Множественный регрессионный анализ

Множественная регрессия – это уравнение связи с несколькими независимыми переменными:

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

часть значения y, которая объяснена уравнением регрессии с несколькими факторами

необъясненная часть значения y (или возмущение)

Построение уравнения регрессии
1. Постановка задачи

По имеющимся данным n наблюдений за совместным изменением параметров y, xj и ((yi,xj,i); j=1, 2, …, p; i=1, 2, …, n) необходимо определить аналитическую зависимость $\hat{y} = f(x1,x2,\ldots,xp)$, наилучшим образом описывающую данные наблюдений.

Данные наблюдений

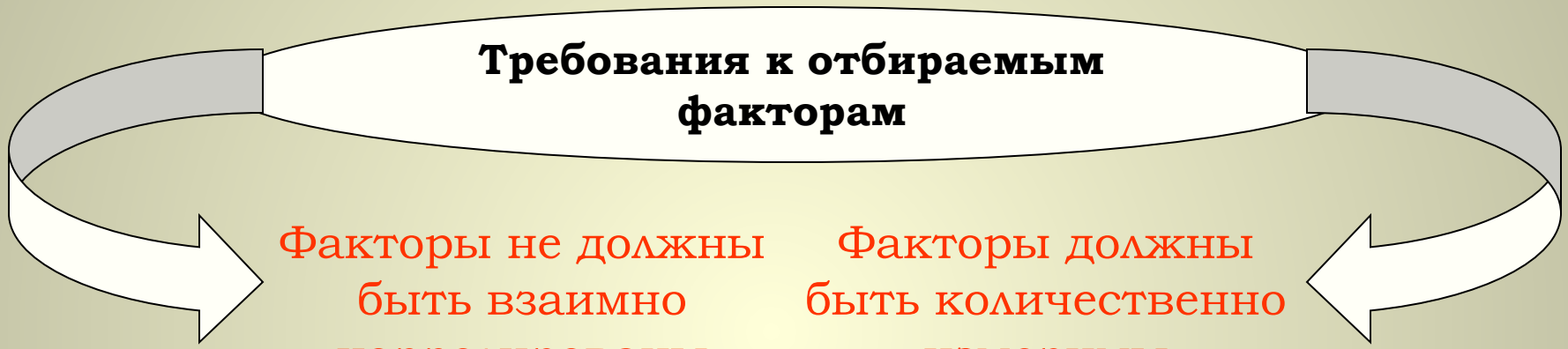
	y	x_1	x_2	…	x_p
1	y_1	x_{11}	x_{21}	…	x_{p1}
2	y_2	x_{12}	x_{22}	…	x_{p2}
…	…	…	…	…	…
n	y_n	x_{1n}	x_{2n}	…	x_{pn}

Критерий качества выбранной зависимости:

$$S = \sum (y_i - \hat{y}_i)^2 \rightarrow \min$$

2. Спецификация модели

2.1. Отбор факторов, подлежащих включению в модель



Пример:

y – себестоимость единицы продукции

x – заработная плата работника

z – производительность труда

$$y = 22600 - 5x - 10z + \varepsilon$$

$$r_{xz} = 0{,}95$$

- ✔ *целесообразность включения каждого нового фактора оценивается с помощью коэффициента детерминации;*
- ✔ *при возникновении необходимости добавить в уравнение качественный фактор вводится «фиктивная» переменная*

Парная коллинеарность и мультиколлинеарность

Две переменные считаются ***явно коллинеарными***, т.е. находятся между собой в линейной зависимости, если ***коэффициент интеркорреляции*** (корреляции между двумя объясняющими переменными) ≥ 0,7.

Если факторы явно коллинеарны, то они дублируют друг друга и один из них рекомендуется исключить из уравнения.

Мультиколлинеарность – *линейная зависимость между более чем двумя переменными, т.е. совокупное воздействие факторов друг на друга.*

Включение в модель мультиколлинеарных факторов нежелательно

по следующим причинам:

- затрудняется интерпретация параметров множественной регрессии; параметры линейной регрессии теряют экономический смысл;
- оценки параметров не надежны, имеют большие стандартные ошибки и меняются с изменением количества наблюдений (не только по величине, но и по знаку), что делает модель непригодной для анализа и прогнозирования.

Оценка мультиколлинеарности

Для оценки мультиколлинеарности используется ***определитель матрицы парных коэффициентов интеркорреляции***:

(!) *Если факторы не коррелируют между собой*, то матрица коэффициентов интеркорреляции является единичной, поскольку в этом случае все недиагональные элементы равны 0.

Например, для уравнения с тремя переменными

$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \varepsilon$$

$$Det(R) = |R| = \begin{vmatrix} r_{x_1 x_1} & r_{x_2 x_1} & r_{x_3 x_1} \\ r_{x_1 x_2} & r_{x_2 x_2} & r_{x_3 x_2} \\ r_{x_1 x_3} & r_{x_2 x_3} & r_{x_3 x_3} \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = 1$$

(!) *Если между факторами существует полная линейная зависимость* и все коэффициенты корреляции равны 1, то определитель такой матрицы равен 0.

$$Det(R) = |R| = \begin{vmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{vmatrix} = 0$$

Чем ближе к 0 определитель матрицы коэффициентов интеркорреляции, тем сильнее мультиколлинеарность и ненадежнее результаты множественной регрессии.

Чем ближе к 1 определитель матрицы коэффициентов интеркорреляции, тем меньше мультиколлинеарность факторов.

Способы преодоления мультиколлинеарности факторов:

- исключение из модели одного или нескольких факторов;
- переход к совмещенным уравнениям регрессии, т.е. к уравнениям, которые отражают не только влияние факторов, но и их взаимодействие. *Например*,

  если $y = f(x_1, x_2, x_3)$, то можно построить следующее совмещенное уравнение:

  $$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_{12} x_1 x_2 + b_{13} x_1 x_3 + b_{23} x_2 x_3 + \varepsilon$$

- переход к уравнениям приведенной формы (в уравнение регрессии подставляется рассматриваемый фактор, выраженный из другого уравнения).

2. Спецификация модели

2.2. Выбор формы уравнения регрессии

- **Линейная регрессия**

$$y = a + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p + \varepsilon$$

- **Линеаризуемые регрессии**
  - **Степенная регрессия**

$$y = a x_1^{b_1} x_2^{b_2} \ldots x_p^{b_p} \varepsilon$$

  - **Экспоненциальная регрессия**

$$y = e^{a + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p + \varepsilon}$$

  - **Гиперболическая регрессия**

$$y = \frac{1}{a + b_1 x_1 + b_2 x_2 + \ldots b_p x_p + \varepsilon}$$

Например, зависимость спроса на товар (Qd) от цены (P) и дохода (I) характеризуется следующим уравнением:

Qd = 2,5 - 0,12P + 0,23 I.

Коэффициенты данного уравнения говорят о том, что при увеличении цены на единицу, спрос уменьшится в среднем на 0,12 единиц, а при увеличении дохода на единицу, спрос возрастет в среднем 0,23 единицы.

Например, зависимость выпуска продукции Y от затрат капитала K и труда L:

$$Y = 0.89K^{0.23}L^{0.81}$$

говорит о том, что увеличение затрат капитала K на 1% при неизменных затратах труда вызывает увеличение выпуска продукции Y на 0,23%. Увеличение затрат труда L на 1% при неизменных затратах капитала K вызывает увеличение выпуска продукции Y на 0,81%.

3. Оценка параметров модели

3.1. МНК

$$S=\sum(y_i-\hat{y}_i)^2\to\min \quad \text{или} \quad \sum\varepsilon^2\to\min$$

$$S=\sum(y_i-a-b_1x_1-b_2x_2-\ldots-b_px_p)^2=S(a,b_1,\ldots,b_p)$$

$$S'_a=-2\sum(y_i-a-b_1x_1-b_2x_2-\ldots-b_px_p)=0$$

$$S'_{b_1}=-2b_1\sum(y_i-a-b_1x_1-b_2x_2-\ldots-b_px_p)=0$$

$$\ldots$$

$$S'_{b_p}=-2b_p\sum(y_i-a-b_1x_1-b_2x_2-\ldots-b_px_p)=0$$

Отсюда получаем систему уравнений:

$$\begin{cases}\sum y=na+b_1\sum x_1+b_2\sum x_2+\ldots+b_p\sum x_p,\\ \sum yx_1=a\sum x_1+b_1\sum x_1^2+b_2\sum x_1x_2+\ldots+b_p\sum x_px_1,\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ \sum yx_p=a\sum x_p+b_1\sum x_1x_p+b_2\sum x_2x_p+\ldots+b_p\sum x_p^2.\end{cases}$$

Решение системы уравнений с помощью метода определителей:

$$a = \frac{\Delta a}{\Delta}, \quad b_1 = \frac{\Delta b_1}{\Delta}, \quad \ldots, \quad b_p = \frac{\Delta b_p}{\Delta}$$

где Δ – определитель системы:

$$\Delta = \begin{vmatrix} n & \sum x_1 & \sum x_2 & \ldots & \sum x_p \\ \sum x_1 & \sum x_1^2 & \sum x_2 x_1 & \ldots & \sum x_p x_1 \\ \sum x_2 & \sum x_1 x_2 & \sum x_2^2 & \ldots & \sum x_p x_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \sum x_p & \sum x_1 x_p & \sum x_2 x_p & \ldots & \sum x_p^2 \end{vmatrix}$$

Δa, Δb_1, Δb_p – частные определители (Δj) , которые получаются из основного определителя путем замены j-го столбца на столбец свободных членов

$$\begin{pmatrix} \sum y \\ \sum yx_1 \\ \ldots \\ \sum yx_p \end{pmatrix}$$

3. Оценка параметров модели

3.2. Метод оценки параметров через стандартизованные коэффициенты β

Уравнение регрессии в стандартизованном (нормированном) масштабе:

$$t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \ldots + \beta_p t_{x_p}$$

где $t_y = \frac{y - \bar{y}}{\sigma_y}$, $t_{x_i} = \frac{x_i - \bar{x}_i}{\sigma_{x_i}}$ - стандартизованные переменные

β - стандартизованные коэффициенты регрессии.

β-коэффициенты показывают, на сколько сигм (средних квадратических отклонений) изменится в среднем результат за счет изменения соответствующего фактора xi на одну сигму при неизменном среднем уровне других факторов.

Взаимосвязь b_i и β

Связь коэффициентов «чистой» регрессии *bi* с коэффициентами *βi* описывается соотношением:

$$b_i = \beta_i \frac{\sigma_y}{\sigma_{x_i}} \quad \text{или} \quad \beta_i = b_i \frac{\sigma_{x_i}}{\sigma_y} \ (i = 1,2,\ldots, p)$$

Коэффициенты β определяются при помощи МНК из следующей системы уравнений методом определителей:

$$\begin{cases} r_{yx_1} = \beta_1 + \beta_2 r_{x_2x_1} + \beta_3 r_{x_3x_1} + \ldots + \beta_p r_{x_px_1}, \\ r_{yx_2} = \beta_1 r_{x_2x_1} + \beta_2 + \beta_3 r_{x_3x_2} + \ldots + \beta_p r_{x_px_2}, \\ \\ r_{yx_p} = \beta_1 r_{x_px_1} + \beta_2 r_{x_px_2} + \beta_3 r_{x_3x_p} + \ldots + \beta_p \end{cases}$$

Параметр a определяется как:

$$a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \ldots - b_p\bar{x}_p$$

4. Проверка качества уравнения регрессии

H0: уравнение статистически не значимо

$$yi = \hat{y}i + \varepsilon i$$

$$D(y) = D(\hat{y}) + D(\varepsilon)$$

$$\frac{1}{n}\sum(y-\bar{y})^2 \quad \frac{1}{n}\sum(\hat{y}-\bar{y})^2 \quad \frac{1}{n}\sum(y-\hat{y})^2$$

полная (общая) сумма квадратов отклонений = **сумма квадратов отклонений, объясненная регрессией** + **(остаточная) сумма квадратов отклонений, не объясненная регрессией**

F-критерий Фишера:

$$F = \frac{\frac{D(\hat{y})}{k}}{\frac{D(\varepsilon)}{n-m-1}} \quad \textit{или} \quad \frac{R^2}{1-R^2} \times \frac{n-m-1}{m}$$

где m – число независимых переменных в уравнении регрессии;

n – число единиц совокупности.

Если **Fфакт > Fтабл**, то *H0* о случайной природе связи отклоняется и признается статистическая значимость и надежность уравнения.

Если **Fфакт < Fтабл**, то *H0* не отклоняется и признается статистическая незначимость уравнения регрессии.

Частный F-критерий:

$$F_{частx_i} = \frac{R^2_{yx_1 \ldots x_i \ldots x_p} - R^2_{yx_1 \ldots x_i - 1 x_i + 1 \ldots x_p}}{1 - R^2_{yx_1 \ldots x_i \ldots x_p}} \cdot \frac{n - m - 1}{1}$$

- оценивает статистическую значимость присутствия каждого из факторов в уравнении.