# Оценка
## значимости уравнения парной линейной регрессии
## (идентификация )

После того, как получено уравнение линейной регрессии, обязательно проводится оценка его качества и значимости коэффициентов на основе проверки гипотез

- Статистическая гипотеза (SH) – это предположение о величине параметра распределения генеральной совокупности.

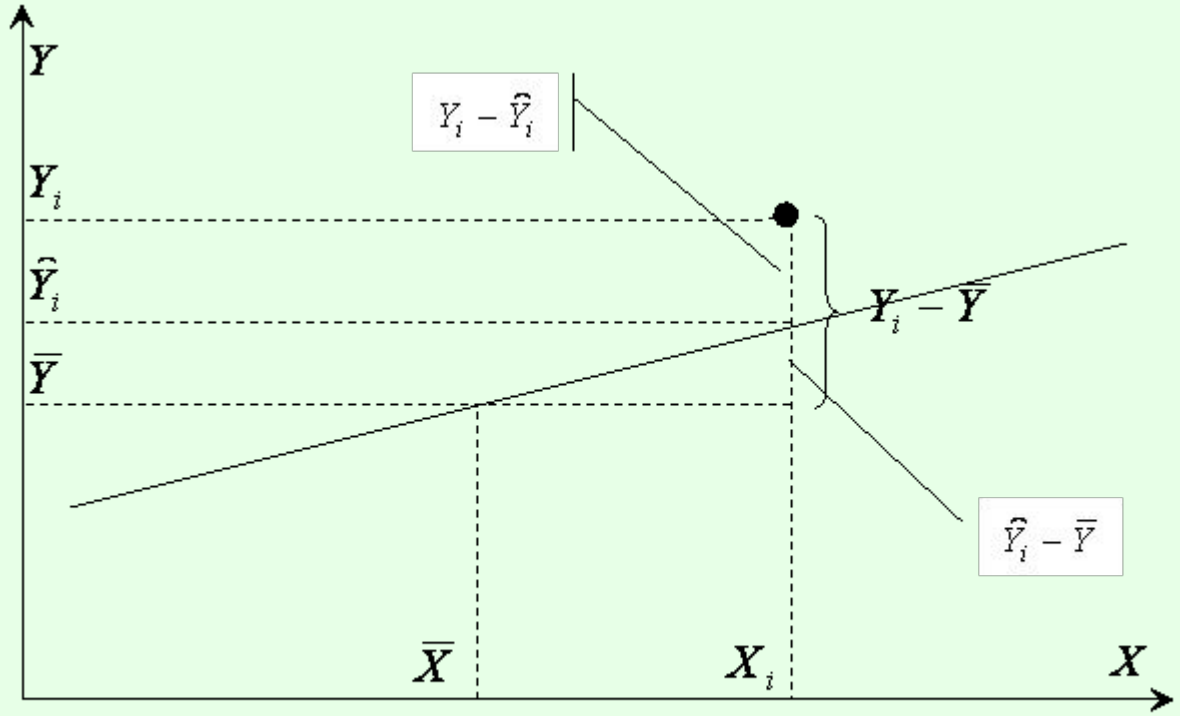
Проверка (SH) осуществляется на базе двух типов гипотез:

нулевая H0 – допущение, которое считается верным до тех пор, пока не будет доказано обратное, исходя из результатов статистической проверки. В частности, предположение о случайной природе оцениваемых параметров, т.е. о незначимом их отличии от нуля.

альтернативная $H_1$ - гипотеза, которая принимается, если в результате проверки отвергается нулевая гипотеза. В частности, это принятие предположения о неслучайной природе оцениваемых параметров, т.е. их статистическая значимость и надежность: не случайно отличаются от нуля и сформировались под влиянием систематически действующего фактора.

- Ошибки 1-го рода – вероятность отвержения гипотезы $H_0$, когда она должна быть принята.
- Ошибка 2-го рода – вероятность принятия гипотезы $H_0$, когда она должна быть отвергнута .

# Разложение отклонения от среднего



# Общая вариация переменной $Y$

величина,
являющаяся мерой
вариации
переменной $Y$ вокруг
ее среднего значения

$$\sum_{i=1}^{N} (Y_i - \bar{Y})^2$$

# Центральное место при этом занимает анализ трех сумм:

$$TSS = \sum_i (y_i - \bar{y})^2$$

- общая сумма квадратов отклонений изучаемого показателя y от его среднего арифметического значения (*total sum of squares*)

$$RSS = \sum_i (\hat{y}_i - \bar{y})^2$$

- сумма квадратов отклонений y, объясняемая регрессией, от среднего арифметического значения изучаемого показателя y (*regression sum of squares*)

$$ESS = \sum_i (y_i - \hat{y}_i)^2$$

- остаточная сумма квадратов отклонений y, объясняемая влиянием неучтенных при моделировании факторов (*error sum of squares*)

# Разложение общей вариации переменной *Y*

$$\sum_{i=1}^{N}(Y_i-\bar{Y})^2=\sum_{i=1}^{N}(Y_i-\hat{Y}_i+\hat{Y}_i-\bar{Y})^2=$$

$$=\sum_{i=1}^{N}(Y_i-\hat{Y}_i)^2-2\sum_{i=1}^{N}(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y})+\sum_{i=1}^{N}(\hat{Y}_i-\bar{Y})^2$$

I II III

В этой сумме II = 0. Тогда:

$$\sum_{i=1}^{N}(Y_i-\bar{Y})^2=\sum_{i=1}^{N}(Y_i-\hat{Y}_i)^2+\sum_{i=1}^{N}(\hat{Y}_i-\bar{Y})^2$$

TSS ESS RSS

$$\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2$$

TSS ESS RSS

- TSS – total sum of squares – вся дисперсия или вариация *Y*, характеризует степень случайного разброса значений функции регрессии около среднего значения *Y*
- ESS – error sum of squares – есть сумма квадратов остатков регрессии, та величина, которую мы минимизируем при построении прямой, часть дисперсии, которая нашим уравнением не объясняется
- RSS – regression sum of squares – объясненная часть общей вариации

Для линейной регрессии :

$$TSS = RSS + ESS$$

Для оценки качества линейной регрессии используют **коэффициент детерминации**

-это величина:

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$$

- показывает долю дисперсии, объясняемую регрессией, в общей дисперсии ***У***

# Связь коэффициента детерминации с коэффициентом корреляции

$$R^2 = r_{xy}^2 = b^2 \frac{\sigma_x^2}{\sigma_y^2}$$

# Свойства коэффициента детерминации

$$0 \le R^2 = \frac{b^2 \sigma_x^2}{\sigma_y^2} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \frac{RSS}{TSS} \le 1$$

Суммы квадратов отклонений (TSS, RSS, ESS) имеют определенное число степеней свободы

**Число степеней свободы K** связано с числом наблюдений и числом определяемых по ним констант

# Распределение дисперсии на одну степень свободы

| Источники вариации | Суммы квадратов отклонений | Число степеней свободы | Дисперсия на одну степень свободы |
|---|---|---|---|
| Общая | $TSS = \sum_i (y_i - \bar{y})^2$ | $n-1$ | $S_T^2 = \frac{TSS}{n-1}$ |
| Регрессия | $RSS = \sum_i (\hat{y}_i - \bar{y})^2$ | $1$ | $S_R^2 = \frac{RSS}{1}$ |
| Остаточная | $ESS = \sum_i (y_i - \hat{y}_i)^2$ | $n-2$ | $S_E^2 = \frac{ESS}{n-2}$ |

Оценка **значимости уравнения** регрессии в целом делается с помощью F-критерия Фишера

$$F = \frac{S_R^2}{S_E^2} = \frac{RSS \cdot (n-2)}{ESS \cdot 1} = \frac{RSS \cdot (n-2)}{TSS - RSS} = \frac{R^2}{1-R^2}(n-2)$$

Гипотеза $H_0$ (нулевая) об отсутствии связи изучаемого показателя с фактором отклоняется и делается вывод о существенности этой связи с уровнем значимости α, если

$$F > F_{\alpha;k_1;k_2}$$

- Итак, если $F_{\text{факт(рассчет.)}} > F_{\text{табл.}}$ ,
то гипотеза $H_0$ о случайной природе оцениваемых характеристик отклоняется и признается их статистическая значимость и надежность.
- Для оценки статистической **значимости коэффициентов регрессии и коэффициента корреляции** рассчитывается t-критерий Стьюдента.

- $\boldsymbol{F_{табл}}$ – это максимально возможное значение критерия, которое могло сформироваться 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 принимается равной 0,05 или 0,01.

Имеются таблицы критических (табличных) значений F-критерия: $\boldsymbol{F(\alpha; k_1; k_2)}$, где , $k_1$=m; $k_2$=n-m-1,

где $n$ – число единиц совокупности;

$m$ – число параметров при переменных $x$.

Например, для линейного уравнения парной регрессии с уровнем значимости α = 0,05 необходимо в таблице значений (см.приложение) найти значение $\boldsymbol{F(0,05; 1; n – 2)}$.

# Регрессия с ограничениями

- Модель, в которой мы проверяем гипотезу о коэффициентах, называется регрессией без ограничений (*unrestricted, UR*)
- Регрессия с ограничениями строится из регрессии без ограничений в предположении, что нулевая гипотеза верна (*restricted*, *R*)
- Сравнение объясняющих способностей регрессии с ограничениями и регрессии без ограничений при помощи *F*-теста – очень распространенный прием в эконометрике.