2. Линейная модель МР

Модель множественной регресии (МР)

1) $M(Y|x_1, x_2, ..., x_p) = f(x_1, x_2, ..., x_p)$

2) $y = f(x_1, x_2, ..., x_p) + \varepsilon$

3) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon$

$\beta_0 \Rightarrow x_0 = 1 - \text{фиктивная переменная}$

Матричная запись

$$\mathbf{Y} = \mathbf{X}\mathrm{B} + \mathrm{E}$$

$$\mathbf{Y} = [y_1, y_2, \ldots, y_n]'$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$\mathrm{B} = [\beta_0, \beta_1, \beta_2, \ldots, \beta_p]'$$

$$\mathrm{E} = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n]'$$

Выборочная модель МР

1) $y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p + e$

2) $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$

Матричная запись

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

$$\hat{\mathbf{Y}} = \mathbf{XB}$$

$$\mathbf{B} = [b_0, b_1, b_2, \ldots, b_p]'$$

$$E = [e_1, e_2, \ldots, e_n]'$$

Число степеней свободы:

$$\nu = n - p - 1$$

Рекомендация:

$$n \geq 3(p+1)$$

Целесообразно:

объем выборки должен быть в 6-7 раз больше числа независимых переменных.

Классическая нормальная линейная модель МР при выполнении требований

1. Независимые переменные – величины неслучайные, возмущения - есть СВ.
2. МО возмущений равны 0.
3. Дисперсия возмущений постоянна (условие гомоскедастичности).

Классическая нормальная линейная модель МР при выполнении следующих требований

4. Отсутствие автокорреляции в возмущениях и их некоррелированность со всеми НП.

5. Возмущения распределены по нормальному закону.

6. Отсутствие мультиколлинеарности (между НП отсутствует сильная линейная связь).

Для классической нормальной линейной модели МР МНК-оценки параметров модели имеют наименьшую дисперсию в классе линейных несмещенных оценок

2. Оценка параметров линейной модели МР

Метод наименьших квадратов

Оценки параметров ЛММР согласно МНК будем искать из условия:

$$Q = \sum e_i^2 =$$

$$= \sum \left(y_i - b_0 - b_1 x_1 - b_2 x_2 - \ldots - b_p x_p \right)^2 \to \min$$

Условиями минимума функции являются равенство нулю первых производных по коэффициентам УМР.

Матричный метод

1) $$\mathbf{B} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} =$$

$$= \mathbf{A}^{-1}\mathbf{X}'\mathbf{Y}$$

2) $$\mathbf{A} = \mathbf{X}'\mathbf{X}$$

3) $$Rank(\mathbf{X}) = p + 1 < n$$

Гомоскедастичность

1) $D(\varepsilon_i) = D(\varepsilon_j)$ для любых i и j

Гетероскедастичность

2) $D(\varepsilon_i) \neq D(\varepsilon_j)$ для любых i и j

Понятие автокорреляции

Определение. *Автокорреляцией* (serial correlation) *порядка p в ошибках модели регрессии* называется зависимость ошибок вида

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \cdots + \rho_p \varepsilon_{t-p} + u_t,$$

где u_t удовлетворяет условиям теоремы Гаусса – Маркова.

Обобщенный метод наименьших квадратов

Теорема. Если в схеме Гаусса-Маркова не выполняется предпосылка о гомоскедастичности и некорелированности случайных возмущений, то наилучшей линейной процедурой оценки параметров модели является:

$$\mathbf{b} = \left(\mathbf{X}^T \mathbf{P}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{P}^{-1} \mathbf{Y}$$

P - матрица ковариаций случайных возмущений

(положительно определенная матрица)

Взвешенный метод наименьших квадратов

Теорема. Если в схеме Гаусса-Маркова не выполняется предпосылка о гомоскедастичности случайных возмущений, то наилучшей линейной процедурой оценки параметров модели является:

$$\mathbf{b} = \left(\mathbf{X}^T \mathbf{P}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{P}^{-1} \mathbf{Y}$$

P - матрица ковариаций случайных возмущений :

$$\mathbf{P} = \begin{pmatrix} \sigma^2(e_1) & 0 & \dots & 0 \\ 0 & \sigma^2(e_2) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma^2(e_n) \end{pmatrix}$$

Взвешенный метод наименьших квадратов

Определение. Пусть заданы положительные числа $\{w_i\}_{i=1}^{n}$ (веса). Оценки коэффициентов модели регрессии, полученные минимизацией *взвешенной суммы квадратов отклонений*

$$\sum_{i=1}^{n} w_i \left(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik}\right)^2 \longrightarrow \min,$$

называются *WLS-оценками* (WLS = Weighted Least Squares) или *оценками взвешенного метода наименьших квадратов*.

Мультиколлинеарность (МТК) – это явление высокой взаимной коррелированности НП.

Два вида МТК:

1) совершенная (строгая, полная)

2) несовершенная (частичная)

Полная МТК при наличии функциональных связей между НП.

Это нарушение требования к рангу матрицы:

1) $$\operatorname{rank} \mathbf{X} < p+1$$

2) $$\det \mathbf{X}'\mathbf{X} = 0$$

$$\mathbf{B} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$$

Последствия МТК:

- Оценки коэффициентов УМР ненадежны и неустойчивы (увеличиваются стандартные ошибки оценок и уменьшаются t-статистики МНК-оценок)
- МНК-оценки коэффициентов неустойчивы (чувствительны к изменениям данных и размерности выборки)
- Возможность получения неверного знака у коэффициентов регрессии

Последствия МТК:

- Оценки коэффициентов УМР становятся очень чувствительными к ошибкам спец.
- Осложнение процесса определения наиболее существенных факторов
- Затрудняет экономическую интерпретацию коэффициентов УМР (выделение характеристик влияния факторов на ЗП в чистом виде)

ОДНАКО:

- Оценки коэффициентов остаются несмещенными
- Оценки коэффициентов немультикол.

факторов не ухудшаются

Причины возникновения МТК:

- НП характеризуют одну и туже сторону экон. процесса
- Использование в модели НП, суммарное значение которых есть постоянная величина
- НП, являющиеся элементами друг друга
- НП могут иметь общий временный тренд, относительно которой они совершают малые колебания

Наблюдается фиктивная (ложная) линейная связь.

Методы устранения мультиколлинеарности

Переход к смещенным методам оценивания

<u>«Ридж – регрессия» («гребневая регрессия»)</u>

$$\mathbf{B}_\tau = \left(\mathbf{X}'\mathbf{X} + \tau \mathbf{I}_{p+1}\right)^{-1} \mathbf{X}'\mathbf{Y}$$

$$\tau_0 \approx 0.1 - 0.4$$

Регрессия на главных компонентах

2. **Метод главных компонент** – переход к новым объясняющим переменным, линейным комбинациям старых:

1) Центрирование переменных $X_ц = X - \overline{X}$, $Y_ц = Y - \overline{Y}$;

2) Решение характеристического уравнения $\left|\sum - \lambda E\right| = 0$:

a) Нахождение собственных чисел $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$,

b) Нахождение для каждого собственного числа λ_j собственного вектора $l^{(j)}$;

3) Переход к новым переменным $Z = X_ц L$, $X_ц = ZL^{-1} = ZL^T$;

4) Построение линейной регрессии $Y_ц = ZC$, вычисление оценок с помощью МНК

$\hat{C} = \left(Z^T Z\right)^{-1} Z^T Y = diag\left\{1/\lambda_j\right\} Z^T Y$;

5) Проверка гипотез H_{0j}: $c_j = 0$, $j = 1, \ldots, p$, исключение несущественных переменных;

6) При необходимости переход к исходной модели $\hat{\theta}_j = \sum\limits_{k \in K_{сущ}} \hat{c}_k l_k^{(j)}$, $\hat{\theta}_0 = \overline{y} - \sum\limits_{j=1}^{p} \hat{\theta}_j \overline{x}^{(j)}$.

Тема №1. Линейные эконометрические модели

4. Оценивание параметров ЭМ с учетом ограничений

Ограничения

1) $$b_i > 0$$

2) $$\mathbf{RB} = \mathbf{r}$$

3) $$b_2 = b_3 \qquad b_2 + b_3 + 2b_4 = 2$$

$$r = \begin{bmatrix} 0 \\ 2 \end{bmatrix} \qquad R = \begin{bmatrix} 0 & 1 & -1 & 0 & \cdots \\ 0 & 1 & 1 & 2 & \cdots \end{bmatrix}$$

Целевая функция

$$1)\quad \min_{b_i}\left(\sum_{j=1}^{n} e_j^2 = (\mathbf{Y}-\mathbf{XB})'(\mathbf{Y}-\mathbf{XB})\right)$$

при ограничениях

$$2)\quad b_{i\,\min} \le b_3 \le b_{i\,\max}, \qquad \mathbf{RB} = \mathbf{r}$$

$$3)\quad \hat{\mathbf{b}} = \mathbf{b} + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{R}^T \times$$

$$\times\left[\mathbf{R}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{R}^T\right]^{-1}\left(\mathbf{r}-\mathbf{Rb}\right)$$

Тема №1. Линейные эконометрические модели

5. Проверка качества линейных моделей МР

(самостоятельная проработка)

Проверка статических гипотез

Виды статистических гипотез:

1) Нулевые (основные);

2) альтернативные (конкурирующие).

Нулевая гипотеза H_0 - проверяемая гипотеза

Альтернативная гипотеза H_1 - гипотеза, которая противоречит нулевой.

Проверка статических гипотез

Ошибка первого рода – отвергается нулевая гипотеза, когда она верна. Обозначают α и наз. уровнем значимости (размером критерия)

Ошибка второго рода –принимается нулевая гипотеза, когда верна альтернативная гипотеза. Обозначают β и вероятность 1- β не совершить ошибку второго рода наз. мощностью критерия

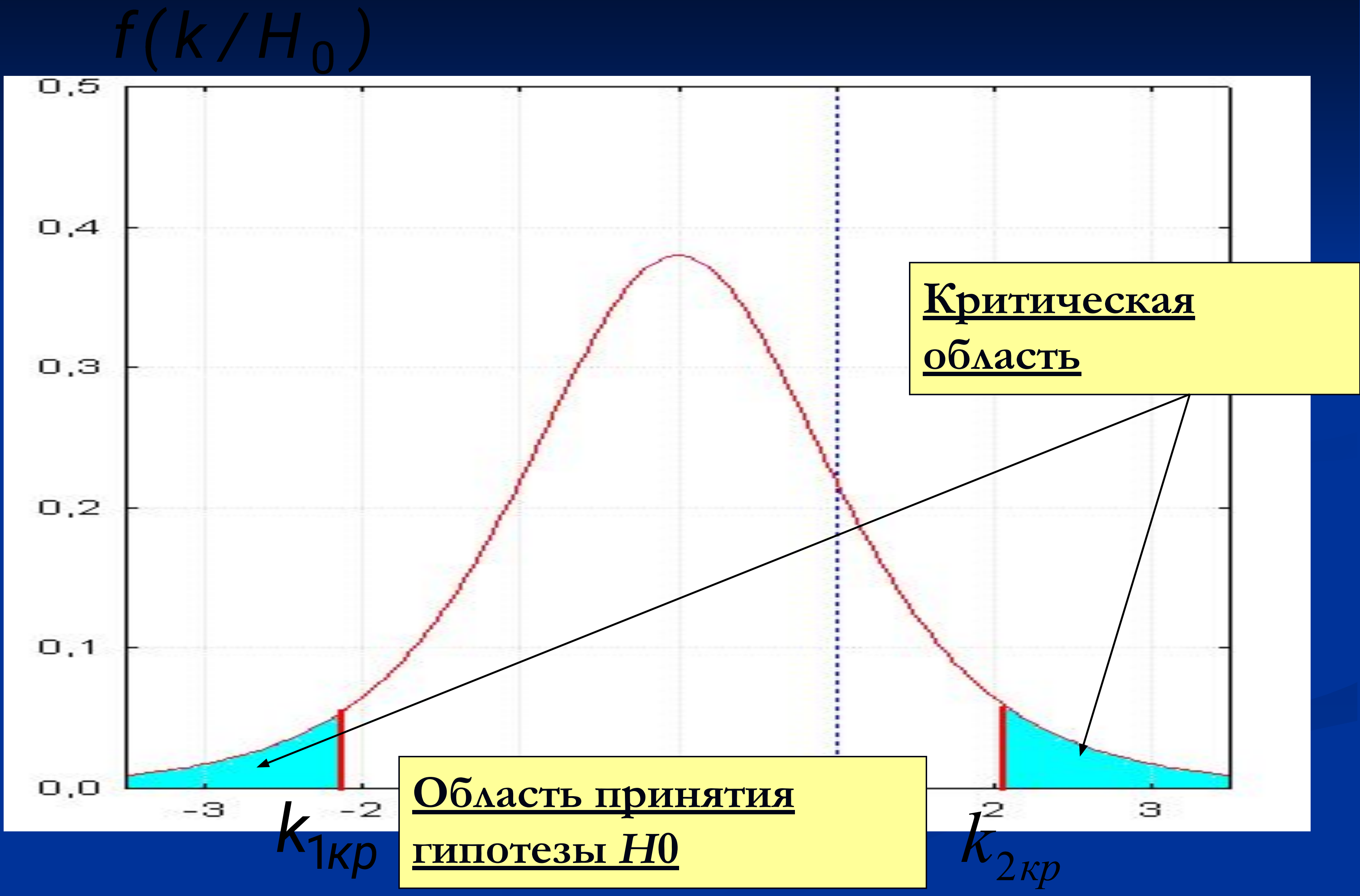
Проверки статистических гипотез

Случайная величина К, построенная по результатам наблюдений для проверки нулевой гипотезы, наз. статистическим критерием.

Основа схемы построения статистического критерия – разделение выборочного пространства на две области:

1) область отклонения нулевой гипотезы (критическая обл.)
2) область принятия нулевой гипотезы

Статистическая проверка гипотез



Проверка статических гипотез

Статистический критерий определяется заданием:

1) статистической гипотезы $H0$;
2) уровня значимости α;
3) статистики критерия;
4) критической области.

Критерий наз. наиболее мощным, если из всех критериев с заданным уровнем значимости он обладает наибольшей мощностью

Два способа проверки гипотез:

1. Нахождением критических точек, соответствующих заданным уровням значимости α .

2. Нахождением уровня значимости *P* (*значимой вероятности*) , соответствующего наблюдаемому значению статистического критерия.

Если значимость Р меньше заданного стандартного уровня значимости α , то нулевая гипотеза отвергается в пользу альтернативной.

Общая схема проверки гипотез

1. Формулировка нулевой и альтернативной гипотез

2. Задается уровень значимости α

3. Определение объема выборки *n*.

4. Выбор статистического критерия для проверки нулевой гипотезы.

5. Определение критической области и области принятия гипотезы.

6. Вычисление наблюдаемого значения статистического критерия.

7. Принятие статистического решения.

Проверка гипотез при двусторонней КО тесно связано с интервальным оцениванием.

Два способа проверки гипотез:

1. Нахождением критических точек, соответствующих заданным уровням значимости α .

2. Нахождением уровня значимости *Р* (*значимой вероятности*) , соответствующего наблюдаемому значению статистического критерия.

Если значимость Р меньше заданного стандартного уровня значимости α , то нулевая гипотеза отвергается в пользу альтернативной.

Качество подгонки данных моделью

$$\sum\left(y_i-\bar{y}\right)^2=\sum\left(\hat{y}_i-\bar{y}\right)^2+\sum\left(y_i-\hat{y}_i\right)^2$$

Общая СумКО (TSS)

$$df=n-1$$

Факторная СумКО (ESS)

$$df=p$$

Остаточная СумКО (RSS)

$$df=n-p-1$$

Коэффициент множественной детерминации

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TTS} \quad \in [0,1]$$

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

Скорректированный коэффициент множественной детерминации

1) $$\overline{R}^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)} \le 1$$

2)

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} = R^2 - \frac{p}{n-p-1}(1 - R^2)$$

Скорректированные коэффициенты в разных моделях с разным числом наблюдений (и переменных) ограниченно сравнимы

Множественный коэффициент корреляции

1) $$R = \sqrt{R^2} \in [0,1]$$

2) $$R \geq \max\{r_{yx_i}, i = \overline{1, p}\}$$

Проверка значимости коэффициентов

Проверка статических гипотез

Статистический критерий определяется заданием:

1) **статистической гипотезы *H*0;**
2) **уровня значимости *α*;**
3) **статистики критерия;**
4) **критической области.**

Статистическая проверка гипотез

Проверка статических гипотез

Ошибка первого рода – отвергается нулевая гипотеза, когда она верна. Обозначают α и наз. уровнем значимости (размером критерия)

Ошибка второго рода –принимается нулевая гипотеза, когда верна альтернативная гипотеза. Обозначают β и вероятность 1- β не совершить ошибку второго рода наз. мощностью критерия

Проверка значимости коэффициент b_0

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.

1) $H_0: \beta_j = 0; \quad H_1: \beta_j \neq 0$

2) $t_{b_j} = \dfrac{b_j}{m_{b_j}} \quad , j = \overline{0, p}$

3) $m_{b_j} = S_e \sqrt{z_{jj}}$

$$S_e = \sqrt{\sum \left(y - \hat{y}_x\right)^2 / \left(n - p - 1\right)}$$

z_{jj} – диагональные элементы матрицы

$$\mathbf{Z} = \left[z_{ij}\right]_{i,j=\overline{0,p}} = (\mathbf{X}'\mathbf{X})^{-1}$$

Проверка значимости коэф. УМР

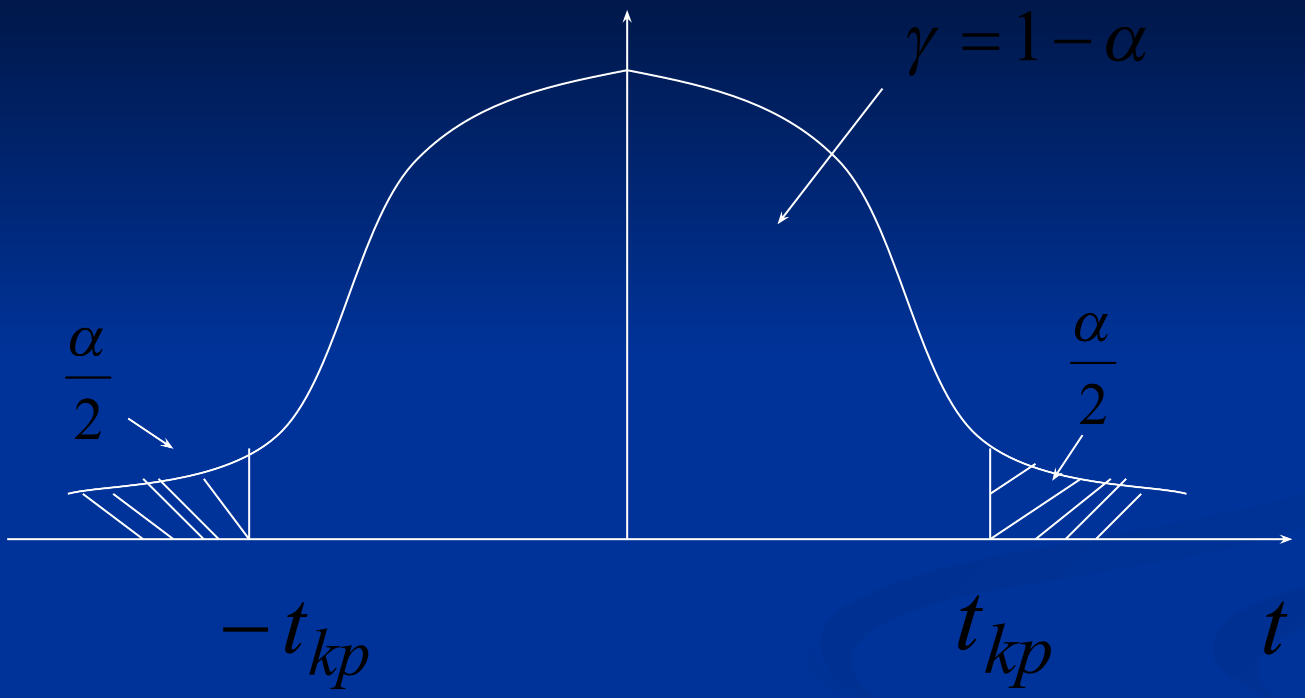
1) $\left|t_{b_j}\right| > t_{\text{табл}}(\alpha; n-p-1)$ - параметр значим

Общий случай

2) $H_0 : b_j = \beta_j; \quad H_1 : b_j \neq \beta_j$

3) $t_{b_j} = \dfrac{b_j - \beta_j}{m_{b_j}} \quad , j = \overline{0, p}$

Вычисление критической точки РС



Доверительный интервал

1) $$b_j - \Delta_{b_j} < \beta_j < b_j + \Delta_{b_j}$$

2) **Предельная ошибка**

$$\Delta_{b_j} = t_{\text{табл}}(\alpha; n-p-1) \cdot m_{b_j}$$

Односторонние проверка значимости коэф. УМР

1) $H_0 : b_j \le \beta_j; \quad H_1 : b_j > \beta_j$

2) $t_{b_j} = \dfrac{b_j - \beta_j}{m_{b_j}}, j = \overline{0, p}$

3) $t_{b_j} > t_{\text{табл}}(2\alpha; n - p - 1)$ - прин. H1

Односторонние гипотезы

1) $H_0 : b_j \le \beta_j; \quad H_1 : b_j > \beta_j$ - правосторонняя

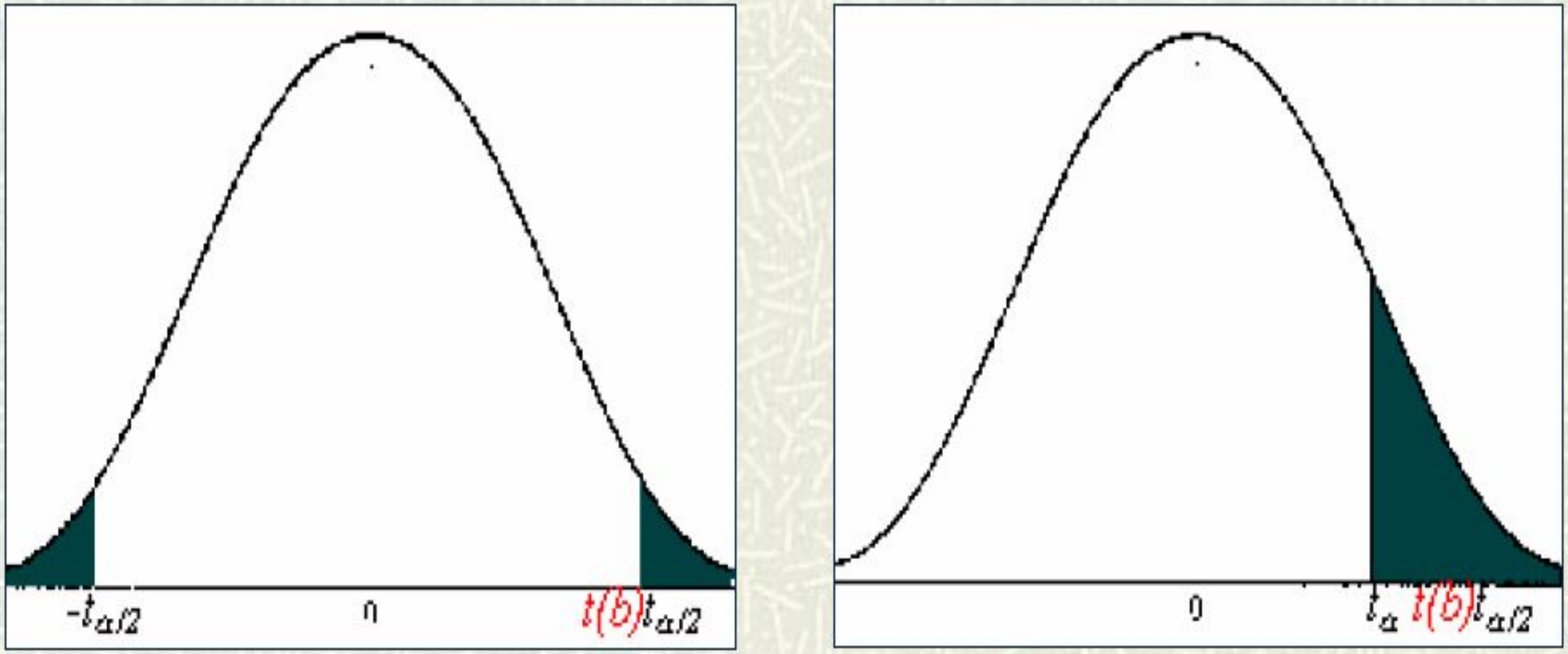
2) $H_0 : b_j \ge \beta_j; \quad H_1 : b_j < \beta_j$ - левосторонняя

Если имеется информация о знаках коэф.

3) Знак «+» $H_0 : b_j \le 0; \quad H_1 : b_j > 0$

Знак «-» $H_0 : b_j \ge 0; \quad H_1 : b_j < 0$

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



Это требует обязательного экономического обоснования

Оценка значимости УМР

1) $H_0 : D_{факт.} = D_{ост.} \quad H_1 : D_{факт.} > D_{ост.}$

2) $F = \dfrac{D_{факт.}}{D_{ост.}} \sim F(\nu_1 = p; \nu_2 = n - p - 1)$

$$D_{факт.} = \frac{ESS}{p} \qquad D_{ост.} = \frac{RSS}{n - p - 1}$$

3) $F > F_{табл}(\alpha; p; n - p - 1) \Rightarrow (модель\ значима)$

Тестирование одного линейного ограничения

1) $$\beta_1 + \beta_2 + \ldots + \beta_k = 1$$

Общий случай

2) $$r_1\beta_1 + r_2\beta_2 + \ldots + r_k\beta_k = \mathbf{r}^T\boldsymbol{\beta}$$

$$H0: \mathbf{r}^T\boldsymbol{\beta} = q \qquad H1: \mathbf{r}^T\boldsymbol{\beta} \neq q$$

4) $$t_{набл} = \frac{\mathbf{r}^T\mathbf{b} - q}{S_e\sqrt{\mathbf{r}^T(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{r}}}$$

$$\left|t_{набл}\right| > t_{\text{табл}}\left(\alpha; n-k\right)$$ - прин. H1

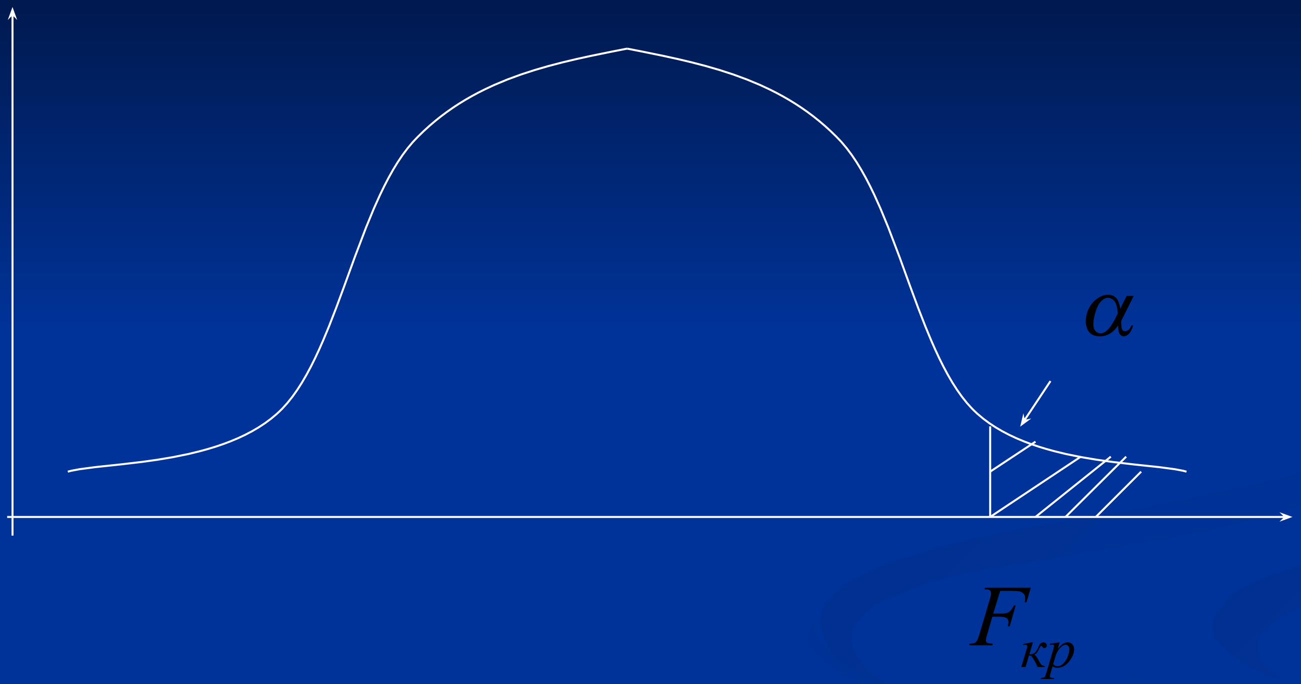
Анализ значимости коэффициента множественной детерминации

1) $H_0 : R^2 = 0 \qquad H_1 : R^2 > 0$

2) $$F = \frac{R^2}{1 - R^2} \cdot \frac{n - p - 1}{p}$$

3) $F > F_{табл}(\alpha; p; n - p - 1)$ - модель значима

Вычисление критической точки РФ для правосторонней критической области



$$F_{кр} = F_{табл}\left(\alpha;\, p;\, n - p - 1\right)$$

Определение точности РМ

Меры точности РМ:

1. Средняя квадратическая ошибка остаточной компоненты на 1 степень свободы (стандартная ошибка регрессии RMSE)

$$RMSE = S_e = \sqrt{\sum \left(y - \hat{y}_x\right)^2 / \left(n - p - 1\right)}$$

Основная величина для измерения качества модели (чем она меньше, тем лучше)

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

Определение точности РМ

2. Средняя относительная ошибка аппроксимации

$$E_{относ.} = \frac{1}{n} \cdot \sum \left| \frac{e_i}{y_i} \right| \times 100$$

$$E_{относ.} < 5 - 7\%$$ **- хороший подбор РМ**

Типичные ошибки в использовании показателей качества ПР

- 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

Конец лекции