Тема 2. Множественная линейная регрессия

Модель множественной линейной регрессии:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon$$

Уравнение множественной линейной регрессии со свободным членом и ***k*** независимыми переменными (факторами):

$$\hat{y} = b_0 + b_1 x_1 + \dots + b_k x_k$$

МНК и основные гипотезы

Применение МНК даёт систему ***k+1*** линейных алгебраических уравнений с ***k+1*** неизвестными (системы нормальных уравнений):

$$\boxed{S^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \to \min}$$

$$X^T X b = X^T y \quad ,$$

откуда:

$$b = (X^T X)^{-1} X^T y = \begin{bmatrix} b_0 \\ \vdots \\ b_k \end{bmatrix}$$

Гипотезы гомоскедастичности и независимости:

$$V(\varepsilon) = E(\varepsilon \varepsilon^T) = \sigma^2 I_n$$

Оценка дисперсии ошибок σ^2

Несмещённая оценка σ^2 равна:

$$\boxed{\hat{\sigma}^2 = s^2 = \frac{1}{n-k-1}\sum \varepsilon_i{}^2 = \frac{1}{n-k-1}\sum (y_i - \hat{y}_i)^2}$$

Числа степеней свободы (df)

Пусть ***n*** – число наблюдений, ***k*** – число факторов. Разность $\boxed{n-k-1>0}$ называется **числом степеней свободы**

(*разность между числом наблюдений и числом оцененных параметров*).

Для надёжной оценки формулы связи требуется: (*как минимум*) $\boxed{n \geq 3(k+1)}$

Если $n = k + 1$, то коэффициенты регрессии оцениваются единственным образом.

Если $n > k + 1$, то нельзя найти ***точную*** формулу связи, а необходимо выбрать наилучшее приближение для имеющихся наблюдений – ***устойчивую*** формулу связи.

Коэффициент детерминации

Для модели регрессии со свободным членом справедливо соотношение:

$$S^2_{общ.} = S^2_{регр.} + S^2_{ост.}$$

или

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

откуда

$$R^2 = \frac{S^2_{регр.}}{S^2_{общ.}} = 1 - \frac{S^2_{ост.}}{S^2_{общ.}}$$

Свойства коэффициента детерминации:

1. При добавлении фактора (регрессора) в модель величина R^2 не убывает.
2. При преобразовании зависимой переменной R^2 изменяется.

Для устранения эффекта возрастания R^2 при увеличении числа регрессоров используют **скорректированный** (**adjusted**) $R^2_{adj}\,(\overline{R^2})$

$$R^2_{adj} = 1 - \frac{S^2_{ост.}/(n-m-1)}{S^2_{общ.}/(n-1)}$$

$$R^2 \geq R^2_{adj}$$

Индекс корреляции *R*

R характеризует тесноту связи между набором всех факторов x_j и результативным признаком y:

$$R = \sqrt{1 - \frac{S^2_{ост.}}{S^2_{общ.}}} \qquad 0 < R < 1$$

Данная формула не зависит от вида уравнения и от факторов x_j .

Особенности спецификации множественной регрессии

- Отбор факторов
- Выбор вида уравнения

Отбор – I стадия: на основе качественного теоретико-экономического анализа, исходя из природы взаимосвязи изучаемых явлений.

Отбор – II стадия: анализ взаимосвязи всех признаков и целесообразности их включения в модель.

Условие качественной регрессии: независимость факторов между собой (анализируется матрица попарных коэффициентов корреляции $r_{x_i x_j} = r_{ij}$)

Отбор факторов. Коллинеарность и мультиколлинеарность

- **Коллинеарность** – линейная взаимосвязь двух регрессоров (выявляется с помощью матрицы парных корреляций: $|r_{ij}| > 0,7$)
- **Мультиколлинеарность** – линейная связь (корреляция) более 2х регрессоров (определяется с помощью матрицы межфакторной корреляции: $R_x = \|r_{ij}\|$;

$\boxed{|R_x| \to 0}$ – критерий наличия мультиколлинеарности: чем ближе $|R_x|$ к нулю, тем сильнее мультиколлинеарность.

Матрица межфакторной корреляции $R_x = \left\| r_{ij} \right\|$

$$R_x = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{pmatrix}$$

Последствия мультиколлинеарности

При наличии мультиколлинеар-ности матрица $X^T X$ является **вырожденной** (обратная матрица не существует) $\Rightarrow$

- МНК-оценки имеют большую вариацию и являются ненадёжными
- Интерпретация параметров затрудняет-ся, они теряют экономический смысл

Внешние признаки наличия мультиколлинеарности

- Некоторые из МНК-оценок имеют непра-вильные (с точки зрения экономической теории) значения или знаки
- Небольшое изменение исходных данных приводит к существенному изменению оценок
- Большинство оценок параметров являются статистически незначимыми, а модель в целом – значимой

Методы устранения мультиколлинеарности

1. Удаление из модели факторов, ответственных за мультиколлинеарность (задача их выявления)
2. Преобразование факторов, уменьшающее корреляцию между ними
3. Построение совмещённого уравнения регрессии, например:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2$$

Выявление факторов, ответст-венных за мультиколлинеарность

- Экпериментальные методы отбора (перебора) факторов ($k < n$ в 6-7 раз)
- Использование индексов детерминации

$$R^2_{x_1|x_2,x_3,\dots,x_k}; \quad R^2_{x_2|x_1,x_3,\dots,x_k}; \dots$$

(переменные, ответственные за мультиколлинеарность, дают значения R^2, близкие к 1)

Отбор факторов с помощью частных корреляций

Парные коэффициенты корреляции могут давать завышенные оценки связи из-за взаимосвязи факторов.

Частные корреляции элиминируют влияние других факторов, т.е. оценивают парные связи в «чистом» виде:

$$r_{yx_1|x_2 \ldots x_k}$$

- коэффициент (*k-1*)-го порядка

Так как при включении в уравнение связи нового фактора величина R^2 увеличивается, то следовательно величина остаточной дисперсии будет уменьшаться.

Показатель частной корреляции выражается отношением уменьшения остаточной дисперсии к её величине, рассчитанной до этого.

Если $y = f(x_1, x_2)$, то в частности:

$$r_{yx_2|x_1} = \sqrt{\frac{S^2_{yx_1} - S^2_{yx_1x_2}}{S^2_{yx_1}}} = \sqrt{1 - \frac{1 - R^2_{yx_1x_2}}{1 - r^2_{yx_1}}}$$

Коэффициенты частной корреляции различных порядков связаны рекуррентным соотношением:

$$r_{yx_1|x_2 \ldots x_k} =$$

$$= \frac{r_{yx_1|x_2 \ldots x_{k-1}} - r_{yx_k|x_2 \ldots x_{k-1}} r_{x_1 x_k|x_2 \ldots x_{k-1}}}{\sqrt{\left(1 - r^2_{yx_k|x_2 \ldots x_{k-1}}\right)\left(1 - r^2_{x_1 x_k|x_2 \ldots x_{k-1}}\right)}}$$

В частности:

- $y = f(x_1, x_2):$

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

- $y = f(x_1, x_2, x_3):$

$$r_{yx_1|x_2x_3} = \frac{r_{yx_1|x_2} - r_{yx_3|x_2} r_{x_1x_3|x_2}}{\sqrt{\left(1 - r_{yx_3|x_2}^2\right)\left(1 - r_{x_1x_3|x_2}^2\right)}}$$

Фиктивные переменные

используются, когда в модель необходимо включить **качественные** признаки, оценить их влияние на ***y***, исследовать структурные изменения и т. п.

Если качественный признак ***z*** имеет **два** значения, то их обозначают числами

0 и 1 (бинарная переменная).

Если качественный признак имеет **несколько** значений (L градаций), то для его описания используют несколько бинарных переменных ($L-1$).

Пример:

- Модель 1: $y = x_1\beta_1 + ... + x_k\beta_k + \varepsilon$
- Модель 2: $y = x_1\beta_1 + ... + x_k\beta_k + z \cdot \beta_{k+1} + \varepsilon$

где y - з/плата, $x_1,...,x_k$ - количественные объясняющие переменные.

$$z = \begin{cases} 1 - \textit{работник имеет в/о} \\ 0 - \textit{работник не имеет в/о} \end{cases}$$

Проверяя гипотезу $H_0 : \beta_{k+1} = 0$, можно ответить на вопрос: влияет ли наличие высшего образования на размер з/платы.

Интерпретация результатов регрессии с фиктивными переменными

Коэффициент регрессии (в линейной модели) отражает величину эффекта (прироста) соответствующей градации качественного фактора.

Фиктивная переменная может выступать в роли результативного признака ***y***. При этом (в вероятностной модели) значение признака интерпретируется как доля (вероятность) осуществления соответствующей альтернативы.

Уравнение регрессии в стандартизированной форме. β - коэффициенты

Пусть $\hat{y} = b_0 + b_1 x$. Применяя к исходным данным ***y***, ***x***, нормирующее преобразование (центрирование и нормирование):

$$t_y = \frac{y - \bar{y}}{\sigma_y}; \quad t_x = \frac{x - \bar{x}}{\sigma_x}$$

получим уравнение:

$$\hat{t}_y = \beta\, t_x$$, где $$\beta = r_{yx}$$

Аналогично строится множественное уравнение с бета-коэффициентами:

$$\hat{t}_y = \beta_1 t_{x1} + \beta_2 t_{x2} + \dots + \beta_k t_{xk}$$

Связь между бета-коэффициентами и коэффициентами «чистой» регрессии:

$$\boxed{\beta_j = b_j \cdot \frac{\sigma_j}{\sigma_y}} \qquad \boxed{b_j = \beta_j \cdot \frac{\sigma_y}{\sigma_j}}$$

позволяет перейти от одной формы к другой. При этом $b_0 = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 - \dots - b_k \bar{x}_k$.

β_j – сравнимы между собой,

b_j - не сравнимы.

Связь индекса детерминации с бета-коэффициентами

$$R^2 = \sum_{j=1}^{k} r_{yx_j} \beta_j = \sum_{j=1}^{k} R_j^2$$

R_j^2 – частный индекс детерминации. Он характеризует вклад каждого фактора x_j в общий индекс детерминации.

(справедливо для линейной регрессии)

Анализ качества регрессионной модели

- Содержательная часть
- Статистическая часть

Проверка статистического качества уравнения регрессии:

1) проверка статистической значимости каждого коэффициента регрессии (t-критерий)
2) проверка значимости регрессии в целом (F-критерий)
3) проверка выполнения основных гипотез (предпосылок МНК)

Содержательная проверка качества модели

- Интерпретация коэффициентов регрессии:

    коэффициент регрессии b_j показывает, на сколько единиц изменяется в среднем y при изменении x_j на 1 единицу (при неизменности остальных факторов).

- Сравнение факторов между собой

    с помощью коэффициентов эластичности E_j и бета-коэффициентов β_j:

$$E_j = b_j \cdot \frac{\bar{x}_j}{\bar{y}}$$

$$\beta_j = b_j \cdot \frac{\sigma_j}{\sigma_y}$$

- Прогнозирование по уравнению регрессии

Точечный и интервальный прогнозы по уравнению регрессии

Точечный прогноз $\hat{y}_p$ определяется подстановкой значений вектора $x_p = (x_{1p}, \ldots, x_{kp})$ в уравнение.

Интервальный прогноз:

$$\hat{y}_p - t_\alpha s_{y_p} < y_p < \hat{y}_p + t_\alpha s_{y_p}$$

$$s_y = s\sqrt{1 + \frac{1}{n} + \sum_{i,j=1}^{k} (x_i - \bar{x}_i) C_{ij} (x_j - \bar{x}_j)}$$

Проверка статистической значимости

1) Проверка гипотезы $H_0 : \beta_j = b_j^0 \ (или \ \beta_j = 0)$

Гипотеза отвергается, если $\left| \dfrac{b_j - b_j^0}{s_{bj}} \right| > t_\alpha (n-k-1)$

Доверительный интервал:

$$b_j - t_\alpha \cdot s_{bj} < \beta_j < b_j - t_\alpha \cdot s_{bj}$$

2) Проверка гипотезы $H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$

Гипотеза отвергается, если

$$F = \frac{S^2_{регр.} / k}{S^2_{ост.} / (n-k-1)} = \frac{R^2 \cdot (n-k-1)}{(1-R^2) \cdot k} > F_\alpha (k, n-k-1)$$

Проверка выполнения предпосылок МНК

Основные гипотезы (1-5) касаются поведения **остатков** $\varepsilon_i = y_i - \hat{y}_i$. При их выполнении МНК-оценки коэффициентов регрессии являются:

- **несмещёнными**
- **состоятельными**
- **эффективными**

Если характер остатков не соответствует некоторым гипотезам, модель следует корректировать

- Гипотеза случайности остатков и равенства нулю их средней величины гарантирует **несмещённость** МНК-оценок
- Гетероскедастичность сказывается на уменьшении **эффективности** МНК-оценок
- Выполнение гипотезы независимости обеспечивает **состоятельность** и **эффективность** МНК-оценок

Несмещённость оценок обеспечивается также независимостью случайных остатков

ε_i и переменных x

Графический способ проверки гипотез

- Определяются оценки случайных остатков: $\varepsilon_i = y_i - \hat{y}_i$
- Строится график зависимости остатков от теоретических значений результативного признака $\hat{y}$ либо от значений факторов x
- Если расположение точек на графике не имеет определённой направленности (т.е. точки можно поместить в горизонтальную полосу), то проверяемая гипотеза выполняется

- Проверка **случайности** остатков и их **гомоскедастичности** осуществляется по графику в системе координат $(\hat{y}_i, \varepsilon_i)$
- Проверка **независимости** остатков **от регрессоров** осуществляется по графику в системе координат (x_i, ε_i)
- Проверка **независимости** остатков – отсутствия автокорреляции соседних наблюдений – осуществляется

  с помощью расчёта и оценки значимости парных коэффициентов корреляции:

$$r_{\varepsilon_i \varepsilon_j} = \frac{\operatorname{cov}(\varepsilon_i, \varepsilon_j)}{\sigma_{\varepsilon_i} \sigma_{\varepsilon_j}}$$

Нарушение гипотезы гомоскедастичности

- Этап 1: **визуальная проверка** наличия гетероскедастичности (*график остатков*)
- Этап 2: **статистическая проверка** наличия гетероскедастичности
  - (тест Гольфельда-Квандта: упорядоченные по x наблюдения разбивают на две группы; по критерию Фишера проверяют гипотезу о равенстве дисперсий остатков в этих группах)
  - оценка зависимости остатков от значений x с помощью ранговой корреляции Спирмена
- Этап 3: **построение регрессии** с учётом гетероскедастичности (**обобщённый метод наименьших квадратов**)

Обобщённый метод наименьших квадратов (ОМНК)

При нарушении гомоскедастичности имеем:

$$V(\varepsilon_i) \neq \sigma^2 = \sigma_i^2$$

Тогда можно записать: $\sigma_i^2 = \sigma^2 K_i$

где K_i - коэффициент неоднородности дисперсии; σ^2 - неизвестно.

Это приводит к взвешенному МНК (ОМНК):

$$S^2 = \sum_{i=1}^{n} \frac{1}{K_i} (y_i - \hat{y}_i)^2 \to \min$$

В частности, парную линейную модель с гетероскедастичными остатками

$$y_i = a + bx_i + \sqrt{K_i} \cdot \varepsilon_i$$

можно привести к уравнению с гомоскедастичными остатками ($\sigma^2 = const$)

$$\frac{y_i}{\sqrt{K_i}} = \frac{a}{\sqrt{K_i}} + b\frac{x_i}{\sqrt{K_i}} + \varepsilon_i$$

и новыми переменными $\frac{y_i}{\sqrt{K_i}}; \frac{x_i}{\sqrt{K_i}}$.

Необходимо определить величины K_i и внести поправки в исходные данные.

Часто предполагается, что остатки пропорциональны значениям фактора.

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

y – издержки производства

***x*₁** – объём продукции

***x*₂** – основные фонды

***x*₃** – численность работников

- Пусть $V(\varepsilon_i) = \sigma_i^2 \cdot x_3^2 \Rightarrow$ новые факторы:

$\frac{x_1}{x_3}$ - производительность труда

$\frac{x_2}{x_3}$ - фондовооружённость

- Пусть $V(\varepsilon_i) = \sigma_i^2 \cdot x_1^2 \Rightarrow$ новые факторы :

x_2 / x_1 - фондоёмкость и x_3 / x_1 - трудоёмкость продукции

Количественная оценка гетероскедастичности

Для количественной оценки зависимости дисперсии остатков от соответствующих значений факторов используют тесты Уайта, Парка, Глейзера и др. **Тест Уайта** (White) включен в программу эконометри-ческого анализа **«Econometric Views»**.

Согласно тесту Уайта зависимость дисперсии остатков от x определяется с помощью квадратичной функции (напри-мер: $\varepsilon^2 = a + bx + cx^2 + \delta$) и проверяется по критериям Фишера и Стьюдента