СПЕЦИАЛЬНЫЕ ПРИЕМЫ МОДЕЛИРОВАНИЯ РЕГРЕССИИ

КАЧЕСТВЕННЫЕ ПРИЗНАКИ
и их учет в регрессионных моделях

РОЛЬ КАЧЕСТВЕННЫХ ПРИЗНАКОВ

- Качественные признаки приводят к неоднородности совокупности наблюдений по изучаемому признаку

УЧЕТ НЕОДНОРОДНОСТИ

1. Регрессия строится для каждой качественно отличной группы в отдельности

2. Регрессия строится для совокупности в целом, учитывая неоднородность данных с помощью ввода фиктивных переменных

Тест Чоу (первый путь)

Производится выборка объема n_1. По ней строится уравнение регрессии $y = b_{01} + b_{11}x_1 + \ldots + b_{m1}x_m + e_1$. Производится выборка объема n_2. По ней строится уравнение регрессии $y = b_{02} + b_{12}x_1 + \ldots + b_{m2}x_m + e_2$. Будет ли уравнение регрессии одним и тем же для обеих выборок?

Для каждой выборки находим сумму квадратов остатков $S_1 = \sum_{i=1}^{n} e_{i1}^2$ и $S_2 = \sum_{i=1}^{n} e_{i2}^2$ соответственно. По объединенной выборке объема $n_1 + n_2$ строим уравнение регрессии, для которого также находим сумму квадратов остатков $S_0 = \sum_{i=1}^{n} e_i^2$.

H_0: $b_{i1} = b_{i2}$, $i = 0, 1, \ldots, m$, то есть коэффициенты уравнений линейной регрессии одинаковы.

H_1: коэффициенты уравнений линейной регрессии различны.

Доверительная вероятность p. Правосторонняя проверка. $\alpha = 1 - p$. Из таблиц F-распределения находим граничную точку $F_{\alpha;m+1;n_1+n_2-2m-2}$.

Статистика $F = \dfrac{S_0 - S_1 - S_2}{S_1 + S_2} \times \dfrac{n_1 + n_2 - 2m - 2}{m + 1}$.

Пример 32. По $n = 25$ наблюдениям построено уравнение линейной регрессии, содержащее $m = 2$ фактора. Есть основания предполагать, что модель будет более реалистичной, если весь интервал наблюдений разбить на два подынтервала и оценивать уравнение линейной регрессии для каждого из них отдельно. Это связано с изменением институциональных условий между 10-м и 11-м наблюдениями. Суммы квадратов остатков для общей выборки $S_0 = 140$, для 1-го подынтервала $S_1 = 100$, для 2-го подынтервала $S_2 = 30$. Есть ли основания считать, что это разбиение целесообразно? Доверительная вероятность $p = 95\%$.

H_0: $b_{i1} = b_{i2}$, $i = 0, 1, \ldots, m$, то есть коэффициенты уравнений линейной регрессии одинаковы.

H_1: коэффициенты уравнений линейной регрессии различны.

Доверительная вероятность $p = 0{,}95$. Правосторонняя проверка. $\alpha = 1 - p = 1 - 0{,}95 = 0{,}05$. $n_1 = 10$, $n_2 = 25 - 10 = 15$.

Из таблиц F-распределения находим граничную точку $F_{\alpha;m+1;n_1+n_2-2m-2} = F_{0{,}05;2+1;10+15-2\times2-2} = 3{,}13$.

ПРОВЕРКА

Статистика $F = \dfrac{S_0 - S_1 - S_2}{S_1 + S_2} \times \dfrac{n_1 + n_2 - 2m - 2}{m + 1} =$

$= \dfrac{140 - 100 - 30}{100 + 30} \times \dfrac{10 + 15 - 2 \times 2 - 2}{2 + 1} \approx 0{,}49 < 3{,}13.$

Мы принимаем гипотезу H_0 на уровне значимости 5%. Для всего рассматриваемого периода нужно строить единое уравнение регрессии.

ФИКТИВНЫЕ ПЕРЕМЕННЫЕ в регрессии (второй путь) (dummy variables)

Это сконструированные переменные, позволяющие качественные признаки вводить в уравнение регрессии, в литературе их еще называют «структурные переменные»

Они отражают неоднородность данных как в пространстве, так и во времени

МОДЕЛИ КОВАРИАЦИОННОГО АНАЛИЗА

Модели регрессии, в которых объясняющие переменные носят как количественный, так и качественный характер, называются

ANCOVA - модели

ИСПОЛЬЗОВАНИЕ ФИКТИВНЫХ ПЕРЕМЕННЫХ

- Можно строить регрессию только для фиктивных переменных
- Можно для зависимой переменной, представленной фиктивной переменной
- Можно использовать для учета фактора сезонности
- Можно вводить в нелинейные модели, и после преобразовывать их к линейному виду

НАГРУЗКА МОДЕЛИ

- Чем больше градаций у качественной переменной, тем большим числом фиктивных переменных она вводится
- Например, m – число градаций, вводится m-1 числом независимых переменных
- Значения фиктивной переменной можно менять на противоположные, суть модели от этого не изменится
- Напомню, что число независимых переменных должно быть меньше или равно n/6 или n/7, иначе незначимые будут коэффициенты регрессии

Пример 40. Исследуется надежность станков трех производителей a, b, c. При этом учитывается возраст станка M (в месяцах) и время H (в часах) безаварийной работы до последней поломки.

Фирма	H	M	F	R	Фирма	H	M	F	R
a	280	23	0	0	a	200	52	0	0
b	230	30	1	0	b	265	20	1	0
c	112	65	1	1	c	148	70	1	1
a	176	69	0	0	c	150	62	1	1
c	90	75	1	1	b	176	40	1	0
a	176	63	0	0	a	123	66	0	0
b	216	25	1	0	a	245	20	0	0
c	110	75	1	1	c	176	39	1	1
b	45	75	1	0	b	260	25	1	0

У уравнения регрессии $H = \beta_0 + \beta_1 M + \varepsilon$ без учета различия станков различных фирм невысокий коэффициент детерминации $R^2 = 0{,}686$. Поэтому нужно учитывать производителя станков.

Качественная переменная «Производитель станков» может принимать $k = 3$ значения (a, b, c).

Поэтому нужно ввести в модель $k - 1 = 3 - 1 = 2$ фиктивных переменных F и R.

$$F = \begin{cases} 0, \text{ если производитель } a, \\ 1, \text{ если производитель } b \text{ или } c. \end{cases}$$

$$R = \begin{cases} 0, \text{ если производитель } a \text{ или } b, \\ 1, \text{ если производитель } c. \end{cases}$$

Для производителя a $F = R = 0$, для производителя b $F = 1$, $R = 0$, для производителя c $F = R = 1$.

Теперь нужно оценить коэффициенты уравнения $H = \beta_0 + \beta_1 M + \gamma_1 F + \gamma_2 R$

Способы ввода dummy variables

Пусть рассматривается уравнение $y = \beta_0 + \beta_1 x$ и в модель решено ввести фиктивную переменную D.

Это можно сделать двумя способами: $y = \beta_0 + \beta_1 x + \gamma_1 D$ и $y = \beta_0 + \beta_1 x + \gamma_1 D + \gamma_2 Dx$.

Коэффициенты γ_1 и γ_2 называются *дифференциальным свободным членом* и *дифференциальным угловым коэффициентом* соответственно.

Фиктивная переменная D во втором уравнении используется как в аддитивном ($\gamma_1 D$), так и в мультипликативном виде ($\gamma_2 Dx$), что позволяет фактически разбивать рассматриваемую зависимость на две части, связанные с периодами изменения рассматриваемой в модели переменной.

ОБОБЩЕННЫЙ МНК

Generalized Least Squares (GLS)

Ordinary Least Squares (OLS)

- Традиционный метод наименьших квадратов нельзя использовать при наличии гетероскедастичности и автокорреляции остатков
- В этом случае применяют GLS

ПРЕОБРАЗОВАНИЕ ДАННЫХ

- Обобщенный МНК (GLS) применяется к преобразованным данным и позволяет получать оценки параметров регрессии, которые являются эффективными и несмещенными

Предпосылки применения GLS

- Для гетероскедастичности

Если известна взаимосвязь остатков модели регрессии с фактором x_i, то есть найдены коэффициенты пропорциональности

ПРИМЕНЕНИЕ GLS при наличии гетероскедастичности остатков

Один из возможных методов устранения гетероскедастичности — это *метод взвешенных наименьших квадратов* (ВНК). Для его применения необходима определенная информация либо обоснованные предположения о величине дисперсий σ_i^2 отклонений ε_i, $i = 1, \ldots, n$.

§ 5.3.1. Метод взвешенных наименьших квадратов в случае пропорциональности неизвестных дисперсий отклонений квадратам значений независимой переменной

Рассмотрим случай, когда дисперсии отклонений σ_i^2 неизвестны и пропорциональны x_i^2.

Уравнение линейной регрессии $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

Разделим обе части этого уравнения на x_i.

Тогда $y_i/x_i = (\beta_0 + \beta_1 x_i + \varepsilon_i)/x_i \rightarrow y_i/x_i = \beta_0/x_i + \beta_1 + \varepsilon_i/x_i$.

Обозначим $z_i = y_i/x_i$, $t_i = 1/x_i$, $v_i = \varepsilon_i/x_i$.

Тогда $z_i = \beta_1 + \beta_0 t_i + v_i$.

Для этого уравнения уже выполнено условие гомоскедастичности. Методом наименьших квадратов находим оценки коэффициентов β_0, β_1 и возвращаемся к исходному уравнению $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

В случае, когда число факторов $m > 1$, исходное уравнение делится на переменную, которая в максимальной степени связана с σ_i.

Пример 35. Для предприятий области анализируется зарплата y в зависимости от количества сотрудников x. Данные по $n = 30$ предприятиям приведены в таблице.

x	y					
100	75,5	75,5	77,5	78,5	80	81
200	80,5	82	84,5	85	85,5	86,5
300	85,5	88,5	90	91	95	96
400	93	93,5	97,5	99	102,5	105
500	102	105,5	107	110,5	115	118,5

Уравнение линейной регрессии $y = \beta_0 + \beta_1 x + \varepsilon$.

Мы видим, что с ростом x разброс значений y увеличивается.

Например, при $x = 100$ размах вариации переменной y равен $81 - 75{,}5 = 5{,}5$, а при $x = 500$ размах вариации переменной y равен $118{,}5 - 102 = 16{,}5$. Поэтому можно ожидать наличие гетероскедастичности.

Проверим с помощью теста Голдфелда-Квандта гипотезу о наличии гетероскедастичности. Возьмем $k = 12$.

Доверительная вероятность $p = 0{,}95$. Тогда $\alpha = 1 - p = 1 - 0{,}95 = 0{,}05$. У нас число факторов $m = 1$.

По F-таблицам находим граничную точку $F_{\alpha;k-m-1;k-m-1} = F_{0{,}05;12-1-1;12-1-1} = F_{0{,}05;10;10} = 2{,}98$.

x_i	y_i	e_i	e_i^2
100	75,5	−1,15	1,32
100	75,5	−1,15	1,32
100	77,5	0,85	0,72
100	78,5	1,85	3,42
100	80	3,35	11,22
100	81	4,35	18,92
200	80,5	−3,94	15,54
200	82	−2,44	5,96
200	84,5	0,06	0,00
200	85	0,56	0,31
200	85,5	1,06	1,12
200	86,5	2,06	4,24
300	85,5	−6,73	45,34
300	88,5	−3,73	13,94
300	90	−2,23	4,99
300	91	−1,23	1,52
300	95	2,77	7,65
300	96	3,77	14,19
400	93	−7,03	49,35
400	93,5	−6,53	42,58
400	97,5	−2,53	6,38
400	99	−1,03	1,05
400	102,5	2,47	6,13
400	105	4,97	24,75
500	102	−5,82	33,83
500	105,5	−2,32	5,37
500	107	−0,82	0,67
500	110,5	2,68	7,20
500	115	7,18	51,60
500	118,5	10,68	114,13

Поясним, как заполняется таблица. Значения первых двух столбцов взяты из условия. В третьем столбце указаны отклонения e_i (получены с помощью надстройки *Пакет анализа* пакета Excel). 4-й столбец – это квадраты чисел 3-го столбца. Результаты округляем до двух цифр после запятой.

Суммы квадратов отклонений равны соответственно $S_1 = \sum_{i=1}^{k} e_i^2 = \sum_{i=1}^{12} e_i^2 \approx 64{,}11$ и $S_3 = \sum_{i=n-k+1}^{n} e_i^2 = \sum_{i=30-12+1}^{30} e_i^2 = \sum_{i=19}^{30} e_i^2 \approx$ $\approx 343{,}03$. Статистика $F = S_3/S_1 = 343{,}03/64{,}11 \approx 5{,}35$.

Так как $F > F_{\alpha;k-m-1;k-m-1}$ $(5{,}35 > 2{,}98)$, то на уровне значимости 5% принимается гипотеза о наличии гетероскедастичности.

Устраним гетероскедастичность. Предположим, что неизвестные дисперсии отклонений σ_i^2 пропорциональны x_i^2.

Уравнение линейной регрессии $y = \beta_0 + \beta_1 x$. Разделим обе части этого уравнения на x.

Тогда $y/x = (\beta_0 + \beta_1 x)/x \rightarrow y/x = \beta_0 \times 1/x + \beta_1$.

Обозначим $z = y/x$, $t = 1/x$ и перейдем к уравнению $z = \beta_1 + \beta_0 t$. Заполним таблицу.

x	y	t	z
100	75,5	0,010	0,76
100	75,5	0,010	0,76
100	77,5	0,010	0,78
100	78,5	0,010	0,79
100	80	0,010	0,80
100	81	0,010	0,81
200	80,5	0,005	0,40
200	82	0,005	0,41
200	84,5	0,005	0,42
200	85	0,005	0,43
200	85,5	0,005	0,43
200	86,5	0,005	0,43
300	85,5	0,003	0,29
300	88,5	0,003	0,30
300	90	0,003	0,30
300	91	0,003	0,30
300	95	0,003	0,32
300	96	0,003	0,32
400	93	0,003	0,23
400	93,5	0,003	0,23
400	97,5	0,003	0,24
400	99	0,003	0,25
400	102,5	0,003	0,26
400	105	0,003	0,26
500	102	0,002	0,20
500	105,5	0,002	0,21
500	107	0,002	0,21
500	110,5	0,002	0,22
500	115	0,002	0,23
500	118,5	0,002	0,24

Поясним, как заполняется таблица. Значения первых двух столбцов взяты из условия. В 3-м столбце указываются обратные величины чисел 1-го столбца (результат округляется до трех цифр после запятой). 4-й столбец равен частному 1-го и 2-го столбцов (результат округляется до двух цифр после запятой).

По данным 3-го и 4-го столбцов с помощью пакета Excel найдем $\beta_0 \approx 70{,}66$ и $\beta_1 \approx 0{,}07$. Тогда $y = \beta_0 + \beta_1 x$, то есть $y = 70{,}66 + 0{,}07x$.

§ 5.3.2. Метод взвешенных наименьших квадратов в случае пропорциональности неизвестных дисперсий отклонений значениям независимой переменной

Рассмотрим случай, когда дисперсии отклонений σ_i^2 неизвестны и пропорциональны x_i.

Уравнение линейной регрессии $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

Разделим обе части этого уравнения на $\sqrt{x_i}$.

Тогда $y_i/\sqrt{x_i} = (\beta_0 + \beta_1 x_i + \varepsilon_i)/\sqrt{x_i} \rightarrow y_i/\sqrt{x_i} = \beta_0/\sqrt{x_i} + \beta_1\sqrt{x_i} + \varepsilon_i/\sqrt{x_i}$.

Обозначим $z_i = y_i/\sqrt{x_i}$, $t_{1i} = 1/\sqrt{x_i}$, $t_{1i} = \sqrt{x_i}$, $v_i = \varepsilon_i/\sqrt{x_i}$.

Тогда $z_i = \beta_0 t_{1i} + \beta_1 t_{2i} + v_i$.

Для этого уравнения уже выполнено условие гомоскедастичности. Методом наименьших квадратов находим оценки коэффициентов β_0, β_1 и возвращаемся к исходному уравнению $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

На практике имеет смысл применить несколько методов определения гетероскедастичности и способов ее устранения.

ПРИМЕНЕНИЕ GLS
при наличии автокорреляции остатков

Возможно, автокорреляция вызвана отсутствием в модели важной объясняющей переменной. Нужно попытаться определить данный фактор и включить его в модель. Также можно попробовать изменить форму зависимости. Но если все разумные процедуры изменения спецификации модели исчерпаны, а автокорреляция имеет место, то можно воспользоваться авторегрессионным преобразованием.

Для простоты ограничимся моделью парной линейной регрессии и авторегрессионной схемой первого порядка AR(1).

$y = \beta_0 + \beta_1 x + \varepsilon.$

Вместо переменных y, x рассмотрим переменные y^*, x^*, значения которых вычисляются по правилу $y_i^* = y_i - \rho y_{i-1}$, $x_i^* = x_i - \rho x_{i-1}$, $i = 2, \ldots, n$, $\rho \approx 1 - DW/2$.

Поправки Прайса-Винстена:

$$x_1^* = x_1\sqrt{1-\rho^2}, \quad y_1^* = y_1\sqrt{1-\rho^2}.$$

Положим $\beta_0^* = \beta_0(1-\rho)$. Тогда по таблице значений переменных y^*, x^* оцениваются коэффициенты уравнения $y^* = \beta_0^* + \beta_1 x^*$. Затем получаем $\beta_0 = \beta_0^*/(1-\rho)$.

Пример 38. Оцениваются коэффициенты уравнения $y = a + bx$, где значения переменных x, y — первые два столбца таблицы.

x_i	y_i	$x_i^* = x_i - 0{,}31x_{i-1}$	$y_i^* = y_i - 0{,}31y_{i-1}$
1,31	1,12	1,25	1,06
2,21	−0,36	1,80	−0,71
1,37	1,41	0,68	1,52
1,87	0,79	1,45	0,35
1,53	0,87	0,95	0,63
2,14	−0,11	1,67	−0,38
2,26	0,1	1,60	0,13
1,31	1,63	0,61	1,60
1,76	−0,07	1,35	−0,58
1,28	0,93	0,73	0,95
1,88	0,44	1,48	0,15
1,46	1,24	0,88	1,10
2,22	0,09	1,77	−0,29
1,75	0,77	1,06	0,74
1,29	1,64	0,75	1,40
1,99	0,54	1,59	0,03
2,27	−0,3	1,65	−0,47
1,29	1,43	0,59	1,52
2,28	−0,07	1,88	−0,51
1,84	0,58	1,13	0,60
2,05	0,22	1,48	0,04
2,17	0,11	1,53	0,04
1,98	0,25	1,31	0,22
1,28	2	0,67	1,92
1,29	1,67	0,89	1,05

На основании критерия Дарбина-Уотсона гипотеза об отсутствии автокорреляции остатков не может быть ни принята, ни отвергнута. Применяется авторегрессионная схема первого порядка AR(1).

$DW = 1{,}381$. $\rho \approx 1 - DW/2 = 1 - 1{,}381/2 \approx 0{,}31$.

$$y_1^* = y_1\sqrt{1-\rho^2} = 1{,}12\sqrt{1-0{,}31^2} \approx 1{,}06,$$

$$x_1^* = x_1\sqrt{1-\rho^2} = 1{,}31\sqrt{1-0{,}31^2} \approx 1{,}25.$$

Заполняем таблицу. Из каждого элемента 1-го столбца вычитаем предыдущее число 1-го столбца, умноженное на 0,31, и результат пишем в 3-м столбце (округляем до двух цифр после запятой). Аналогично для 2-го и 4-го столбцов.

По МНК находим коэффициенты уравнения $y^* = a^* + bx^*$. Тогда $a = a^*/(1-\rho) = a^*/(1-0{,}31) = a^*/0{,}69$.

Thank You !