Парная регрессия и корреляция

Специфика экономических измерений состоит в наличии большого числа разнородных данных – ресурсов и результатов (например, товаров и услуг). Отсюда большое значение имеют стоимостные метрики, далеко не всегда отвечающие поставленным задачам. Это не исключает потребность в натуральных метриках.

Количественная определенность функционирования экономики имеет объемные и структурные характеристики.

Объемные характеристики определяют масштаб явления, тогда как структурны – его разнообразие, организацию и соподчиненность. Количественные и структурные меры дополняют друг друга.

Главное, что определяет специфику точности экономических измерений, - это неконтролируемость погрешности наблюдений.

Точность измерения – это его адекватность.

По объективным причинам для социально-экономических измерений характерна низкая контролируемость их точности.

- В области экономических измерений проблема точности связана со следующими показателями:
- Определением понятия «экономическая величина»;
- Формированием системы принципов, постулатов и других теоретических положений, формирующих базис точности экономических измерений;
- Определением экономических показателей;
- Разработкой принципов конструирования измерителей и измерений;
- Основанием выбора типа шкал при конструировании измерителя;
- Разработкой правил формирования систем показателей;
- Выявлением типов и определением методов устранения ошибок экономического измерения;
- Разработкой правил агрегирования и сверки экономических показателей;
- Выявлением условий сравнимости экономических величин (показателей);
- Разработкой правил и методов измерений.

Спецификация моделей

Номер наблюдения	Доход Долл. DPI	Потреб Долл CONS	Номер наблюдения	Доход Долл.	Потреб долл
1	2508	2406	11	2432	2311
2	2572	2564	12	2354	2278
3	2408	2336	13	2404	2240
4	2522	2281	14	2381	2183
5	2700	2641	15	2581	2408
6	2531	2385	16	2529	2379
7	2390	2297	17	2562	2378
8	2595	2416	18	2624	2554
9	2524	2460	19	2407	2232
10	2685	2549	20	2448	2356

Результаты наблюдений за расходами

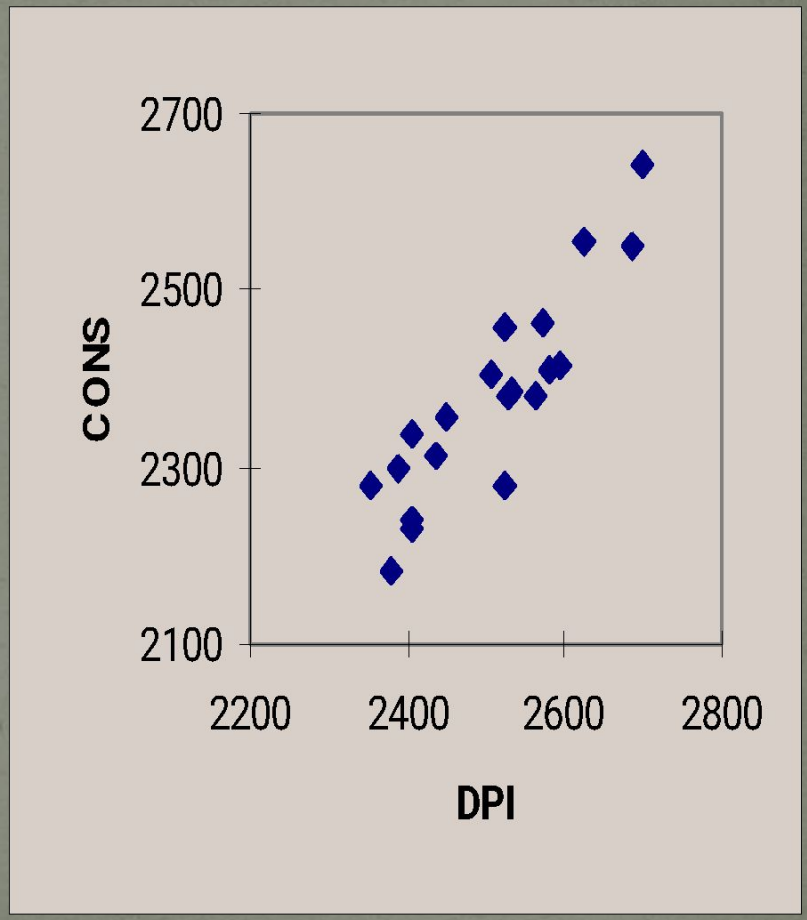

Диаграмма рассеяния.

Спецификация моделей

- Причина неоднозначной связи между располагаемым доходом и расходами:
  - Индивидуальные особенности домашних хозяйств
  - Влияние неучтенных факторов.
- **Выводы:**
  - Невозможно построить модель вида Y=f(x), с помощью которой возможно однозначно определить связь между расходами и доходами.
  - Зависимость между доходами и расходами домашних хозяйств имеет элемент случайности.

- Для учета случайного характера экономических процессов, модель записывают в виде:

  $Y = f(X) + \varepsilon$

  где: Y – эндогенная переменная;

  X – вектор предопределенных переменных;

  f(X) – детерминированная математическая функция, определяющая закономерность между эндогенной и предопределенными переменными;

  ε – случайная величина, учитывающая влияние неучтенных факторов и индивидуальные особенности конкретного объекта.

Парная регрессия

- **Парная регрессия** – уравнение связи двух переменнных

$$y = \hat{f}(x)$$

- *y* – зависимая переменная (результативный признак);
- *x* – независимая, объясняющая переменная (признак-фактор)

Различают линейные и нелинейные регрессии

- Линейная регрессия:

$$y = a + b \cdot x + \varepsilon$$

- Нелинейные регрессии делятся на **два класса**:
- Регрессии, нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам;
- Регрессии, нелинейные по оцениваемым параметрам

Регрессии, нелинейные по объясняющим переменным:

- Полиномы разных степеней

$$y = a + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3 + \varepsilon$$

- Равносторонняя гипербола

$$y = a + \frac{b}{x} + \varepsilon$$

Регрессии, нелинейные по оцениваемым параметрам:

- Степенная

$$y = a \cdot x^{b} \cdot \varepsilon$$

- Показательная

$$y = a \cdot b^{x} \cdot \varepsilon$$

- Экспоненциальная

$$y = e^{a+b \cdot x} \cdot \varepsilon$$

Линеаризация нелинейных по оцениваемым параметрам уравнений парной регрессии

1. степенная функция

$$y = a \cdot x^{b} \cdot \varepsilon$$

$$\lg y = \lg a + b \cdot \lg x$$

$$Y = C + bX$$

$$Y = \lg y$$

$$; \quad X = \lg x \quad ;$$

$$c = \lg a$$

2. показательная функция

$$y = a \cdot b^{x} \cdot \varepsilon$$

$$\lg y = \lg a + x \lg b$$

$$Y = C + Bx$$

$$Y = \lg y \;;$$

$$C = \lg a \;;$$

$$B = \lg b$$

3. экспоненциальная функция

$$y = e^{a+bx} \cdot \varepsilon$$

$$\ln y = a + bx$$

4. равносторонняя гипербола

$$y = a + b \cdot \frac{1}{x}$$

замена $z = \frac{1}{x}$

$$y = a + bz$$

- **РЕГРЕССИЯ ЛИНЕЙНАЯ ПАРНАЯ** - причинная модель статистической связи линейной между двумя количественными переменными x и y, представленная уравнением $y = a + bx$

- Существуют два подхода к интерпретации коэффициента регрессии b.
- Согласно первому из них, b представляет собой величину, на которую изменяется предсказанное по модели значение $\hat{y}_i = a + bx_i$ при увеличении значения независимой переменной x на одну единицу измерения, согласно второй - величину, на которую в среднем изменяется значение переменной y_i при увеличении независимой переменной x на единицу.
- На диаграмме рассеяния коэффициент b представляет тангенс угла наклона линии регрессии $y = a + bx$ к оси абсцисс. Знак коэффициента регрессии совпадает со знаком коэффициента линейной корреляции: значение $b > 0$ свидетельствует о прямой линейной связи, значение $b < 0$ - об обратной. Если $b = 0$, линейная связь между переменными отсутствует (линия регрессии параллельна оси абсцисс).

- Свободный член уравнения регрессии *a* интерпретируется, если для независимой переменной значение $x = 0$ имеет смысл. В этом случае $y = a$, если $x = 0$.

Построение уравнения регрессии сводится к оценке ее параметров

- Для оценки параметров регрессий, линейных по параметрам, используют метод наименьших квадратов (МНК).
- МНК позволяет получить такие оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака y от теоретических $\hat{y}$ минимальна, то есть
- $$\sum (y - \hat{y}_x)^2 \to \min$$

- Для линейных и нелинейных уравнений, приводимых к линейным, решается следующая система относительно a и b:
- :
$$\begin{cases} na + b\sum x = \sum y, \\ a\sum x + b\sum x^2 = \sum yx. \end{cases}$$

Можно воспользоваться готовыми формулами, которые вытекают из этой системы:

$$a = \bar{y} - b \cdot \bar{x},$$

$$b = \frac{\operatorname{cov}(x, y)}{\sigma_x^2} = \frac{\overline{y \cdot x} - \bar{y} \cdot \bar{x}}{\overline{x^2} - \bar{x}^2},$$

где $\sigma_x{}^2 = \frac{1}{n-1}\sum (x - \bar{x})^2$ - несмещённая (исправленная) дисперсия величины x .

Тесноту связи изучаемых явлений оценивает **линейный коэффициент парной корреляции** r_{xy} для линейной регрессии $\left(-1 \le r_{xy} \le 1\right)$:

$$r_{xy} = b\frac{\sigma_x}{\sigma_y} = \frac{\mathrm{cov}(x,y)}{\sigma_x\sigma_y} = \frac{\overline{yx} - \bar{y}\cdot\bar{x}}{\sigma_x\sigma_y},$$

и **индекс корреляции** p_{xy} - для нелинейной регрессии $(0 \le p_{xy} \le 1)$:

$$p_{xy} = \sqrt{1 - \frac{\sigma^2_{ост}}{\sigma^2_y}} = \sqrt{1 - \frac{\sum (y - \hat{y}_x)^2}{\sum (y - \bar{y})^2}},$$

где $\sigma_{ост}^{\ 2} = \frac{1}{n-1} \sum (y - \hat{y}_x)^2$,

$$\sigma_y^{\ 2} = \frac{1}{n-1} \sum (y - \bar{y})^2.$$

Оценку качества построенной модели даст **коэффициент (индекс) детерминации**, а также **средняя ошибка аппроксимации**.

Средняя ошибка аппроксимации – среднее отклонение расчетных значений от фактических:

$$\overline{A} = \frac{1}{n}\sum\left|\frac{y-\hat{y}}{y}\right| \cdot 100\%.$$

Допустимый предел значений $\overline{A}$ - не более 8-10%.

Средний коэффициент эластичности $\overline{Э}$ показывает, на сколько процентов в среднем по совокупности изменится результат *y* от своей средней величины при изменении фактора *x* на 1% от своего среднего значения:

$$\overline{Э} = f'(x)\frac{\overline{x}}{\overline{y}}.$$

Задача дисперсионного анализа состоит в анализе дисперсии зависимой переменной:

$$\sum (y - \bar{y})^2 = \sum (\hat{y}_x - \bar{y})^2 + \sum (y - \hat{y}_x)^2,$$

где $\sum (y - \bar{y})^2$ - общая сумма квадратов отклонений;

$\sum (\hat{y}_x - \bar{y})^2$ - сумма квадратов отклонений, обусловленная регрессией («объясненная» или «факторная»);

$\sum (y - \hat{y}_x)^2$ - остаточная сума квадратов отклонений.

Долю дисперсии, объясняемую регрессией, в общей дисперсии результативного признака *y* характеризует коэффициент (индекс) детерминации R^2:

$$R^2 = \frac{\sum(\hat{y}_x - \bar{y})^2}{\sum(y - \bar{y})^2}.$$

Коэффициент детерминации – квадрат коэффициента или индекса корреляции.

F-критерий Фишера

- **Критерий Фишера** (F-критерий, φ*-критерий, критерий наименьшей значимой разности) — апостериорный статистический критерий, используемый для сравнения дисперсий двух вариационных рядов, то есть для определения значимых различий между групповыми средними в установке дисперсионного анализа.

F-критерий Фишера

F-тест – оценивание качества уравнения регрессии – состоит в проверке гипотезы H_0 о статистической незначимости уравнения регрессии и показателя тесноты связи.

Для этого выполняется сравнение фактического $F_{факт}$ и критического (табличного) $F_{табл}$ значений F-критерия Фишера. $F_{фак.}$ определяется из соотношения значений факторной и остаточной дисперсии, рассчитанных на одну степень свободы:

$$F_{факт} = \frac{\sum (\hat{y} - \bar{y})^2 / m}{\sum (y - \hat{y})^2 / (n - m - 1)} = \frac{r_{xy}^2}{1 - r_{xy}^2}(n - 2),$$

где n – число единиц совокупности;

m – число параметров при переменных x.

k_1 - число степеней свободы (факторная вариация результата)

Для парной линейной регрессии

$$k_1 = m = 1$$

m - число параметров при переменной *x*)

$$k_2 = n - m - 1$$

n-m - число степеней свободы *df*

число степеней свободы для парной линейной регрессии

$$df = n - 1$$

$F_{табл}$ - это максимально возможное значение критерия под влиянием случайных факторов при данных степенях свободы и уровне значимости α. Уровень значимости α – вероятность отвергнуть правильную гипотезу при условии, что она верна. Обычно α принимается равной 0,05 или 0,01.

Если $F_{табл} < F_{факт}$, то H_0 - гипотеза о случайной природе оцениваемых характеристик отклоняется и признается их статическая значимость и надежность. Если $F_{табл} > F_{факт}$, то гипотеза H_0 не отклоняется и признается статистическая незначимость, ненадежность уравнения регрессии.

Оценка статистической значимости коэффициентов регрессии и корреляции

Для оценки статистической значимости коэффициентов регрессии и корреляции рассчитываются *t*-критерии Стьюдента и доверительные интервалы каждого из показателей.

Выдвигается гипотеза H_0 о случайной природе показателей, т.е. о незначимом их отличии от нуля.

Оценка значимости коэффициентов регрессии и корреляции с помощью t-критерия Стьюдента проводится путем сопоставления их значений с величиной случайной ошибки:

$$t_b = \frac{b}{m_b};$$

$$t_a = \frac{a}{m_a};$$

$$t_r = \frac{r}{m_r}.$$

Случайные ошибки параметров линейной регрессии и коэффициента корреляции определяются по формулам:

$$m_b = \sqrt{\frac{\sum (y - \hat{y}_x)^2 / (n-2)}{\sum (x - \bar{x})^2}} = \sqrt{\frac{S_{ост}^2}{\sum (x - \bar{x})^2}} = \frac{S_{ост}}{\sigma_x \sqrt{n}};$$

$$m_a = \sqrt{\frac{\sum (y - \hat{y}_x)^2}{(n-2)} \cdot \frac{\sum x^2}{n \sum (x - \bar{x})^2}} = \sqrt{S_{ост}^2 \frac{\sum x^2}{n^2 \sigma_x^2}} = S_{ост} \frac{\sqrt{\sum x^2}}{n \sigma_x};$$

$$m_{r_{xy}} = \sqrt{\frac{1 - r_{xy}^2}{n-2}}.$$

Сравнивая фактическое и критическое (табличное) значение t-статистики - $t_{табл}$ и $t_{факт}$ - принимаем или отвергает гипотезу H_0.

Связь между F-критерием Фишера и t-статистикой Стьюдента выражается равенством

$$t_r^2 = t_b^2 = \sqrt{F}.$$

Если $t_{табл} < t_{факт}$, то гипотеза H_0 отклоняется, т.е. a,b и r_{xy} не случайно отличаются от нуля и сформировались под влиянием систематически действующего фактора x. Если $t_{табл} > t_{факт}$, то гипотеза H_0 не отклоняется и признается случайная природа формирования a,b или r_{xy} .

- **Доверительный интервал** — это интервал, построенный с помощью случайной выборки из распределения с неизвестным параметром, такой, что он накрывает данный параметр с заданной вероятностью.

Для расчета доверительного интервала определяем предельную ошибку Δ для каждого показателя:

$$\Delta_a = t_{табл} m_a,$$

$$\Delta_b = t_{табл} m_b.$$

Формулы для расчета доверительных интервалов имеют следующий вид:

$$\gamma_a = a \pm \Delta_a;$$

$$\gamma_{a_{\min}} = a - \Delta_a;$$

$$\gamma_{a_{\max}} = a + \Delta_a;$$

$$\gamma_b = b \pm \Delta_b;$$

$$\gamma_{b_{\max}} = b + \Delta_b.$$

$$\gamma_{b_{\min}} = b - \Delta_b;$$

Если в границы доверительного интервала попадает ноль, т.е. нижняя граница отрицательна, а верхняя положительна, то оцениваемый параметр принимается нулевым, так как он не может одновременно принимать и положительное, и отрицательное значения.

Прогнозное значение y_p определяется путем подстановки в уравнение регрессии $\hat{y}_x = a + b \cdot x$ соответствующего (прогнозного) значения x_p .

Вычисляется средняя стандартная ошибка прогноза $m_{\hat{y}_p}$:

$$m_{\hat{y}_p} = \sigma_{ост} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x - \bar{x})^2}},$$

где $\sigma_{ост} = \sqrt{\frac{\sum (y - \hat{y})^2}{n - m - 1}}$;

и строится доверительный интервал прогноза:

$$\gamma_{\hat{y}_p} = \hat{y}_p \pm \Delta\hat{y}_p;$$

$$\gamma_{\hat{y}_{p\min}} = \hat{y}_p - \Delta\hat{y}_p;$$

$$\gamma_{\hat{y}_{p\max}} = \hat{y}_p + \Delta\hat{y}_p;$$

где $\Delta\hat{y}_p = t_{табл} \cdot m_{\hat{y}_p}$.

- Все изложенное в данном разделе понадобится Вам при выполнении контрольной работы.

- Желаю удачи!