma_{11}(t) & \dots & \gamma_{1n}(t) \\ \dots & \dots & \dots \\ \gamma_{n1}(t) & \dots & \gamma_{nn}(t) \end{pmatrix} \quad \text{или} \quad \begin{pmatrix} \gamma_{11}(t) & \dots & \gamma_{1p}(t) \\ \dots & \dots & \dots \\ \gamma_{p1}(t) & \dots & \gamma_{pp}(t) \end{pmatrix}$$

Часто, но не всегда симметричная

$\gamma_{ij}(t)$ – попарное сравнение объектов или признаков в момент времени t.
Часто, но не всегда, $\gamma_{ij}(t) = \gamma_{ji}(t)$ – симметричная матрица.

Расстояние, поток продукции (экспорт, импорт, торговый оборот), коэффициенты корреляции, отношения предпочтения,…

$$\hat{R} = \begin{pmatrix} 1 & -0{,}389 & 0{,}613 & 0{,}088 \\ -0{,}389 & 1 & 0{,}152 & -0{,}269 \\ 0{,}613 & 0{,}152 & 1 & -0{,}391 \\ 0{,}088 & -0{,}269 & -0{,}391 & 1 \end{pmatrix} \begin{matrix} y \\ x^{(1)} \\ x^{(2)} \\ x^{(3)} \end{matrix}$$

$$\begin{matrix} y & x^{(1)} & x^{(2)} & x^{(3)} \end{matrix}$$

Объем продаж
Цена
Рекламный бюджет
Число праздников

Типовые задачи эконометрики

1. Нормирование:

$\hat{y} = f(X;\Theta)$. Задача найти вектор параметров Θ.

Оценивание величины постоянных и предельных издержек.

Решение задач массового обслуживания (супермаркет, такси).

2. Прогнозирование:

$y_i,\ x_i^{(1)},...,x_i^{(p)},\ i=1,...,n$ - значения в прошлом или на аналог. объектах

Нужно оценить y_{n+1} по известным $x_{n+1}^{(1)},...,x_{n+1}^{(p)}$.

Прогнозирование спроса.

Диагностика эффективности рекламы.

Прогнозирование динамики валютного курса и курса акций.

3. Оценка труднодоступных для наблюдения показателей:

Выявление предпочтений потребителей и их реакции на стимулы.

Оценка денежных сбережений по доходу.

Типовые задачи эконометрики

4. Оценка не подлежащих измерению показателей:

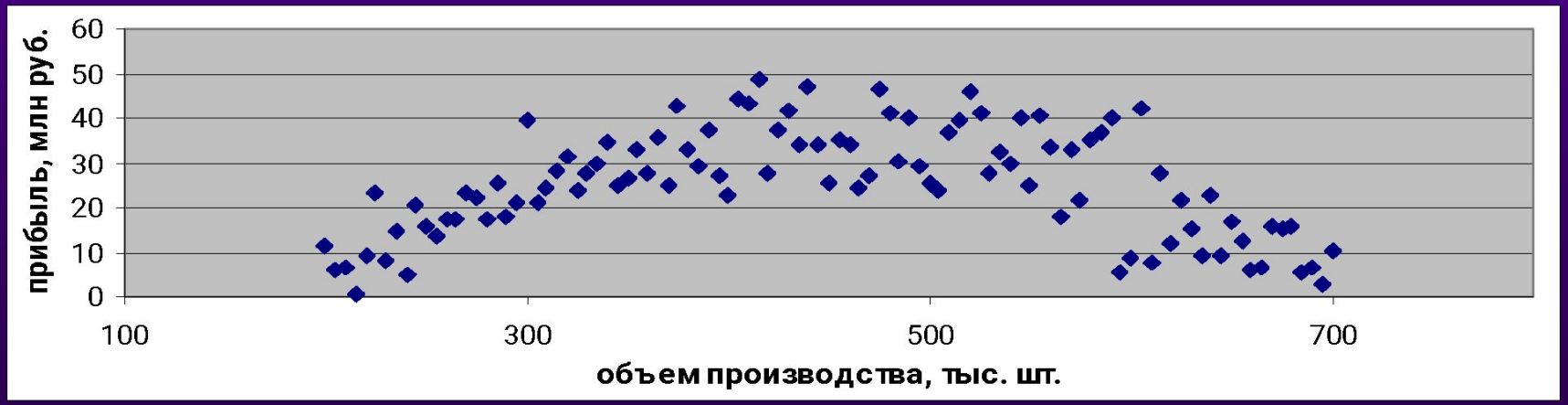
Ранжирование стран по качеству жизни (материальный достаток, экологическая ситуация, безопасность, уровень образования и медицины, качество институтов,… → совокупный индикатор).

Оценка эффективности менеджмента.

5. Оптимальное управление:

Оптимальная фискальная и монетарная политика (налоговая ставка, ключевая ставка ЦБ, объем интервенций на валютном рынке,…).

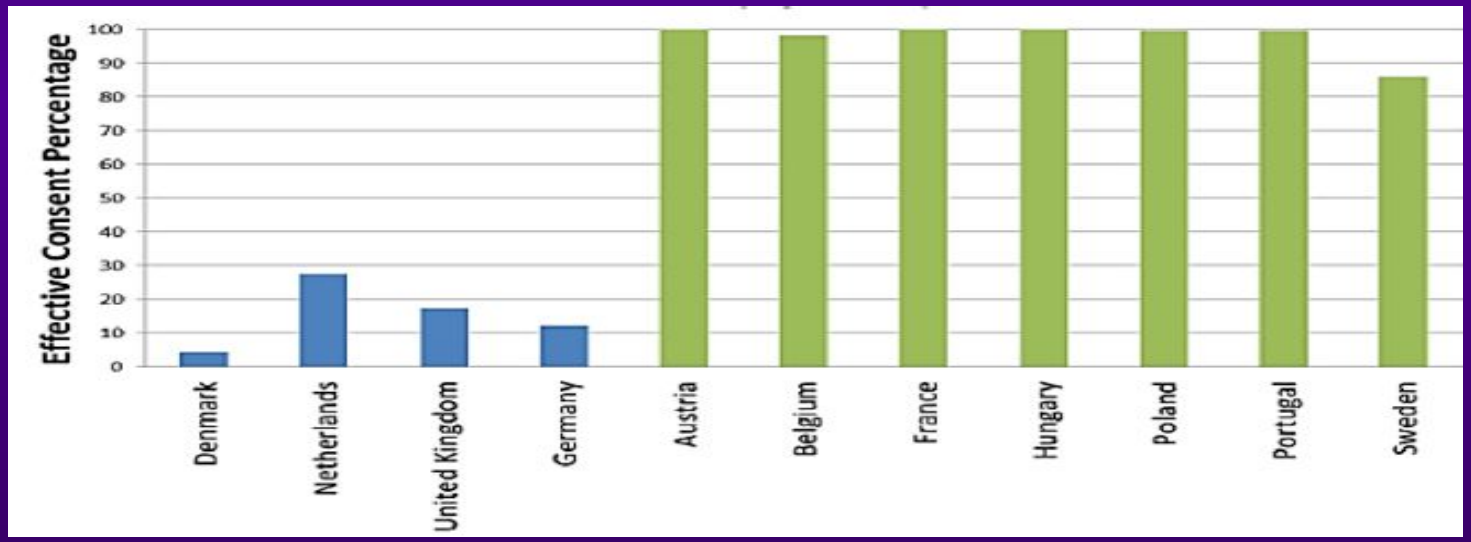
Поиск оптимального объема производства и ценовой политики, которые максимизируют прибыль.



Этапы эконометрического анализа

1. Предварительный анализ исследуемой экономической системы

1) Определение основных целей исследования.
2) Отбор переменных $x^{(1)},\ldots x^{(p)}$.
3) Формализация переменных, единицы измерения.
4) Определения форм, используемых для сбора информации.



2. Составление плана сбора информации, определение баз данных, формирование репрезентативной выборки, сбор данных и ввод в компьютер.

Этапы эконометрического анализа

3. Первичная обработка данных

1) Отображение переменных, описанных текстом (количественная шкала; шкала с n градациями; категории).
2) Унификация типов переменных (количественные, порядковые, категоризованные).
3) Статистическое описание популяций с указанием пределов варьирования переменных.
4) Обработка аутлаеров – резко выделяющихся наблюдений: (исключение; меньший вес; преобразование данных:

$$x \in [1; +\infty) \rightarrow \tilde{x} \in [1; 2): \qquad \tilde{x} = \frac{x-1}{x} + 1$$

5) Восстановление пропущенных данных.
6) Проверка однородности порций данных $(n_1 \times p) + \ldots + (n_k \times p) = (n \times p)$
7) Проверка статистической независимости переменных.

Этапы эконометрического анализа

4. Предварительный экспериментальный анализ

1) Выборочное среднее, дисперсия, асимметрия, эксцесс.
2) Выборочная корреляционная матрица.
3) Учет априорной информации об экономической сущности связи: монотонная или имеет экстремум, стремление к асимптотам, аддитивное или мультипликативное воздействие, прохождение графика через определенные точки пространства.
4) Построение корреляционных полей – парных зависимостей $x^{(k)}(x^{(j)})$ в количестве $(p+1)\cdot p/2$.
5) Визуальное прослеживание каждого поля: линейное / нелинейное монотонное / с одним или несколькими экстремумами.
6) Изучение условных средних (диапазон переменной по оси абсцисс разбивается на интервалы группировки).

5. Составление детального плана анализа с определением методов и критерия качества, вычислительная реализация.

6. Интерпретация результатов и подведение итогов.

Корреляционный анализ количественных переменных

1. **Выбрать подходящий измеритель** статистической связи (коэффициент корреляции, корреляционное отношение и т. д.).
2. **Оценить** (с помощью точечной и интервальной оценок) его числовое значение по выборочным данным.
3. **Проверить гипотезу** о том, что полученное числовое значение действительно свидетельствует о наличии статистической связи (корреляционная характеристика значимо отлична от нуля).

Рассматриваемая зависимость: $y(X) = f(X) + \varepsilon(X)$

$X = \left(x^{(1)}, \ldots, x^{(p)}\right)$ – объясняющие переменные, y – результирующая.

$Dy = Df + D\varepsilon$ – связь безусловных характеристик.

Теснота связи максимальна, если по заданному X можно восстановить y без всякой ошибки: $\varepsilon(X) \equiv 0$, $D\varepsilon = 0$, $Dy = Df$.

Теснота связи минимальна, если значения X не несут никакой информации об y: $f(X) \equiv \text{const}$, $Df = 0$, $Dy = D\varepsilon$.

Коэффициент детерминации – наиболее общий показатель связи

Коэффициент детерминации отражает долю общей вариации y, объясненную функцией регрессии $f(X)$:

$$K_d(y,X)=\frac{Df}{Dy}=1-\frac{D\varepsilon}{Dy}\in[0;1].$$

$K_d(y, X) = 1$, теснота связи максимальна, если $Df = Dy$, $D\varepsilon = 0$, $\varepsilon(X) \equiv 0$, $y = f(X)$ – функциональная зависимость.

$K_d(y, X) = 0$, теснота связи минимальна, если $Df = 0$, $D\varepsilon = Dy$, $f(X) \equiv \text{const}$ – полное отсутствие связи.

Выборочное значение коэффициента детерминации:

$$s_y^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2,\quad \bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$$

$$\hat{K}_d(y,X)=1-\frac{s_\varepsilon^2}{s_y^2},\qquad s_\varepsilon^2=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{f}(X_i)\right)^2,$$

если есть оцененное в точке значение функции регрессии

$$s_\varepsilon^2=\frac{1}{n}\sum_{j=1}^{s}\sum_{i=1}^{n_j}(y_{ji}-\bar{y}_j)^2,$$

если есть группировка

Основные показатели тесноты статистической связи

Наиболее общий показатель тесноты связи – коэффициент детерминации $K_d(X)$

- Показатели парной связи
  - Линейная связь
    - Парный коэффициент корреляции r_{xy}
  - Произвольная связь
    - Корреляционное отношение ρ_{yx}
- Показатели множественной связи
  - Частные коэффициенты корреляции $r_{ij(-ij)}$
  - Множественный коэффициент корреляции $R_{y.X}$

Парный коэффициент корреляции

Парные корреляционные характеристики измеряют тесноту связи без учета опосредованного или совместного влияния других показателей, только на основе наблюдения значений двух переменных.

Коэффициент корреляции измеряет тесноту парной линейной связи:

$$r_{xy} = \frac{E\big((x - Ex)(y - Ey)\big)}{\sigma_x \sigma_y}$$

Свойства парного коэффициента корреляции:

1. $r_{xy} \in [-1; 1]$.
   Если $r_{xy} > 0$, то монотонно возрастающая парная линейная связь.
   Если $r_{xy} < 0$, то монотонно убывающая парная линейная связь.
2. Если x и y статистически независимы, то $r_{xy} = 0$.
3. $|r_{xy}| = 1$ тогда и только тогда, когда имеется функциональная связь.
4. Коэффициент корреляции – симметричная характеристика: $r_{xy} = r_{yx}$.

Если x и y распределены нормально или связаны только линейно:

5. Если $r_{xy} = 0$, то x и y статистически независимы.
6. $K_d(y, x) = r_{xy}^2$.

Проверка гипотезы о наличии парной линейной связи

Выборочный коэффициент корреляции:

$$\hat{r}_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \text{КОРРЕЛ}(x_1 : x_n; y_1 : y_n)$$

Вопрос: какую величину выборочного коэффициента корреляции можно считать достаточной для статистически обоснованного вывода о наличии связи между исследуемыми переменными?

Величина зависит от размерности, поскольку с уменьшением объема выборки ослабевает надежность статистических характеристик, и **уровня значимости** – вероятности ошибки первого рода (отвергнуть истинную гипотезу.

Статистика: $\hat{t} = \dfrac{\hat{r}_{xy}\sqrt{n-2}}{\sqrt{1-\hat{r}_{xy}^2}} \sim t(n-2)$ – закон распределения Стьюдента.

Проверка гипотезы о наличии парной линейной связи

Гипотеза о статистической независимости *x* и *y*: H_0: $r_{xy} = 0$.

1. Выбираем уровень значимости α.

Типичные значения $\alpha = 0{,}05$; 0,1; 0,01, 0,001.

2. Вычисляем эмпирическое значение критерия:

$$t_{\text{эмп}} = \frac{\left|\hat{r}_{xy}\right|\sqrt{n-2}}{\sqrt{1-\hat{r}_{xy}^2}}$$

3. Вычисляем критическую точку:

$t_{\text{крит}} = t(\alpha; n-2) =$ СТЬЮДРАСПОБР $(\alpha; n-2)$.

4. Сравниваем эмпирическое и критическое значение и делаем вывод:

Если $t_{\text{эмп}} > t_{\text{крит}}$, то гипотеза H_0 отвергается при уровне значимости α, между переменными наблюдается связь, близкая к линейной.

Возможно решение обратной задачи – найти такое значение α, при котором эмпирическое и критическое значение совпадают. Это граничное значение уровня значимости называется ***p*-value**.

Доверительный интервал для истинного значения коэффициента корреляции

Доверительный интервал для истинного значения коэффициента корреляции асимметричен и смещен относительно оценки $\hat{r}_{xy}$.

1. Выбираем доверительную вероятность γ.

Типичные значения **$\gamma = 0{,}95$**; 0,9; 0,99, 0,999.

2. Убираем асимметричность преобразованием Фишера:

$$z = \frac{1}{2}\ln\frac{1+r}{1-r} = \text{ФИШЕР}(r)$$

3. Убираем смещение: $\tilde{z} = \hat{z} - \dfrac{\hat{r}}{2(n-1)}$.

4. Находим доверительный интервал для переменной z:

$$z \in [z_1; z_2] = \left[\tilde{z} + \frac{u_{(1-\gamma)/2}}{\sqrt{n-3}};\ \tilde{z} + \frac{u_{(1+\gamma)/2}}{\sqrt{n-3}}\right],$$

$U_\alpha = \text{НОРМСТОБР}(\alpha)$ – квантили норм. станд. распределения.

5. Возвращаемся в исходные координаты:

$$r = \frac{e^z - e^{-z}}{e^z + e^{-z}} = \text{ФИШЕРОБР(z)}, \quad r \in \left[\text{ФИШЕРОБР}(z_1);\ \text{ФИШЕРОБР}(z_2)\right]$$

Влияние ошибок измерения анализируемых переменных на величину коэффициента корреляции

При уменьшении доверительной вероятности или увеличении объема выборки **интервал сужается**, а при увеличении доверительной вероятности или сокращении объема выборки – **расширяется**!

Если переменные x и y измерены с ошибками ε_x и ε_y, эти ошибки независимы между собой, не зависят от x и y, распределены по нормальному закону с нулевыми математическими ожиданиями и стандартными ошибками σ_1 и σ_2, коэффициент корреляции корректируется по формуле

$$r'_{xy} = \frac{r_{xy}}{\sqrt{\left(1+\frac{\sigma_1^2}{\sigma_x^2}\right)\left(1+\frac{\sigma_2^2}{\sigma_y^2}\right)}}$$

Ошибки измерения ослабляют исследуемую корреляционную связь между переменными. Это искажение тем меньше, чем меньше отношение дисперсий ошибок к дисперсиям самих исходных переменных.

Парные нелинейные связи: корреляционное отношение

Если исследуемая зависимость отклоняется от линейного вида, то парный коэффициент корреляции r теряет смысл как характеристика степени тесноты связи.

Двумерные выборочные данные $(x_1; y_1),\ldots,(x_n; y_n)$.

По переменной x производится разбиение на s интервалов группировки.

Корреляционное отношение y по x:

$$\hat{\rho}^2_{yx} = \frac{s^2_{\bar{y}(x)}}{s^2_y} = \frac{\frac{1}{n}\sum_{j=1}^{s} n_j \left(\bar{y}_j - \bar{y}\right)^2}{\frac{1}{n}\sum_{j=1}^{s}\sum_{i=1}^{n_j} \left(y_{ji} - \bar{y}\right)^2}$$ – оценка коэффициента детерминации.

n – общий объем выборки; s – число интервалов группировки;

n_j – число точек, попавших в j-интервал;

$\bar{y}_j$ – условное среднее из ординат j-интервала; $\bar{y}$ – общее среднее;

y_{ji} – ордината i-точки из j-интервала.

Свойства корреляционного отношения

Свойства корреляционного отношения:

1. $\rho_{yx} \in [0; 1]$.
2. $\rho_{yx} = 1$ тогда и только тогда, когда имеется функциональная связь.
3. $\rho_{yx} = 0$ тогда и только тогда, когда наблюдается полное отсутствие связи, то есть $\bar{y} = \bar{y}_j = \text{const}$.
4. Корреляционное отношение – асимметричная характеристика: $\rho_{yx} \neq \rho_{xy}$.

   ## $y = x^2$.

x	-1	0	1
y	1	0	1

   $\rho_{yx} = 1, \quad \rho_{xy} = 0.$
5. $\rho_{yx} \geq | r_{xy} |$. Если наблюдается линейная зависимость, значения близки.

Из свойства 5 следует, что величину $\hat{\rho}^2_{yx} - \hat{r}^2_{xy}$ можно рассматривать как меру отклонения регрессионной зависимости от линейного вида.

Проверка гипотезы о наличии связи произвольного вида

Гипотеза о статистической независимости *x* и *y*: H_0: $\rho_{xy} = 0$.

1. Выбираем уровень значимости α.

Типичные значения $\alpha = 0{,}05$; 0,1; 0,01, 0,001.

2. Вычисляем эмпирическое значение критерия:

$$F_{\text{эмп}} = \frac{\hat{\rho}_{yx}^2}{1-\hat{\rho}_{yx}^2}\frac{n-s}{s-1}$$

3. Вычисляем критическую точку:

$F_{\text{крит}} = F(\alpha; s-1; n-s) =$ FРАСПОБР (α; s – 1; n – s).

4. Сравниваем эмпирическое и критическое значение и делаем вывод:

Если $F_{\text{эмп}} > F_{\text{крит}}$, то гипотеза H_0 отвергается при уровне значимости α, между переменными наблюдается некоторая связь произвольного вида.

Поскольку при вычислении корреляционного отношения используется эмпирическая функция регрессии, построенная по условным средним, никакого конкретного вида зависимости не предполагается.

Доверительный интервал для истинного значения корреляционного отношения

1. Выбираем доверительную вероятность γ.
2. Вычисляем вспомогательное число степеней свободы ν^*:

$$\nu^* = \frac{\left(s-1+n\hat{\rho}_{yx}^2\right)^2}{s-1+2n\hat{\rho}_{yx}^2}.$$

3. Вычисляем критические точки распределения Фишера:

$$F_{\frac{1-\gamma}{2}} = F\left(\frac{1-\gamma}{2}; \nu^*; n-s\right), \quad F_{\frac{1+\gamma}{2}} = F\left(\frac{1+\gamma}{2}; \nu^*; n-s\right).$$

4. Вычисляем доверительный интервал для истинного значения ρ_{yx}:

$$\rho_{yx}^2 \in \left[\frac{(n-s)\hat{\rho}_{yx}^2}{n\left(1-\hat{\rho}_{yx}^2\right)F_{\frac{1-\gamma}{2}}} - \frac{s-1}{n}; \frac{(n-s)\hat{\rho}_{yx}^2}{n\left(1-\hat{\rho}_{yx}^2\right)F_{\frac{1+\gamma}{2}}} - \frac{s-1}{n}\right]$$

Не следует использовать при малом объеме выборки. Значения левого и правого концов могут выходить за пределы [0;1]. Нужна корректировка!

Спасибо за внимание!

alexander.filatov@gmail.com
http://vk.com/alexander.filatov, http://vk.com/baikalreadings