

Тема 3.

Множественная регрессия и корреляция

План

3.1. Основные гипотезы

3.2. Парная и частная корреляция

3.3. Множественный коэффициент
корреляции и множественный
коэффициент детерминации

3.4. Проверка гипотез

Модель множественной линейной регрессии

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{k-1} X_{k-1} + \beta_k X_k + \varepsilon$$

Уравнение множественной линейной регрессии

$$\hat{y}_x = b_0 + b_1 x_1 + b_2 x_2 + \\ + \square + b_{k-1} x_{k-1} + b_k x_k$$

Парные коэффициенты корреляции

$$r_{yx_1}, \quad r_{yx_2}, \quad \dots \quad r_{yx_k}$$

Частные коэффициенты корреляции

$$r_{yx_1(x_2)} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{(1 - r_{yx_2}^2) \cdot (1 - r_{x_1x_2}^2)}}$$



$$r_{yx_2}(x_1) = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{(1 - r_{yx_1}^2) \cdot (1 - r_{x_1x_2}^2)}}$$

$$r_{x_1x_2}(y) = \frac{r_{x_1x_2} - r_{yx_1} \cdot r_{yx_2}}{\sqrt{(1 - r_{yx_1}^2) \cdot (1 - r_{yx_2}^2)}}$$

Парные коэффициенты корреляции

$$r_{yx_1} = 0,75$$

$$r_{yx_2} = 0,8$$

$$r_{x_1x_2} = 0,9$$

Частные коэффициенты корреляции

$$r_{yx_1(x_2)} = \frac{0,75 - 0,8 \cdot 0,9}{\sqrt{(1 - 0,8^2) \cdot (1 - 0,9^2)}} = 0,115$$

$$r_{yx_2(x_1)} = \frac{0,8 - 0,75 \cdot 0,9}{\sqrt{(1 - 0,75^2) \cdot (1 - 0,9^2)}} = 0,434$$

Частные коэффициенты корреляции

$$r_{yx_1}(x_2)$$

1-го порядка

$$r_{yx_1}(x_2, x_3)$$

2-го порядка

$$r_{yx_1}(x_2, x_3, x_4)$$

3-го порядка

$$r_{yx_1}(x_2, x_3, \dots, x_{k+1})$$

k-го порядка

$$r_{yx_j(x_1x_2\ldots x_k)} =$$

$$= \frac{r_{yx_i(x_1x_2\ldots x_{k-1})} - r_{yx_k(x_1x_2\ldots x_{k-1})} r_{x_ix_k(x_1x_2\ldots x_{k-1})}}{\sqrt{\left(1 - r_{yx_k(x_1x_2\ldots x_{k-1})}^2\right) \cdot \left(1 - r_{x_ix_k(x_1x_2\ldots x_{k-1})}^2\right)}}$$

Проверка статистической значимости частного коэффициента корреляции

Сформулируем гипотезы

$$H_0 : \rho_{yx_i(x_1x_2 \dots x_k)} = 0$$

$$H_1 : \rho_{yx_i(x_1x_2 \dots x_k)} \neq 0$$

Устанавливаем уровень
значимости α

$$\alpha = 0,05$$

Находим наблюдаемое значение
критерия

$$t_{\text{набл.}} = \frac{r_{yx_i(x_1 x_2 \dots x_k)}}{\sigma_r}$$

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

Находим критическое значение критерия по таблице Стьюдента по уровню значимости α и по числу степеней свободы $k=n-2-1$

$$t_{\text{кр.}}(\alpha; k=n-2-1)$$

Если $|t_{набл}| > t_{кр}$, то нулевая гипотеза отклоняется в пользу альтернативной о статистической значимости частного коэффициента корреляции. Иначе, оснований отклонять нулевую гипотезу нет.

Множественный коэффициент корреляции

$$R_{yx_1x_2\dots x_k} = \sqrt{\frac{SSR}{SST}} = \sqrt{R^2}$$

$$0 \leq R_{yx_1x_2\dots x_k} \leq 1$$

Множественный коэффициент детерминации

$$R^2 = \frac{SSR}{SST}$$

$$0 \leq R^2 \leq 1$$

Скорректированный коэффициент детерминации

$$R^2_{\text{сkopp}} = 1 - (1 - R^2) \cdot \frac{n - 1}{n - m}$$

$$m = k + 1$$

Модель множественной линейной регрессии

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{k-1} X_{k-1} + \beta_k X_k + \varepsilon$$

Уравнение множественной линейной регрессии

$$\hat{y}_x = b_0 + b_1 x_1 + b_2 x_2 + \\ + \square + b_{k-1} x_{k-1} + b_k x_k$$

Интерпретация коэффициентов регрессии b_j

b_j показывает на сколько единиц изменится Y , если X_j изменится на единицу, при условии, что все остальные объясняющие переменные не изменятся.

Интерпретация свободного члена b_0

b_0 определяет значение Y при условии, что все объясняющие переменные равны нулю; оценивает агрегированное влияние всех неучтенных в модели факторов.

Метод наименьших квадратов

$$\sum_{i=1}^n (y - b_0 - b_1 x_1 - b_2 x_2 - \dots - b_k x_k)^2 \rightarrow \min$$

- Находим частные производные первого порядка и приравниваем их к нулю:

$$\begin{cases} \frac{dSSE}{db_0} = 0 \\ \frac{dSSE}{db_1} = 0 \\ \vdots \\ \frac{dSSE}{db_k} = 0 \end{cases}$$

Двухфакторная модель множественной линейной регрессии

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Уравнение регрессии

$$\hat{y}_x = b_0 + b_1x_1 + b_2x_2$$

Метод наименьших квадратов

$$\sum_{i=1}^n (y - b_0 - b_1 x_1 - b_2 x_2)^2 \rightarrow \min$$

- Находим частные производные первого порядка и приравниваем их к нулю:

$$\begin{cases} \frac{dSSE}{db_0} = 0 \\ \frac{dSSE}{db_1} = 0 \\ \frac{dSSE}{db_2} = 0 \end{cases}$$

- Оценки параметров двухфакторной модели

$$\begin{cases} b_1 = \frac{\text{cov}(y, x_1)\sigma^2(x_2) - \text{cov}(y, x_2)\text{cov}(x_1, x_2)}{\sigma^2(x_1)\sigma^2(x_2) - [\text{cov}(x_1, x_2)]^2} \\ b_2 = \frac{\text{cov}(y, x_2)\sigma^2(x_1) - \text{cov}(y, x_1)\text{cov}(x_1, x_2)}{\sigma^2(x_1)\sigma^2(x_2) - [\text{cov}(x_1, x_2)]^2} \\ b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 \end{cases}$$

Проверка статистической значимости уравнения регрессии в целом.

Сформулируем гипотезы:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

Y не зависит от всех X,
включенных в модель
(уравнение в целом не
значимо)

Проверка статистической значимости уравнения регрессии в целом.

Сформулируем гипотезы:

$$H_0 : R^2 = 0$$

Y не зависит от всех
X, включенных
в модель (уравнение в
целом не значимо)

$$H_1 : R^2 > 0$$

Y зависит от всех X
(вместе взятых),
включенных в модель
(уравнение в целом
значимо)



Устанавливаем
уровень значимости α

$$\alpha = 0,05$$

Найдем
наблюдаемое значение критерия

$$F_{\text{набл.}} = \frac{SSR / (m - 1)}{SSE / (n - m)}$$

где n — число наблюдений,
 m — число параметров в модели
регрессии

По таблице распределения Фишера
найдем критическое значение критерия:

$$F_{кр} = F(\alpha; k_1 = m - 1; k_2 = n - m)$$

Если $F_{набл.} > F_{кр.}$, то нулевая гипотеза отклоняется в пользу альтернативной о статистической значимости уравнения регрессии в целом. Иначе оснований отклонять нулевую гипотезу нет.

Проверка статистической значимости коэффициентов регрессии

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

Для проверки гипотезы
используется t-критерий

$$t_{набл\ b_j} = \frac{b_j}{S_{b_j}}$$

Стандартная ошибка коэффициентов регрессии

$$S_{bj} = \frac{\sigma_y \sqrt{1 - R_{yx_1x_2 \dots x_k}^2}}{\sigma_{xj} \sqrt{1 - R_{x_jx_1x_2 \dots x_k}^2}} \cdot \frac{1}{n - m - 1}$$

Стандартные ошибки коэффициентов регрессии для двухфакторной модели

$$S_{b_1} = \frac{S_{yx}}{\sigma_{x_1} \sqrt{n}} \cdot \sqrt{\frac{1}{1 - r_{x_1 x_2}^2}}$$

$$S_{b_2} = \frac{S_{yx}}{\sigma_{x_2} \sqrt{n}} \cdot \sqrt{\frac{1}{1 - r_{x_1 x_2}^2}}$$

$t_{кр}$ определяется по таблице
распределения Стьюдента
по уровню значимости α и
числу степеней свободы
 $k=n-m$

$$t_{кр.}(\alpha; k=n-m)$$

Если $|t_{набл}| > t_{кр}$,то нулевая гипотеза отклоняется в пользу альтернативной о статистической значимости коэффициента регрессии. Иначе, оснований отклонять нулевую гипотезу нет.

Уравнение регрессии

$$\hat{y}_x = 116,7 + 0,112x_1 - 0,739x_2$$

(9,6) (0,003) (0,114)

$$R^2 = 0,99$$

Коэффициент эластичности

- $$\bar{\varepsilon}_j = b_j \frac{\bar{x}_j}{\bar{y}}$$

Коэффициент эластичности

Коэффициент эластичности (средний) показывает на сколько процентов в среднем изменится Y от своей средней величины при изменении значения X_j на 1% от своего среднего значения при условии, что остальные объясняющие переменные не изменятся.