Проверка качества уравнения регрессии

Лекция

Цели лекции

- Выполнимость теоретических предпосылок
- Анализ расчетных статистических показателей качества
- Интерпретация регрессии

Случайные составляющие коэффициентов регрессии

После определения оценок b_0 и b_1 возникают вопросы:

- насколько точно эмпирическое уравнение регрессии соответствует уравнению для всей генеральной совокупности;
- насколько близки оценки b_0 и b_1 к своим теоретическим значениям β_0 и β_1;
- как близко оцененное значение $\hat{y}_i$ к условному математическому ожиданию $M[Y/X = x_i]$;
- насколько надежны найденные оценки.

Для ответа на эти вопросы необходимы дополнительные исследования.

Свойства оценок коэффициентов регрессии

Оценки b_0 и b_1 представляют собой случайные величины, зависящие от случайного члена в уравнении регрессии.

Рассмотрим теоретическую модель парной линейной регрессии и ее оценку по выборке из n наблюдений:

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad \hat{Y} = b_0 + b_1 X$$

Справедлива формула: $b_1 = \dfrac{S_{xy}}{S_x^2}$

Свойства оценок коэффициентов регрессии

Представим выборочную ковариацию S_{xy} в виде:

$$S_{xy} = \mathrm{Cov}(X,\beta_0+\beta_1 X+\varepsilon) = \mathrm{Cov}(X,\beta_0) + \mathrm{Cov}(X,\beta_1 X) + \mathrm{Cov}(X,\varepsilon) =$$

$$= \beta_1 S_x^{\,2} + \mathrm{Cov}(X,\varepsilon).$$

Следовательно, $b_1 = \dfrac{S_{xy}}{S_x^2} = \beta_1 + \dfrac{S_{x\varepsilon}}{S_x^2},$

где β_1 – постоянная составляющая; $\dfrac{S_{x\varepsilon}}{S_x^2}$ – случайная компонента.

Тот же результат можно получить и для коэффициента b_0.

Свойства оценок коэффициентов регрессии

Т.о. показано, что

Свойства оценок коэффициентов регрессии, а следовательно, и качество построенного уравнения регрессии существенно зависят от свойств случайной составляющей.

Свойства оценок коэффициентов регрессии

Доказано, что для получения по МНК наилучших результатов (при этом оценки b_i обладают свойствами ***состоятельности***, ***несмещенности*** и ***эффективности***) необходимо ***выполнение ряда предпосылок относительно случайного отклонения***.

Предпосылки использования МНК (условия Гаусса – Маркова)

1^0. Случайное отклонение имеет ***нулевое*** математическое ожидание.

2^0. Дисперсия случайного отклонения ***постоянна***.

3^0. Наблюдаемые значения случайных отклонений ***независимы*** друг от друга.

4^0. Случайное отклонение д.б. ***независимо*** от объясняющей переменной.

5^0. Регрессионная модель является ***линейной относительно параметров***, корректно специфицирована и содержит аддитивный случайный член.

Предпосылки использования МНК (условия Гаусса – Маркова)

1^0. Случайное отклонение имеет ***нулевое*** математическое ожидание.

$$M[\varepsilon] = 0$$

Данное условие означает, что случайное отклонение в среднем не оказывает влияния на зависимую переменную.

Предпосылки использования МНК (условия Гаусса – Маркова)

2^0. Дисперсия случайного отклонения ***постоянна***.

$$D[\varepsilon] = \sigma^2 = const$$

Из данного условия следует, что несмотря на то, что при каждом конкретном наблюдении случайное отклонение ε_i может быть различным, но не должно быть причин, вызывающих большую ошибку.

Предпосылки использования МНК (условия Гаусса – Маркова)

3^0. Наблюдаемые значения случайных отклонений ***независимы*** друг от друга.

$$\sigma_{\varepsilon_i \varepsilon_j} = Cov(\varepsilon_i, \varepsilon_j) = \begin{cases} 0, & i \neq j \\ \sigma^2, & i = j \end{cases}$$

Если данное условие выполняется, то говорят об отсутствии автокорреляции.

Предпосылки использования МНК (условия Гаусса – Маркова)

4^0. Случайное отклонение д.б. ***независимо*** от объясняющей переменной.

$$\sigma_{\varepsilon_i X_i} = Cov(\varepsilon_i, X_i) = 0$$

Это условие выполняется, если объясняющая переменная не является случайной в данной модели.

Предпосылки использования МНК (условия Гаусса – Маркова)

5^0. Регрессионная модель является ***линейной относительно параметров***, корректно специфицирована и содержит аддитивный случайный член.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Предпосылки использования МНК (условия Гаусса – Маркова)

6^0. Наряду с выполнимостью указанных предпосылок при построении линейных регрессионных моделей обычно делаются еще ***некоторые предположения***, а именно:

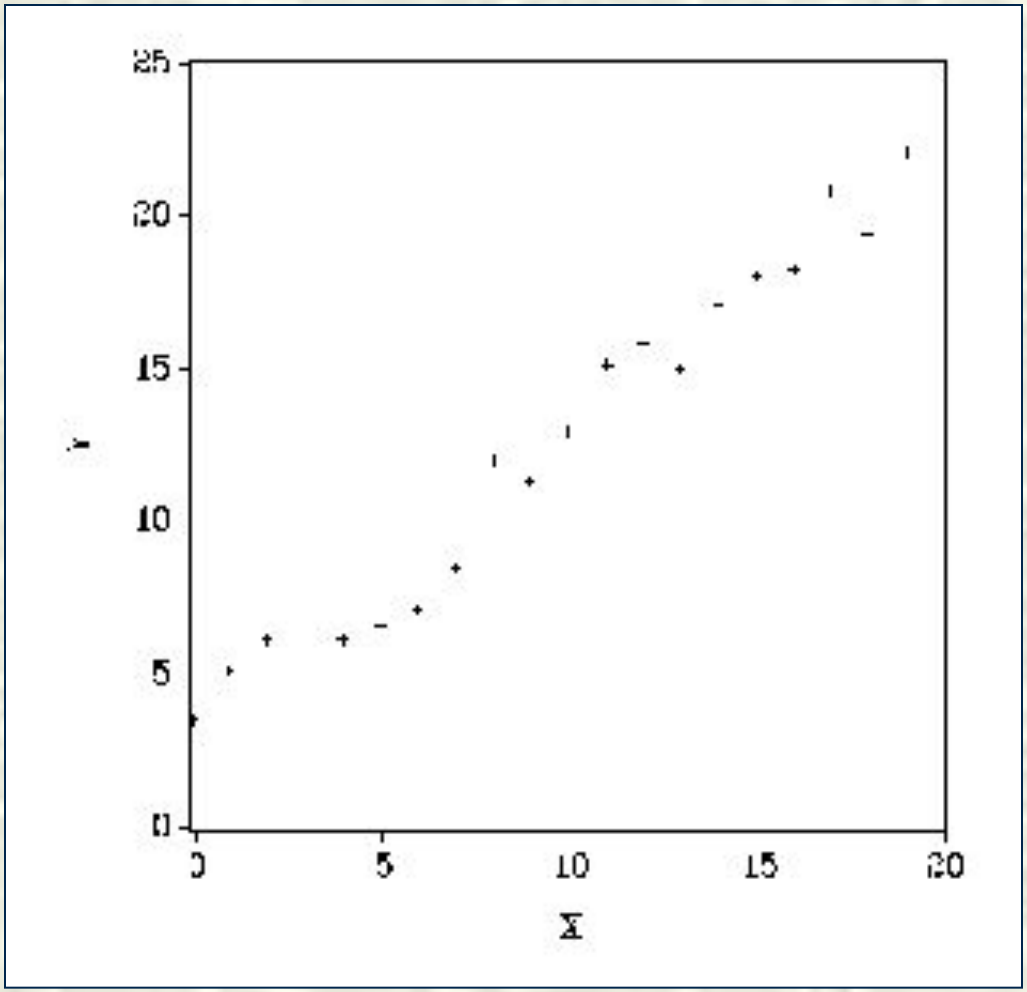
- случайное отклонение имеет нормальный закон распределения;
- число наблюдений существенно больше числа объясняющих переменных;
- отсутствуют ошибки спецификации;
- отсутствует линейная взаимосвязь между двумя или несколькими объясняющими переменными.

Теорема Гаусса - Маркова

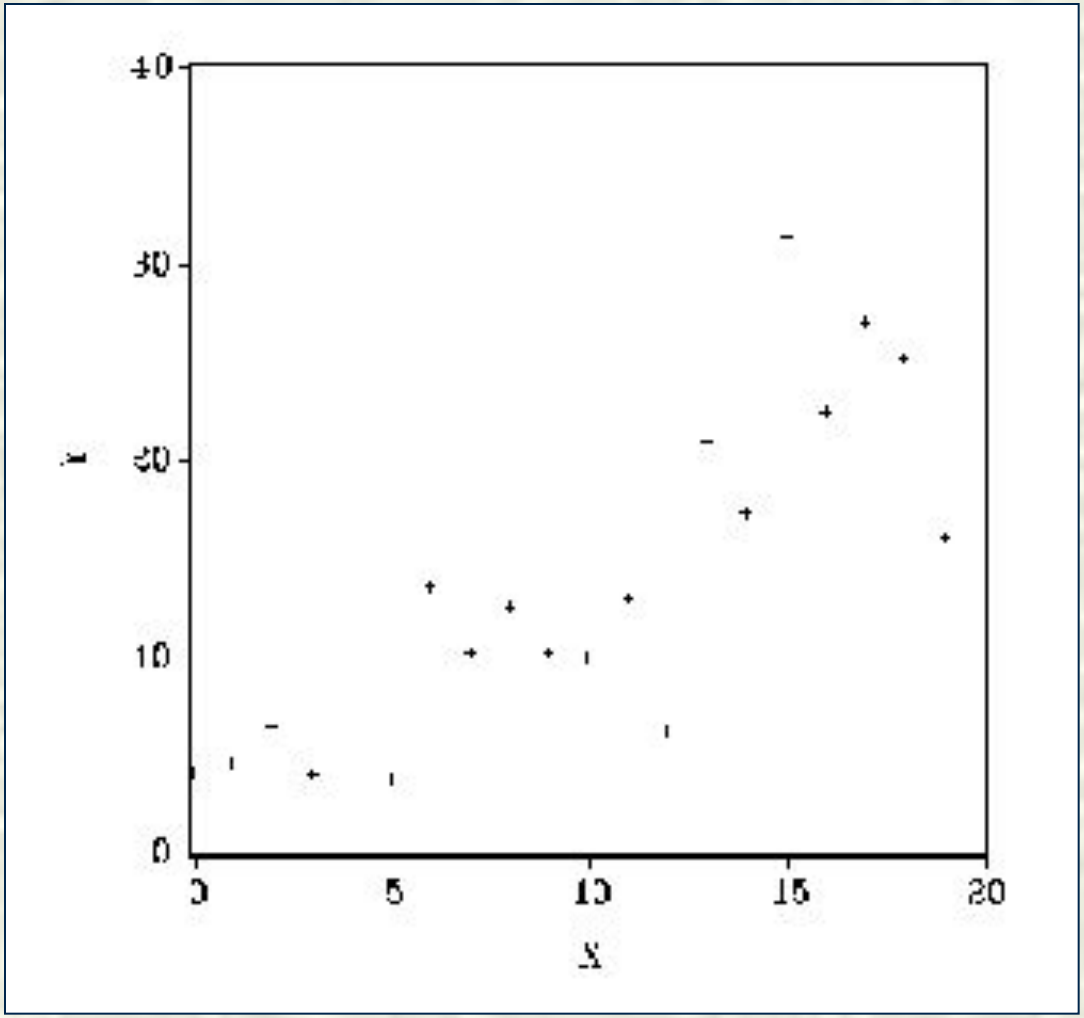
Теорема. Если предпосылки $1^0 – 5^0$ выполнены, то оценки, полученные по МНК, обладают следующими свойствами:

1. Оценки являются ***несмещенными***, т.е. $M[b_0] = \beta_0$, $M[b_1] = \beta_1$. Это говорит об отсутствии систематической ошибки при определении положения линии регрессии.
2. Оценки ***состоятельны***, т.к. при $n \to \infty$ $D[b_0] \to 0$, $D[b_1] \to 0$. Это означает, что с ростом n надежность оценок возрастает.
3. Оценки ***эффективны***, т.е. они имеют наименьшую дисперсию по сравнению с любыми другими оценками данных параметров, линейными относительно величин y_i.

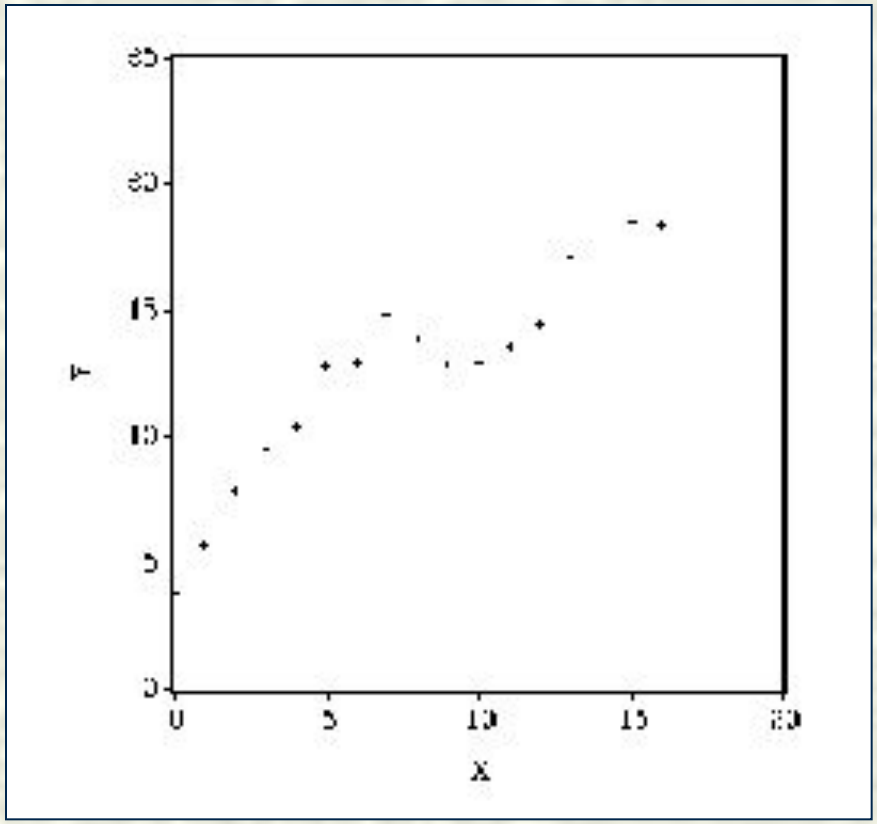
Типичная картина выполнения условий Гаусса – Маркова

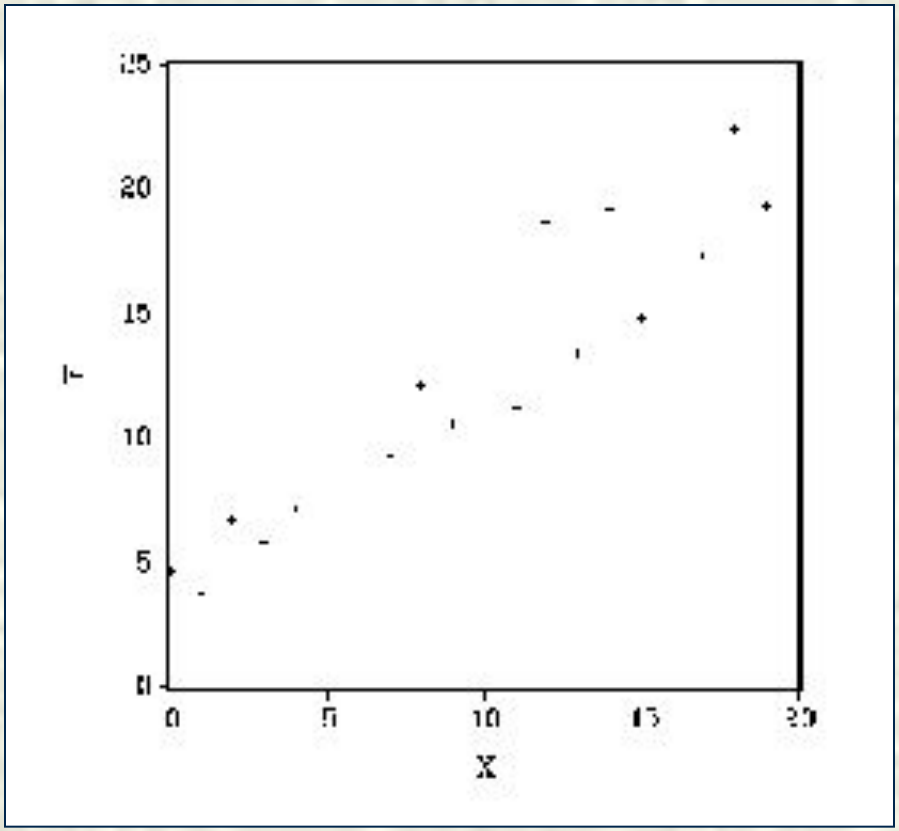


Типичная картина нарушения условий 2^0 и 4^0: $D[\varepsilon] = const$, $Cov(\varepsilon_i, X_i) = 0$



Типичная картина нарушения условия 3^0: $\mathrm{Cov}(\varepsilon_i,\varepsilon_j) = 0,\ i \neq j$





Система показателей качества парной регрессии

1. Показатели качества коэффициентов регрессии

2. Показатели качества уравнения регрессии в целом

3. Адекватность модели – остатки должны удовлетворять условиям теоремы Гаусса-Маркова

Показатели качества коэффициентов регрессии

1. Стандартные ошибки оценок (анализ точности определения оценок).
2. Значения *t*-статистик (проверка гипотез относительно коэффициентов регрессии).
3. Интервальные оценки коэффициентов линейного уравнения регрессии.
4. Доверительные области для зависимой переменной.

Стандартные ошибки оценок

Оценки b_0 и b_1 являются случайными величинами. Отсюда следует, что ***стандартные ошибки*** коэффициентов регрессии – это средние квадратические отклонения коэффициентов регрессии от их истинных значений.

Можно показать, что дисперсии оценок b_0 и b_1 равны:

$$D[b_1] = \frac{\sigma_\varepsilon^2}{\sum (x_i - \bar{x})^2}, \quad D[b_0] = \frac{\sigma_\varepsilon^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}$$

Свойства дисперсий оценок

1. Дисперсии D[b_0] и D[b_1] прямо пропорциональны дисперсии случайного отклонения σ_ε^2. Следовательно, чем больше фактор случайности, тем менее точными будут оценки.
2. Чем больше число наблюдений n, тем меньше дисперсии оценок.
3. Чем больше дисперсия объясняющей переменной, тем меньше дисперсия оценок коэффициентов регрессии. Другими словами, чем шире область изменений объясняющей переменной, тем точнее будут оценки (тем меньше доля случайности в их определении).

Расчет стандартных ошибок

Заменив σ_ε^2 на ее несмещенную оценку

$$\sigma_\varepsilon^2 \approx S_e^2 = \frac{\sum e_i^2}{n-1}$$

получим:

$$D[b_1] \approx S_{b_1}^2 = \frac{S_e^2}{\sum (x_i - \overline{x})^2} \qquad D[b_0] \approx S_{b_0}^2 = \frac{S_e^2 \sum x_i^2}{n \sum (x_i - \overline{x})^2} = \overline{x^2} S_{b_1}^2$$

Формулы расчета стандартных ошибок оценок

Стандартные ошибки коэффициентов регрессии:

$$S_{b_0} = \sqrt{S_{b_0}^2} \qquad S_{b_1} = \sqrt{S_{b_1}^2}$$

Стандартная ошибка является оценкой среднего квадратического отклонения коэффициента регрессии от его истинного значения

Использование стандартных ошибок

Сравнивая значение коэффициента с его стандартной ошибкой, можно судить о значимости коэффициента

Коэффициент называется значимым, если есть достаточно высокая вероятность того, что его истинное значение отлично от нуля

Для стандартных ошибок оценок нет таблиц критических уровней – для точного суждения используются *t*-статистики

Проверка значимости на основе *t*-статистик

Проверка значимости на основе *t*-статистик заключается в установлении наличия линейной зависимости между *Y* и *X*. Данный анализ осуществляется по схеме проверки статистических гипотез. Проверяются альтернативные гипотезы:

$$H_0 : \beta_1 = 0 \quad \text{и} \quad H_1 : \beta_1 \neq 0$$

Проверка значимости на основе *t*-статистик

Если принимается гипотеза H_0, то считают, что величина Y не зависит от X. В этом случае говорят, что коэффициент b_1 ***статистически незначим*** (т.к. слишком близок к нулю). В противном случае говорят, что коэффициент b_1 ***статистически значим***, что указывает на наличие линейной зависимости между Y и X.

Для парной линейной регрессии более важным является анализ статистической значимости коэффициента b_1, т.к. именно в нем скрыто влияние объясняющей переменной X на зависимую переменную Y.

Значимость свободного члена

Аналогично проверяется значимость коэффициента b_0.

Однако мы должны быть осторожны в сильном выделении свободного члена. Почему?

1. Мы обычно не имеем наблюдений вблизи X=0.
2. При отсутствии наблюдений на каком-либо участке оцененная зависимость не может быть данном месте достоверной.

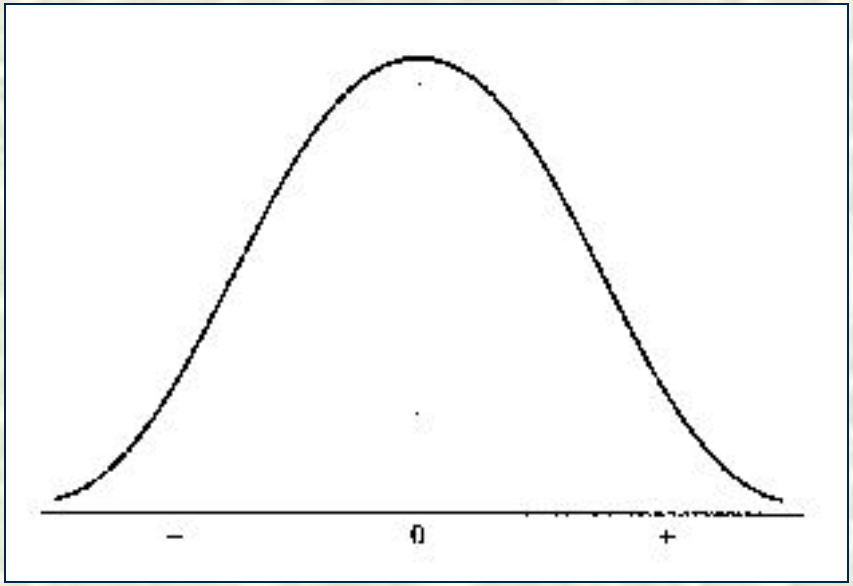
t-статистики для проверки значимости коэффициентов регрессии

t-статистика соизмеряет значение коэффициента с его стандартной ошибкой:

$$t(b_0) = \frac{b_0}{S_{b_0}} \qquad t(b_1) = \frac{b_1}{S_{b_1}}$$

t-статистики для проверки значимости коэффициентов регрессии

t-статистики в парной регрессии по n наблюдениям при справедливости гипотезы H_0 имеют распределение Стьюдента с числом степеней свободы $l = n - 2$

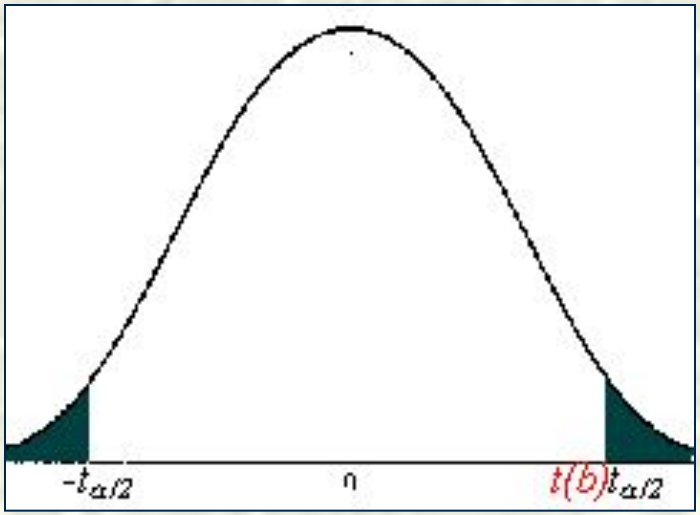


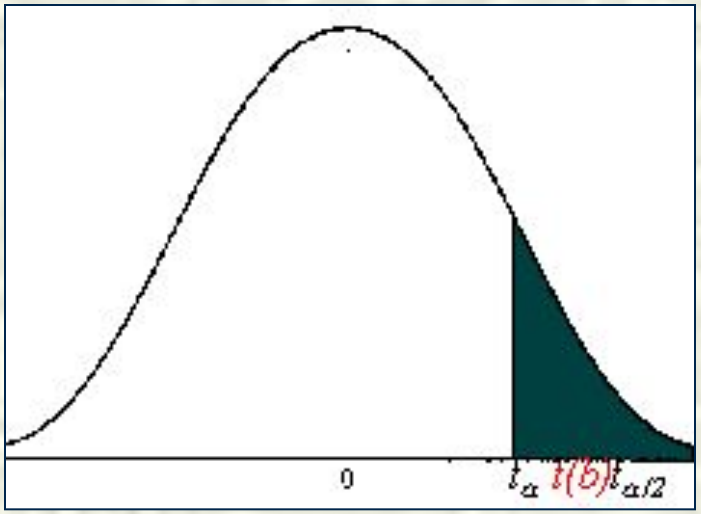
Порядок работы при проверке значимости коэффициента по *t*-статистике

1. Выбираем уровень значимости α (1% или 5%).
2. Вычисляем число степеней свободы ($n-2$).
3. По таблицам распределения Стьюдента определяем критическое значение $t_{\alpha/2;\, n-2}$ (двухсторонний критерий) или $t_{\alpha;\, n-2}$ (односторонний критерий).
4. Если модуль *t*-статистики больше критического значения, то коэффициент является значимым на уровне значимости α.
5. В противном случае коэффициент не значим (на данном уровне α).

Использование односторонних гипотез для проверки значимости коэффициентов

Использование односторонних гипотез иногда позволяет «спасти» значимость коэффициентов регрессии при том же уровне значимости





Это требует обязательного экономического обоснования

Пример (А). Проверка значимости

$$S_{b_1}^2 = \frac{S_e^2}{\sum (x_i - \bar{x})^2} = \frac{\sum e_i^2}{(n-2)\sum (x_i - \bar{x})^2} = \frac{35{,}249}{10 \cdot 2366{,}3} = 0{,}001490$$

$$S_{b_1} = \sqrt{0{,}00149} = 0{,}03860 \Rightarrow t_{b_1} = \frac{b_1}{S_{b_1}} = \frac{0{,}9361}{0{,}0386} = 24{,}25$$

Критическое значение при уровне значимости $\alpha = 0{,}05$:

$$t_{êð} = t_{\frac{\alpha}{2};\, n-2} = t_{0{,}025;10} = 2{,}634$$

Пример (А). Проверка значимости

$$\left|t_{b_1}\right| = 24{,}25 > 2{,}634 = t_{\hat{e}\eth}$$

Поэтому нулевая гипотеза H_0: $\{\beta_1 = 0\}$ отвергается в пользу альтернативной при выбранном уровне значимости. Следовательно, коэффициент регрессии b_1 статистически значим

Аналогично проверяем статистическую значимость коэффициента b_0

Пример (А). Проверка значимости

$$S_{b_0}^2 = \overline{x^2} \cdot S_{b_1}^2 = 0{,}00149 \cdot 15884{,}75 = 23{,}663$$

$$S_{b_0} = \sqrt{23{,}663} = 4{,}864 \qquad t_{b_0} = \frac{b_0}{S_{b_0}} = \frac{3{,}423}{4{,}864} = 0{,}704$$

$$\left|t_{b_0}\right| = 0{,}704 < 2{,}634 = t_{êð}$$

Гипотеза о статистической незначимости b_0 не отклоняется. Это означает, что свободным членом уравнения регрессии можно пренебречь, рассматривая регрессию как $Y = b_1 X$

Правило оценки значимости коэффициентов регрессии без использования таблиц

1. Если $|t_{b_i}| \le 1$, то коэффициент b_i не м.б. признан значимым, т.к. доверительная вероятность менее 0,7.
2. Если $1 < |t_{b_i}| \le 2$, то найденная оценка может рассматриваться как относительно (слабо) значимая. При этом доверительная вероятность лежит между 0,7 и 0,95.
3. Если $2 < |t_{b_i}| \le 3$, то коэффициент значим. Доверительная вероятность лежит между значениями 0,95 и 0,99.
4. Если $|t_{b_i}| > 3$, то это почти полная гарантия значимости коэффициента.

Интервальные оценки коэффициентов линейного уравнения регрессии

Построение доверительных интервалов для коэффициентов линейной регрессии при заданном уровне значимости α:

для β_0: $$\left(b_0 - t_{\frac{\alpha}{2},n-2} \cdot S_{b_0}; b_0 + t_{\frac{\alpha}{2},n-2} \cdot S_{b_0} \right)$$

для β_1: $$\left(b_1 - t_{\frac{\alpha}{2},n-2} \cdot S_{b_1}; b_1 + t_{\frac{\alpha}{2},n-2} \cdot S_{b_1} \right)$$

Доверительные интервалы с надежностью $(1-\alpha)$ накрывают истинные значения β_0 и β_1

Порядок работы при проверке значимости коэффициента по доверительному интервалу

1. Выбираем уровень значимости α (1% или 5%).
2. Вычисляем число степеней свободы ($n-2$).
3. По таблицам распределения Стьюдента определяем критическое значение $t_{\alpha/2;\, n-2}$ (двухсторонний критерий).
4. Вычисляем границы доверительного интервала.
5. Если точка 0 (ноль) не лежит внутри доверительного интервала, то коэффициент является значимым на уровне значимости α.
6. В противном случае коэффициент не значим (на данном уровне α).

Доверительные области для зависимой переменной

Одной из центральных задач эконометрики является прогнозирование значений зависимой переменной при определенных значениях объясняющих переменных. Здесь возможны два варианта:

1. Предсказать условное математическое ожидание зависимой переменной при определенных значениях объясняющих переменных (***предсказание среднего значения***).
2. Предсказать некоторое конкретное значение зависимой переменной (***предсказание конкретного значения***).

Предсказание среднего значения зависимой переменной

Пусть построено уравнение регрессии $\overline{y}(x_i) = b_0 + b_1 x_i$

На его основе необходимо предсказать условное м. о.

$$M[Y / X = x_p] = \beta_0 + \beta_1 x_p$$

переменной Y при $X = x_p$.

Вопрос: Как сильно может уклониться значение $\overline{y}(x_p)$ от $M[Y / X = x_p]$

Предсказание среднего значения зависимой переменной

Доверительная область для условного м. о. $M[Y/X = x_p]$:

$$\left(b_0 + b_1 x_p - t_{\frac{\alpha}{2},n-2} \cdot S_{\overline{y}(x_p)};\quad b_0 + b_1 x_p + t_{\frac{\alpha}{2},n-2} \cdot S_{\overline{y}(x_p)} \right)$$

$$S^2_{\overline{y}(x_p)} = S^2_e \left(\frac{1}{n} + \frac{(x_p - \overline{x})^2}{\sum (x_i - \overline{x})^2} \right)$$

При $x_p = \overline{x}$ она минимальна, а по мере удаления x_p от $\overline{x}$ величина доверительной области увеличивается

Предсказание индивидуальных значений зависимой переменной

Построенная доверительная область для $M_x[Y]$ определяет местоположение модельной линии регрессии (условного м.о.), а не отдельных возможных значений зависимой переменной, которые отклоняются от среднего $\bar{x}$.

Оценка дисперсии индивидуальных значений при $x = x_p$ равна $\hat{y}_p = b_0 + b_1 x_p$

$$S^2_{\hat{y}_p} = S^2_e \left(1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)$$

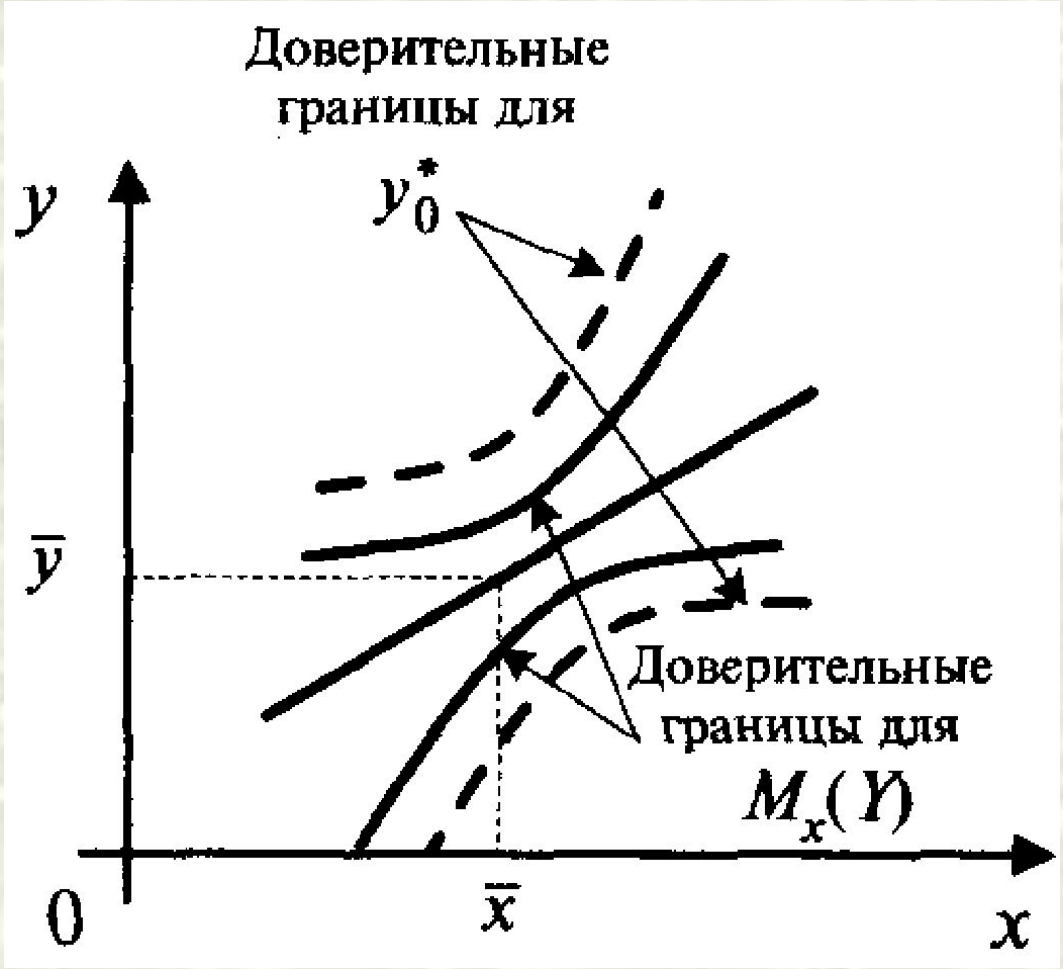
Предсказание индивидуальных значений зависимой переменной

Доверительная область для прогнозов индивидуальных значений y_p^* имеет вид:

$$b_0 + b_1 x_p - t_{\frac{\alpha}{2},n-2} \cdot S_{\hat{y}_p} < y_p^* < b_0 + b_1 x_p + t_{\frac{\alpha}{2},n-2} \cdot S_{\hat{y}_p}$$

Доверительная область для индивидуальных значений y_p^* шире доверительной области для условного м.о. $M[Y / X = x_p]$

Графики доверительных областей для зависимой переменной



Выводы по доверительным областям для зависимой переменной

1. ***Прогноз значений*** зависимой переменной Y по уравнению регрессии ***оправдан***, если значение x объясняющей переменной X ***не выходит за диапазон ее значений по выборке***. Причем, чем ближе x_p к $\bar{x}$, тем точнее прогноз (уже доверительный интервал).

2. ***Использование линии регрессии вне обследованного диапазона значений объясняющей переменной*** (даже если оно оправдано, исходя из смысла решаемой задачи) ***может привести к значительным погрешностям***.

Пример (А). Доверительные области для зависимой переменной

1. Рассчитаем 95%-й доверительный интервал для условного м.о. при $x_p = 160$. Границы интервала равны:

$$3{,}423 + 0{,}9361 \cdot 160 \pm 2{,}634 \cdot 1{,}8775 \cdot \sqrt{\frac{1}{12} + \frac{(160 - 125{,}25)^2}{2366{,}25}}$$

Отсюда среднее потребление при доходе 160 д.е. с вероятностью 95% будет находиться в интервале:

$$(149{,}39; 157{,}01)$$

Пример (А). Доверительные области для зависимой переменной

2. Границы 95%-го доверительного интервала для индивидуальных объемов потребления равны:

$$3{,}423 + 0{,}9361 \cdot 160 \pm 2{,}634 \cdot 1{,}8775 \cdot \sqrt{1 + \frac{1}{12} + \frac{(160 - 125{,}25)^2}{2366{,}25}}$$

Отсюда интервал, в котором будут находиться, по крайней мере 95% индивидуальных объемов потребления при доходе $x_p = 160$, равен:

$$(146{,}96; 159{,}44)$$

Показатели качества уравнения регрессии в целом

Суть проверки общего качества уравнения регрессии – оценить насколько хорошо эмпирическое уравнение регрессии согласуется со статистическими данными.

Основные показатели качества:

1. Коэффициент детерминации R^2.
2. Значение F-статистики.
3. Коэффициент корреляции $r_{xy.}$
4. Сумма квадратов остатков (RSS).
5. Стандартная ошибка регрессии S_e.
6. Средняя ошибка аппроксимации.

Коэффициент детерминации R^2

Коэффициент R^2 показывает долю объясненной вариации зависимой переменной:

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}$$

Используется для предварительной оценки качества модели и как основа для расчета других показателей

Коэффициенты R^2 в разных моделях с разным числом наблюдений (и переменных) несравнимы

Основные свойства коэффициента детерминации

1. $0 \le R^2 \le 1$.
2. Чем ближе R^2 к 1, тем лучше регрессия аппроксимирует статистические данные, тем теснее линейная связь между зависимой и объясняющими переменными.
3. Если $R^2 = 1$, то статистические данные лежат на линии регрессии, т.е. между зависимой и объясняющими переменными имеется функциональная зависимость. Если $R^2 = 0$, то вариация зависимой переменной полностью обусловлена воздействием неучтенных в модели переменных.
4. В случае парной регрессии $R^2 = r_{xy}^{\ 2}$.

Пример (А). Расчет коэффициента детерминации

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{35{,}249}{2108{,}667} = 0{,}983$$

F-тест на качество оценивания уравнения регрессии

Основан на основном тождестве дисперсионного анализа

$$Var(y) = Var(\hat{y}) + Var(e) \qquad \times n \qquad \Rightarrow$$

$$\underbrace{\sum (y_i - \bar{y})^2}_{TSS} = \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{ESS} + \underbrace{\sum e_i^2}_{RSS}$$

TSS – общая сумма квадратов отклонений
ESS – объясненная сумма квадратов отклонений
RSS – необъясненная сумма квадратов отклонений

F-статистика для проверки качества уравнения регрессии

F-статистика представляет собой отношение объясненной суммы квадратов (в расчете на одну независимую переменную) к остаточной сумме квадратов (в расчете на одну степень свободы)

$$F = F_{факт} = \frac{\frac{\sum(\hat{y}_i - \bar{y})^2}{m}}{\frac{\sum(y_i - \hat{y}_i)^2}{n-m-1}} = \frac{\frac{ESS}{m}}{\frac{RSS}{n-m-1}},$$

n – число выборочных наблюдений, *m* – число объясняющих переменных

F-статистика для проверки качества уравнения регрессии

При отсутствии линейной зависимости между зависимой и объясняющими(ей) переменными *F*-статистика имеет *F*-распределение Фишера-Снедекора со степенями свободы $k_1 = m$, $k_2 = n - m - 1$.

Уравнение регрессии значимо на уровне α, если

$$F_{факт} > F_{\alpha;k_1;k_2},$$

где $F_{\alpha;k_1;k_2}$ – табличное значение *F*-распределения, определенное на уровне значимости α при степенях свободы k_1 и k_2

F-статистика для проверки качества парного уравнения регрессии

В парной ($m = 1$) регрессии *F*-статистика является отношением объясненной суммы квадратов к остаточной сумме квадратов (в расчете на одну степень свободы), причем $m = 1$, $n - m - 1 = n - 2$.

$$F = \frac{ESS / m}{RSS / (n - m - 1)} = \frac{(ESS / TSS) / m}{(RSS / TSS) / (n - m - 1)} = \frac{R^2}{(1 - R^2) / (n - 2)}$$

F-статистика в парной регрессии по *n* наблюдениям имеет *F*-распределение с 1 и (n–2) степенями свободы

Порядок работы при проверке значимости парного уравнения по *F*-статистике

1. Выбираем уровень значимости α (1% или 5%).
2. Вычисляем число степеней свободы 1 и $(n-2)$.
3. По таблицам *F*-распределения определяем критическое значение $F_{\alpha;\,1;\,n-2}$ (всегда одностороннее).
4. Если *F*-статистика больше $F_{\alpha;\,1;\,n-2}$, то уравнение в целом является значимым на уровне значимости α.
5. В противном случае уравнение в целом не значимо (на данном уровне α).

Связь между значимостью коэффициента регрессии и уравнения в целом

В парной регрессии *F*-статистика равна квадрату *t*-статистики; то же верно и для их критических уровней (односторонний для *t*-статистики)

$$t^2 = F \qquad \left(t_{\alpha;n-2}\right)^2 = F_{\alpha;1;n-2}$$

В парной регрессии значимость коэффициента регрессии и значимость уравнения в целом эквивалентны

F-статистики в разных моделях с разным числом наблюдений и (или) переменных несравнимы

Коэффициент корреляции r_{xy}

Коэффициент корреляции указывает на наличие (или отсутствие) линейной связи между зависимой и объясняющей переменными

Для проверки гипотезы об отсутствии линейной связи используется тот факт, что величина

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

имеет распределение Стьюдента с $(n–2)$ степенями свободы

Взаимосвязь критериев в парном регрессионном анализе

Коэффициент корреляции по абсолютной величине совпадает с квадратным корнем из коэффициента детерминации

$$\left|r_{xy}\right| = \sqrt{R^2}$$

t-статистики для коэффициента корреляции и коэффициента регрессии b_1 совпадают

Проверка значимости коэффициента регрессии эквивалентна проверке наличия линейной связи

Проверка значимости коэффициента детерминации

Критическое значение R^2 связано с критическим значением F-статистики

$$R^2_{крит} = \frac{mF_{крит}}{mF_{крит} + (n - m - 1)} = \frac{mF_{\alpha;m;n-m-1}}{mF_{\alpha;m;n-m-1} + (n - m - 1)},$$

Проверка значимости коэффициента детерминации эквивалентна проверке значимости уравнения регрессии в целом

Сумма квадратов остатков *RSS*

Является оценкой необъясненной части вариации зависимой переменной

$$RSS = \sum_{i=1}^{n} e_i^2$$

Используется как основная минимизируемая величина в МНК, а также для расчета других показателей

Значения *RSS* в разных моделях с разным числом наблюдений и (или) переменных несравнимы

Стандартная ошибка регрессии S_e

Является оценкой величины квадрата ошибки, приходящейся на одну степень свободы модели

$$S_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n - m - 1}}$$

Используется как основная величина для измерения качества модели (чем она меньше, тем лучше)

Значения S_e в однотипных моделях с разным числом наблюдений и (или) переменных сравнимы

Средняя ошибка аппроксимации *A*

Оценку качества модели дает также средняя ошибка аппроксимации – среднее отклонение расчетных значений $\hat{y}_i$ зависимой переменной от фактических значений y_i

$$A = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\%.$$

Допустимый предел значений *A* – не более 10%. Чем меньше значение A, тем лучше

Значения *A* в моделях с разным числом наблюдений и одинаковым количеством переменных сравнимы

Типичные ошибки в использовании показателей качества регрессии

- Величина коэффициентов регрессии не указывает на силу связи или силу влияния на зависимую переменную
- Значимость коэффициентов по *t*-тестам не позволяет сделать вывод о справедливости тех или иных теорий
- *t*-статистики не указывают на относительную важность коэффициентов регрессии
- *t*-статистики предназначены для использования исключительно для выборки и бесполезны для анализа всей совокупности
- Нельзя сравнивать *t*-статистики, *F*-статистики, коэффициенты детерминации и др. у разных уравнений

Ограниченность простой регрессии

1. Никакая единственная переменная за редкими исключениями не в состоянии хорошо «объяснить» изменения зависимой переменной.
2. Могут существовать несколько одинаково хороших и взаимно противоречивых регрессий.
3. Наконец, линейная форма примитивна.

И тем не менее: Нет ничего лучше по простоте и ясности объяснения парной линейной связи. При равной объясняющей способности из двух моделей мы всегда выбираем более простую.

Конец лекции