Степени свободы для сумм квадратов

- Число степеней свободы некоторой величины – это разность между числом ее исследований (наблюдений) и числом констант, которые установлены (найдены) в результате этих исследований, независимо друг от друга

Применительно к *SST*, *SSR* и *SSE*

Степень свободы – *число, которое показывает, сколько независимых элементов информации, полученных из элементов* $y_1, y_2, \dots, y_n$ *, необходимо для расчета данной суммы квадратов*

SST

$$y_1 - \bar{y}\,,\quad y_2 - \bar{y}\,,\ldots,\, y_n - \bar{y}$$

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right) = 0 \qquad\qquad \sum_{i=1}^{n-1}\left(y_i - \bar{y}\right) \neq 0$$

$n-1$ разность независима

$$DF_{SST} = n-1$$

SSR

$$\hat{a}_1$$

$$\hat{y}_i - \bar{y} = \hat{a}_0 + \hat{a}_1 x_i - \hat{a}_0 - \hat{a}_1 \bar{x} = \hat{a}_1 \left(x_i - \bar{x} \right)$$

$$\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y} \right)^2 = \hat{a}_1^2 \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2$$

$$DF_{SSR} = 1$$

SSE

$$DF_{SSE} = n - 2$$

Определение

- *Средний квадрат – сумма квадратов, поделенная на соответствующее число степеней свободы*

Средний квадрат ошибок

$$MSE = \frac{\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2}{n-2}$$

Средний квадрат, объясняющий регрессию

$$MSR = \frac{\sum_{i=1}^{n}\left(\widehat{y}_i - \overline{y}\right)^2}{1} = SSR$$

- **Проверка адекватности модели и ее параметров**

- **Проверка адекватности модели по *F*-критерию Фишера**

$$F = \frac{MSR}{MSE}$$

$$v_1 = 1 \qquad\qquad v_2 = n - 2$$

$$\hat{a}_1 = 0$$

$$\widehat{y}_i = \widehat{a}_0 + \widehat{a}_1 x_i = \overline{y} - \widehat{a}_1 \overline{x} + \widehat{a}_1 x_i = \overline{y} + \widehat{a}_1 \left(x_i - \overline{x} \right) = \overline{y}$$

Гипотеза $H_0 : \widehat{a}_1 = 0$

Этапы проверки

1. Рассчитываем величину *F*-критерия

$$F_{расч.} = \frac{MSR}{MSE}$$

2. Задаем уровень значимости α

3. Находим табличное (критическое) значение F-критерия $\left(F_{(1,n-2,\alpha)}\right)$ с $v_1 = 1$ и $v_2 = n - 2$ степенями свободы и уровнем значимости α

4. Если $F_{расч.} > F_{(1,n-2,\alpha)}$, то гипотеза

$$H_0 : \widehat{a}_1 = 0$$

отвергается, если , $F_{расч.} \le F_{(1,n-2,\alpha)}$

то гипотеза принимается

Рассчитаем *F*-статистику Фишера

$$MSR = \frac{\sum_{i=1}^{n}\left(\widehat{y}_i - \overline{y}\right)^2}{1} = \frac{274,7935}{1} = 274,7935$$

$$MSE = \frac{\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2}{n-2} = \frac{10,70652}{8 - 2} = 1,78442$$

$$F_{расч.} = \frac{MSR}{MSE} = \frac{274,7935}{1,78442} = 153,996$$

Для

$$v_1 = 1 \quad v_2 = n - 2$$

$$\alpha = 0,05 \qquad F_{(1,6,0.05)} = 5,99$$

$$F_{расч.} = 153,996 > F_{(1,6,0.05)} = 5,99$$

Проверка гипотезы о значимости коэффициента корреляции

Гипотеза

$$H_0 : r_{yx} = 0$$

Используем критерий Стьюдента

$$t_{расч.} = r_{yx} \cdot \sqrt{\frac{n-2}{1-r_{yx}^2}}$$

или

$$t_{расч.} = r_{yx} \cdot \sqrt{\frac{n-2}{1-R^2}}$$

$v = n - 2$ - число степеней свободы

Для заданных v и α находим табличное значение $t_{(v,\alpha)}$

Если $\left|t_{расч.}\right| > t_{(v,\alpha)}$, то гипотеза $H_0 : r_{yx} = 0$ отвергается, если $\left|t_{расч.}\right| \le t_{(v,\alpha)}$ - то принимается

Рассчитаем t-статистику Стьюдента

$$t_{расч.} = r_{yx} \cdot \sqrt{\frac{n-2}{1-r_{yx}^2}} = 0{,}981 \cdot \sqrt{\frac{8-2}{1-0{,}9625}} = 12{,}41$$

При $\alpha=0,05$ и $v=n-2=8-2=6$

$$t_{табл.} = t_{(6,0.05)} = 1,943$$

Так как

$$t_{расч.} = 12,41 > t_{(6,0.05)} = 1,943,$$

то коэффициент корреляции значим

•<u>Критерии оценки качества линейной модели</u>

1. Средняя ошибка прогноза

$$ME = \overline{u} = \frac{1}{n}\sum_{i=1}^{n} u_i$$

$ME \to 0$ при $n \to \infty$

2. Дисперсия и стандартное отклонение ошибок

$$\operatorname{var}(u) = \frac{1}{n}\sum_{i=1}^{n}\left(u_i - \overline{u}\right)^2 = \sigma_u^2$$

$$\sigma = \sqrt{\operatorname{var}(u)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(u_i - \overline{u}\right)^2}$$

3. Среднее абсолютное отклонение

$$MAD(u) = \frac{1}{n}\sum_{i=1}^{n}\left|u_i - \overline{u}\right|$$

4. Средний квадрат ошибки

$$MSE(u) = \frac{1}{n}\sum_{i=1}^{n} u_i^2$$

(сумма квадратов ошибок)

$$SSE = \sum_{i=1}^{n} u_i^2$$

5. Средняя абсолютная процентная ошибка

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|u_i|}{y_i}100\%$$

6. Средняя процентная ошибка аппроксимации

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\frac{u_i}{y_i}100\%$$

7. Средняя абсолютная ошибка

$$MAE(u) = \frac{1}{n}\sum_{i=1}^{n}|u_i|$$

Рассчитаем среднюю процентную ошибку аппроксимации

$$MPE = \frac{1}{n}\sum_{i=1}^{n}\frac{u_i}{y_i}\cdot 100\% =$$

$$= \frac{-0{,}28386}{8}\cdot 100\% = -3{,}55\%$$

Несмещенная оценка истинного значения σ_u^2:

а) для парной зависимости

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{n} u_i^2}{n-2}$$

б) для многофакторной модели

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{n} u_i^2}{n-k}$$

$n-2$ и $n-k$ – число степеней свободы

- **Условия применения метода наименьших квадратов**

1. $$M\left(u_i\right)=0$$

Факторы, не включенные в модель, не влияют систематически на математическое ожидание

2. Гомоскедастичность

$$\operatorname{var}(u_i) = \sigma_u^2, \; i = \overline{1, n}$$

3. Случайная величина u распределена по нормальному закону

$$u \sim N\left(0, \sigma_u^2\right)$$

4. Между случайными величинами u отсутствует автокорреляция

$$\operatorname{cov}(u_i, u_j) = M\{[u_i - M(u_i)] \cdot [u_j - M(u_j)]\} =$$

$$= M(u_i u_j) = 0, \; i \neq j$$

5. Значения независимых переменных x_i не зависят от значений случайных величин u_i

$$\operatorname{cov}(u_i, x_i) = M\{[u_i - M(u_i)] \cdot [x_i - M(x_i)]\} =$$

$$= M\{u_i [x_i - M(x_i)]\} =$$

$$= M(u_i x_i) - M(u_i) \cdot M(x_i) = M(u_i x_i) = 0$$

$$M(u_i) = 0 \qquad\qquad M(x_i) = const$$

6. Независимые переменные не должны быть мультиколлинеарными

7. Регрессионная модель построена (специфицирована) правильно: правильно подобрана линия регрессии, в нее включены только существенно влияющие переменные, модель адекватна экономическому процессу

8. Количество наблюдений больше числа независимых переменных (регрессоров) $(n > k)$

9. Регрессоры – детерминированные величины

10. Набор переменных X не содержит ошибок наблюдений

11. Набор переменных X содержит все существенно влияющие на регрессант факторы

12. Коэффициенты регрессии – детерминированные величины, неизменные во времени и пространстве

- <u>**Распределение независимой переменной**</u> y

Единственный источник изменения ***y*** –

случайная величина ***u***

распределение переменной ***y*** зависит от

распределения случайной величины ***u***, т.е.

<u>является нормальным</u>

$$M(y_i) = M(a_0 + a_1 x_i + u_i) = M(a_0 + a_1 x_i) + M(u_i)$$

$$M(u_i) = 0$$

a_0, a_1 - constants

x_i – детерминированная величина

$$M\left(y_i\right) = M\left(a_0 + a_1 x_i\right) = a_0 + a_1 x_i$$

$$D(y_i) = M\left[y_i - M\left(y_i\right)\right]^2 = \sigma_u^2$$

Здесь и далее в расчетах используется

несмещенная оценка дисперсии !

$$\widehat{\sigma}_u^2$$