Лекция 2

Оценка качества уравнения парной регрессии

1. Оценка качества уравнения регрессии

2. Оценка значимости уравнения регрессии в целом

3. Оценка значимости параметров уравнения

4. Интервальные оценки

5. Нелинейная парная регрессия

Оценка точности уравнения регрессии производится на основе *дисперсионного анализа*. Центральное место в линейном дисперсионном анализе занимает разложение общей суммы квадратов отклонений

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

переменной y от среднего $\bar{y}$ на две части - одна из них вызвана влиянием фактора x , другая – прочими неучтенными факторами:

$$Q=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\dot{\tilde{y}}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\tilde{y}_i)^2 \quad (2)$$

1. Оценка качества уравнения регрессии

После того, как найдено уравнение парной регрессии

$$\tilde{y} = b_0 + b_1 x \quad (1)$$

возникает вопрос – насколько точно оно представляет неизвестную связь между переменными y и x, и насколько можно доверять этому уравнению, чтобы уверенно использовать его на практике?

Здесь $Q_R = \sum_{i=1}^{n} (\tilde{y}_i - \bar{y})^2$ – *факторная* сумма (объясненная с помощью регрессии часть), обусловленная влиянием фактора x ,

$Q_e = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 = \sum_{i=1}^{n} e_i^2$ – *остаточная* сумма (необъясненная часть), обусловленная влиянием прочих неучтенных факторов.

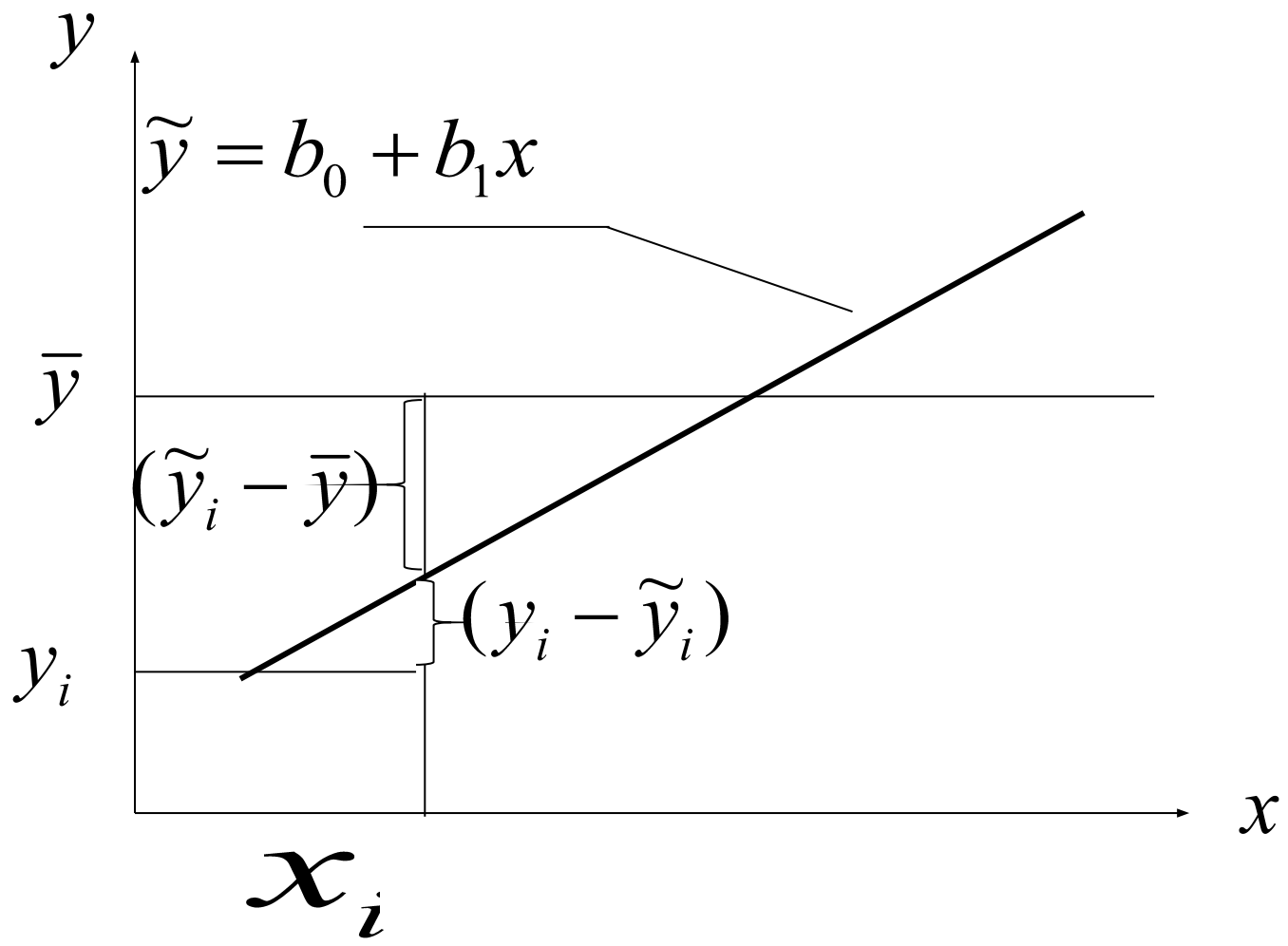

Рис. 1

Если фактор x не оказывает влияния на переменную y, то $Q_R = 0$ и $Q = Q_e$.

Если же $Q_R > Q_e$, то x влияет существенно на признак y.

В связи с этим вводят в рассмотрение одну из наиболее эффективных оценок меры качества уравнения регрессии – *коэффициент детерминации* R^2, который определяется по формуле

$$R^2 = \frac{Q_R}{Q} = \frac{\sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{Q_e}{Q} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \quad (3)$$

Из формулы (3) следует, что $0 \le R^2 \le 1$, а величина R^2 показывает, какая доля вариации переменной y обусловлена вариацией фактора x.

Чем ближе R^2 к единице, тем лучше данная регрессия (1) аппроксимирует модельное уравнение регрессии, тем выше качество модели.

Для линейной парной регрессии (1) коэффициент детерминации можно найти по другой формуле

$$R^2 = r_{xy}^2. \quad (4)$$

Другим критерием оценки качества уравнения регрессии является *средняя относительная ошибка аппроксимации*, определяемая из выражения:

$$A = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \tilde{y}_i}{y_i}\right| \cdot 100\% = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{y_i}\right| \cdot 100\%. \quad (5)$$

Если $A < 10\%$, то это говорит о хорошем качестве модели.

2. Оценка значимости уравнения регрессии в целом

Разделив каждую сумму квадратов соотношения (2) на соответствующее ей число степеней свободы, получим несмещенные оценки этих дисперсий:

$$s_R^2 = \frac{\sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2}{1}; \quad s_e^2 = \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n-2}; \quad s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}.$$

Далее выдвинем гипотезу о равенстве дисперсий

$$H_0 : s_R^2 = s_e^2$$

По-существу это гипотеза об *отсутствии* линейной зависимости между переменными x и y (для наличия такой зависимости требуется, чтобы факторная дисперсия s_R^2 в несколько раз превышала остаточную s_e^2).

Как известно для тестирования данной гипотезы используется статистика

$$F = \frac{s_R^2}{s_e^2} = \frac{Q_R(n-2)}{Q_e} \quad (6)$$

которая представляет собой F –распределение Фишера с $k_1 = 1$ и $k_2 = n-2$ степенями свободы.

Вычисленное по формуле (6) значение статистики F сравнивают с $F_{кр}$, которое находят из таблиц распределения Фишера по заданному уровню значимости α и числам степеней свободы $k_1 = 1$ и $k_2 = n - 2$

Если $F > F_{кр}$, то гипотеза H_0 отклоняется и уравнение регрессии (1) с вероятностью $\gamma = 1 - \alpha$ признаётся *статистически значимым* и его можно использовать на практике. В противном случае ($F \le F_{кр}$) оно не является таковым и, следовательно, непригодно для использования.

На практике для вычисления статистики F применяют другую формулу

$$F = \frac{R^2}{1 - R^2}(n - 2), \quad (7)$$

связывающую величину F и коэффициент детерминации R^2.

3. Оценка значимости параметров уравнения

В линейной регрессии обычно оценивается значимость не только уравнения в целом, но и значимость его параметров. Рассмотрим это на примере параметра b_1, который имеет чёткий экономический смысл. Выдвигаем гипотезу $H_0 : \beta_1 = 0$ (коэффициент регрессии статистически незначим). В качестве альтернативной возьмём $H_1 : \beta_1 \neq 0$, что соответствует двусторонней критической области.

Тогда при выполнении предпосылки 5° МНК доказано, что случайная величина

$$U = \frac{b_1}{\sigma_{b_1}} \quad (8)$$

имеет стандартное нормальное распределение, т.е. $U \sim N(0,1)$.

Нетрудно доказать, что для дисперсии $D(b_1)$ параметра b_1 справедлива формула

$$D(b_1) = \sigma_{b_1}^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad (9)$$

где дисперсия возмущения , т.е.

по предпосылке 3° .

σ_{b_1}

Величина σ^2 неизвестна, а её несмещенной s_e^2 оценкой является выборочная исправленная дисперсия

$$T_{b_1} = \frac{b_1}{s}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

.

Если заменить s в формуле (8) с использованием соотношения (9) на оценку , то получим случайную величину

$k = n - 2$

$$s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

которая имеет распределение Стьюдента с

Введём в рассмотрение величину

$$m_{b_1} = \frac{s_e}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}},$$

которую называют *стандартной ошибкой* параметра b_1 (по-существу она является несмещенной оценкой неизвестного σ_{b_1}). Тогда

$$T_{b_1} = \frac{b_1}{m_{b_1}}. \qquad (10)$$

В итоге проверка гипотезы H_0 сводится к вычислению по формуле (10) наблюдаемого значения статистики t_{b_1} и сравнения её модуля $\left|t_{b_1}\right|$ с критическим значением $t_{кр}$, которое находят из таблицы критических точек распределения Стьюдента по заданному половинному уровню значимости $\alpha/2$ (критическая область двусторонняя) и числу степеней свободы $k = n - 2$

Если выполняется неравенство

$$\left|t_{b_1}\right| = \frac{\left|b_1\right|}{m_{b_1}} > t_{кр},$$

то параметр b_1 считается *статистически значимым* с вероятностью $\gamma = 1 - \alpha$.

В противном случае ($\frac{\left|b_1\right|}{m_{b_1}} \le t_{кр}$) гипотеза H_0 принимается.

Аналогично, если выполняются неравенства

$$\left|t_{b_0}\right| = \frac{\left|b_0\right|}{m_{b_0}} > t_{кр}, \qquad \left|t_r\right| = \frac{\left|r_{xy}\right|}{m_r} > t_{кр},$$

где $m_{b_0} = s_e \sqrt{\dfrac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}} = m_{b_1}\sqrt{\overline{x^2}}$, $m_r = \sqrt{\dfrac{1 - r_{xy}^2}{n-2}} -$

стандартные ошибки параметров b_0 и r_{xy} соответственно, то они признаются *статистически значимыми*.

В заключении отметим, что между наблюдаемыми значениями статистик существует связь:

$$t_{b_1} = t_r = \sqrt{F}.$$

4. Интервальные оценки

Если коэффициент регрессии β_1 является статистически значимым, то для него строят *интервальную оценку*

$$b_1 - t_{кр} m_{b_1} \le \beta_1 \le b_1 + t_{кр} m_{b_1}, \qquad (11)$$

где величины $m_{b_1}, t_{кр}$ уже известны из предыдущих вычислений.

По-существу она является доверительным интервалом, который с доверительной вероятностью $\gamma = 1 - \alpha$ накрывает неизвестное значение коэффициента β_1 и характеризует точность оценивания.

Аналогично строят интервальные оценки для других параметров регрессии:

$$b_0 - t_{кр} m_{b_0} \le \beta_0 \le b_0 + t_{кр} m_{b_0},$$

$$r_{xy} - t_{кр} m_r \le \rho_{xy} \le r_{xy} + t_{кр} m_r.$$

Прогнозирование по адекватному уравнению регрессии представляет собой подстановку в уравнение регрессии прогнозного значения фактора x.

В соответствии с этим зафиксируем некоторое значение объясняющей переменной $x = x_p$ и найдём для неё прогнозное значение зависимой переменной y :

.

$$\tilde{y}_p = b_0 + b_1 x_p .$$

Величина $\tilde{y}_p$ является *точечной* оценкой неизвестного значения y_p, соответствующего значению x_p объясняющей переменной x в природе.

Интервальную оценку для y_p определяют из соотношения

$$\tilde{y}_p - t_{кр} m_{\tilde{y}_p} \le y_p \le \tilde{y}_p + t_{кр} m_{\tilde{y}_p} ,$$

где стандартная ошибка $m_{\tilde{y}_p}$ индивидуального прогнозного значения $\tilde{y}_p$ находится по формуле

$$m_{\tilde{y}_p} = s_e \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \qquad (12)$$

5. Нелинейная парная регрессия

Соотношения между показателями экономических или социальных процессов не всегда можно выразить линейными функциями, ибо при этом могут возникнуть большие ошибки. В этих случаях используют нелинейные регрессии. Различают два класса нелинейных регрессий, используемых в эконометрике:

- регрессии, *линейные* относительно оцениваемых коэффициентов;
- регрессии, *нелинейные* относительно коэффициентов.

Регрессии, линейные относительно коэффициентов

Примерами моделей первого типа являются:

- парабола второго порядка

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon;$$

- равносторонняя гипербола

$$y = \beta_0 + \beta_1 \frac{1}{x} + \varepsilon;$$

- полулогарифмическая функция

$$y = \beta_0 + \beta_1 \ln x + \varepsilon$$

и т.д.

Регрессии, нелинейные относительно коэффициентов

Второй класс представляют функции:

- степенная

$$y = \beta_0 \cdot x^{\beta_1} \cdot \varepsilon;$$

- показательная

$$y = \beta_0 \cdot (\beta_1)^x \cdot \varepsilon;$$

- экспоненциальная

$$y = e^{\beta_0 + \beta_1 x} \cdot \varepsilon$$

и т.п.

Непосредственно МНК для оценки коэффициентов этих моделей применять нельзя, так как *системы нормальных уравнений* уже являются нелинейными и решаются в общем случае только численными приближенными методами.

Для оценки коэффициентов нелинейных моделей используют два подхода. Первый из них основан на *линеаризации модели* и заключается в том, что с помощью подходящих преобразований исходных переменных или (и) исследуемую зависимость представляют в виде линейного соотношения между преобразованными переменными.

Второй подход применяют в том случае, когда линеаризация модели не удаётся и для нахождения оценок коэффициентов приходится применять *численные методы нелинейной оптимизации*.

Вначале рассмотрим пример линеаризации на моделях первого класса, т.е. моделях, линейных по коэффициентам. Возьмём в качестве примера равностороннюю гиперболу

$$y = \beta_0 + \beta_1 \frac{1}{x} + \varepsilon$$

Введём в рассмотрение новую переменную $x' = \frac{1}{x}$

относительно которой уравнение регрессии будет уже линейно

$$y = \beta_0 + \beta_1 x' + \varepsilon.$$

Теперь оценка коэффициентов последнего уравнения может быть выполнена обычным МНК. В итоге получим следующие оценки:

$$b_1 = \frac{\overline{x'y} - \overline{x'} \cdot \overline{y}}{\overline{x'^2} - (\overline{x}')^2},$$

$$b_0 = \overline{y} - b_1 \overline{x}'.$$

Сложнее выполняется линеаризация моделей второго класса. Рассмотрим это на примере степенной регрессии

$$y = \beta_0 \cdot x^{\beta_1} \cdot \varepsilon.$$

Предварительно прологарифмируем обе части уравнения

$$\ln y = \ln \beta_0 + \beta_1 \ln x + \ln \varepsilon$$

и сделаем замену переменных:

$$y' = \ln y, \quad \beta_0' = \ln \beta_0, \quad x' = \ln x, \quad \varepsilon' = \ln \varepsilon.$$

Тогда для новых переменных уравнение будет линейным

$$y' = \beta_0{}' + \beta_1 x' + \varepsilon'.$$

Вновь для оценки его коэффициентов можно применить МНК:

$$b_1 = \frac{\overline{x'y'} - \overline{x'} \cdot \overline{y'}}{\overline{x'^2} - (\overline{x'})^2},$$

$$b_0' = \overline{y}' - b_1 \overline{x}'.$$

Осталось найти оценку β_0

$$b_0' = e^{b_0'}.$$

В итоге осталось получить искомую нелинейную регрессию в виде степенной функции

$$\tilde{y} = b_0 \cdot x^{b_1}.$$

Для оценки тесноты *нелинейной связи* между переменными x и y в моделях, **линейных** по коэффициентам, используют индекс корреляции

$$R = \sqrt{\frac{Q_R}{Q}} = \sqrt{1 - \frac{Q_e}{Q}} = \sqrt{1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}}. \qquad (13)$$

Чем ближе R к единице, тем теснее связь рассматриваемых показателей, тем более надежно уравнение регрессии.

Квадрат R^2 имеет тот же смысл, что и коэффициент детерминации и его называют *индексом детерминации* нелинейной регрессии.

Индекс детерминации R^2 используют для проверки значимости уравнения регрессии в целом по критерию Фишера

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m},$$

где m – число коэффициентов модели при факторе .

Отметим особо, что если модель является **нелинейной по оцениваемым коэффициентам**, то индексы корреляции и детерминации для них *не вычисляются*, ибо для таких моделей **не выполняется** основной постулат линейного дисперсионного анализа о разложении общей суммы квадратов

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

отклонений переменной y от среднего $\bar{y}$ на две части: Q_R и Q_e.

Благодарю за внимание