Эконометрика

Тема 5

Тема 5. Некоторые вопросы практического использования регрессионных моделей

1) Мультиколлинеарность
2) Отбор наиболее существенных объясняющих переменных в регрессионной модели
3) Линейные регрессионные модели с переменной структурой. Фиктивные переменные
4) Критерий Г. Чоу
5) Нелинейные модели регрессии
6) Частная корреляция

1. Мультиколлинеарность

Под **мультиколлинеарностью** понимается высокая взаимная коррелированность объясняющих переменных.

Формы мультиколлинеарности

Функциональная

Матрица X'X особенная и | X'X | = 0

нарушается предпосылка 6 регрессионного анализа

невозможность решения соответствующей системы нормальных уравнений и получения оценок параметров регрессионной модели

Стохастическая ***(встречается чаще)***

Матрица X'X неособенная, но | X'X | очень мал

Вектор оценок b и его ковариационная матрица пропорциональны матрице $(X'X)^{-1}$, т.е. их элементы обратно пропорциональны определителю | X'X |

значительные средние квадратические отклонения коэффициентов регрессии b_0, $b_1, \ldots, b_p$ и ***оценка их значимости по t-критерию не имеет смысла***

1. Мультиколлинеарность

Эвристические подходы по выявлению мультиколлинеарности

Анализ корреляционной матрицы между объясняющими переменными X_1, X_2,…, X_p и выявление пар переменных, имеющих высокие коэффициенты корреляции (обычно больше 0,8). Если такие переменные существуют, то говорят о мультиколлинеарности между ними.

Нахождение множественных коэффициентов детерминации между одной из объясняющих переменных и некоторой группой из них. Наличие высокого множественного коэффициента детерминации (обычно больше 0,6) свидетельствует о мультиколлинеарности.

Исследование матрицы X'X. Если | X'X | либо ее минимальное собственное значение близки к нулю или есть значительное отклонение максимального собственного значения матрицы X'X от ее минимального собственного значения, то это говорит о наличии мультиколлинеарности.

1. Мультиколлинеарность

Методы устранения или уменьшения мультиколлинеарности:

1) Из ***двух объясняющих переменных, имеющих высокий коэффициент корреляции (больше 0,8), одну переменную исключают из рассмотрения.*** Выбор исключаемой переменной:
а) на основании экономических соображений;
б) переменная с меньшим коэффициентом корреляции с зависимой переменной;
в) переменная с большими коэффициентами корреляции с другими независимыми переменными

2) Переход ***от несмещенных оценок, определенных по МНК, к смещенным оценкам, обладающим меньшим рассеянием*** относительно оцениваемого параметра: использование **«ридж-регрессии»** со смещенными оценками, задаваемыми вектором:

$$\hat{\beta}_\tau = \left(X'X + \tau E_{p+1} \right)^{-1} X'Y$$

где τ - некоторое положительное число («гребень» или «хребет»), E_{p+1} — единичная матрица (p+1)-го порядка

3) Переход ***от исходных объясняющих переменных X_1, X_2,…, X_n, связанных между собой достаточно тесной корреляционной зависимостью, к новым переменным, представляющим линейные комбинации исходных***. При этом новые переменные должны быть ***слабокоррелированными*** либо вообще некоррелированными.

2. Отбор наиболее существенных объясняющих переменных в регрессионной модели

Еще одним из возможных методов устранения или уменьшения мультиколлинеарности является использование ***пошаговых процедур отбора наиболее информативных переменных***:

1. **Процедура присоединения объясняющих переменных.** На 1-м шаге рассматривается лишь 1 объясняющая переменная, имеющая с зависимой переменной Y наибольший R^2. На 2-м шаге включается в регрессию новая объясняющая переменная, которая вместе с первоначально отобранной образует пару объясняющих переменных, имеющую с Y наиболее высокий (скорректированный) R^2. На 3-м шаге вводится в регрессию еще 1 объясняющая переменная, которая вместе с двумя первоначально отобранными образует тройку объясняющих переменных, имеющую с Y наибольший (скорректированный) R^2 коэффициент детерминации, и т. д. Процедура введения новых переменных продолжается до тех пор, пока будет увеличиваться соответствующий (скорректированный) R^2.

2. **Процедура исключения факторов**. Сначала в модель включаются все факторы. Затем после построения уравнения регрессии из модели исключают фактор, коэффициент при котором незначим и имеет наименьшее значение t-критерия. После этого получают новое уравнение регрессии и снова проводят оценку значимости всех оставшихся коэффициентов регрессии. Процесс исключения факторов продолжается до тех пор, пока модель не станет удовлетворять определенным условиям и все коэффициенты регрессии не будут значимы.

3. Линейные регрессионные модели с переменной структурой. Фиктивные переменные

Качественные признаки могут существенно влиять на структуру линейных связей между переменными. В этом случае говорят об исследовании регрессионных моделей ***с переменной структурой*** или ***построении регрессионных моделей по неоднородным данным***.

Для этого используются ***фиктивные переменные*** – обычно дихотомические (бинарные, булевы) переменные, которые принимают всего ***два значения: «0» или «1»*** *(например, значение переменной Z_1 по фактору «пол»: Z_1= 0 для работников-женщин и Z_1=1 - для мужчин).*

Пример 1 модели множественной линейной регрессии с переменной структурой - зависимость размера заработной платы Y работников от количественных факторов X_1, X_2,…, X_n и от качественного фактора Z_1 - «пол работника»:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \alpha_1 z_{i1} + \varepsilon_i\,,\ i=1,\ldots,\ n,$$

$$\text{где } z_{i1} = \begin{cases} 1, & \text{если } i\text{-й работник мужского пола;} \\ 0 & \text{если } i\text{-й работник женского пола.} \end{cases}$$

3. Линейные регрессионные модели с переменной структурой. Фиктивные переменные

Если рассматриваемый качественный признак имеет ***несколько (k) уровней (градаций)***, то в уравнение вводят ***(k–1) бинарных переменных.***

Пример 2 модели множественной линейной регрессии с переменной структурой - зависимость размера заработной платы Y работников от количественных факторов X_1, X_2,…, X_n, от качественного фактора Z_1 - «пол работника» и качественных бинарных переменных Z_{21} и Z_{22}, учитывающих уровень образования работников *(k=3 – уровни образования работников: высшее, среднее, начальное)*:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \alpha_1 z_{i1} + \alpha_{21} z_{i21} + \alpha_{22} z_{i22} + \varepsilon_i \ ,$$

где $z_{i21} = \begin{cases} 1, & \text{если } i\text{-й работник имеет высшее образование;} \\ 0 & \text{во всех остальных случаях;} \end{cases}$

$$z_{i22} = \begin{cases} 1, & \text{если } i\text{-й работник имеет среднее образование;} \\ 0 & \text{во всех остальных случаях.} \end{cases}$$

если i-й работник имеет начальное образование, это будет отражено в модели парой значений $z_{i21} = 0$, $z_{i22} = 0$.

3. Линейные регрессионные модели с переменной структурой. Фиктивные переменные

Пример 1 и Пример 2 рассмотренные выше отражали влияние качественного признака (фиктивных переменных) только на значения переменной Y, т. е. ***на свободный член*** уравнения регрессии. В более сложных моделях может быть отражена также зависимость фиктивных переменных ***на сами параметры при переменных*** регрессионной модели.

Пример 3: при наличии в модели объясняющих переменных количественной X_1 и фиктивных Z_{11}, Z_{12}, Z_{21}, Z_{22}, из которых Z_{11}, Z_{12} влияют только на значение коэффициента при X_1, а Z_{21}, Z_{22} - только на величину свободного члена уравнения, такая регрессионная модель примет вид:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_{11}(z_{i11} x_{i1}) + \beta_{12}(z_{i12} x_{i1}) + \\ + \alpha_{21} z_{i21} + \alpha_{22} z_{i22} + \varepsilon_i, \quad i = 1, \ldots, n.$$

Модели такого типа используются, например, при исследовании зависимости объема потребления Y некоторого продукта от дохода потребителя X, когда одни качественные признаки (например, фактор сезонности) влияют лишь на количество потребляемого продукта (свободный член уравнения регрессии), а другие (например, уровень доходности домашнего хозяйства) - на параметр β_1 при X, интерпретируемый как «склонность к потреблению».

4. Критерий Г. Чоу

при одних условиях при несколько измененных условиях.

однородны можно ли объединить две выборки в одну

тест (критерий) Г. Чоу

4. Критерий Г. Чоу

По каждой выборке строятся две линейные регрессионные модели:

$$y_i = \beta'_0 + \sum_{j=1}^{p} \beta'_j x_{ij} + \varepsilon'_i\,, \quad i = 1, \ldots, n_1;$$

$$y_i = \beta''_0 + \sum_{j=1}^{p} \beta''_j x_{ij} + \varepsilon''_i\,, \quad i = n_1+1, \ldots, n_1+n_2.$$

Проверяемая нулевая гипотеза имеет вид — H_0: $\beta'=\beta''$; $D(\varepsilon')=D(\varepsilon'')=\sigma^2$, где $\beta'=\beta''$ — векторы параметров двух моделей; $\varepsilon', \varepsilon''$ — их случайные возмущения.

Если нулевая гипотеза H_0 верна, то две регрессионные модели можно объединить в одну объема $n = n_1+n_2$:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i\,, \quad i = 1,2 \ldots, n.$$

4. Критерий Г. Чоу

Согласно **критерию Г. Чоу** нулевая гипотеза H_0 ***отвергается*** на уровне значимости α , если статистика:

$$F = \frac{\left(\sum_{i=1}^{n} e_i^2 - \sum_{i=1}^{n_1} e_i^2 - \sum_{i=n_1}^{n} e_i^2\right)(n-2p-2)}{\left(\sum_{i=1}^{n_1} e_i^2 + \sum_{i=n_1}^{n} e_i^2\right)(p+1)} > F_{\alpha;p+1;n-2p-2},$$

где $\sum_{i=1}^{n} e_i^2,$ $\sum_{i=1}^{n_1} e_i^2,$ $\sum_{i=n_1}^{n} e_i^2$ — остаточные суммы квадратов соответственно для объединенной, первой и второй выборок; $n=n_1+n_2$.

p – количество факторов в каждой регрессии;

$F_{\alpha;p+1;n-2p-2}$ - критерий Фишера для соответствующего уровня значимости и количества степеней свободы.

5. Нелинейные модели регрессии

Виды нелинейных моделей регрессии

Модели нелинейные *по переменным*

Модели нелинейные *по параметрам*

Пример:

$$y_i = \beta_0 + \beta_1 x_{i1}^2 + \beta_2 \sqrt{x_{i2}} + \varepsilon_i,$$

$$i = 1, \ldots, n$$

Пример:

$$Y = AK^{\alpha} L^{\beta} \varepsilon$$

(производственная функция Кобба—Дугласа)

5. Нелинейные модели регрессии

Методы оценки параметров нелинейных моделей

Линеаризация модели:

с помощью подходящих преобразований исходных переменных исследуемую зависимость представляют в виде линейного соотношения между преобразованными переменными.

Нелинейная оптимизация на основе исходных переменных:

применяется в случае, когда подобрать соответствующее линеаризующее преобразование не удается.

5. Нелинейные модели регрессии

Пример 1: решение модели нелинейной по переменным с помощью линеаризации:

$$y_i = \beta_0 + \beta_1 x_{i1}^2 + \beta_2 \sqrt{x_{i2}} + \varepsilon_i, \qquad i = 1, \ldots, n$$

Вводим новые переменные: $Z_1 = X_1^2, \quad Z_2 = \sqrt{X_2}$

Получаем линейную модель: $y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \varepsilon_i, \quad i = 1, \ldots, n,$

Параметры линеаризованной модели находятся обычным ММК (необходимо определенное уточнение полученных оценок ***для получения оценок по исходным переменным***)

5. Нелинейные модели регрессии

Пример 2: решение модели нелинейной по параметрам с помощью линеаризации:

$$Y = AK^{\alpha} L^{\beta} \varepsilon$$

- производственная функция Кобба-Дугласа

Логарифмирование обеих частей уравнения

Получаем линейную модель:

$$\ln y_i = \ln A + \alpha \ln K_i + \beta \ln L_i + \ln \varepsilon_i, \quad i = 1, \ldots, n.$$

Параметры линеаризованной модели находятся обычным ММК (необходимо определенное уточнение полученных оценок ***для получения оценок по исходным переменным***)

6. Частная корреляция

Если переменные коррелируют друг с другом, то на значении выборочного коэффициента корреляции частично сказывается влияние ***другof*** переменных. В связи с этим часто возникает необходимость исследовать ***частную корреляцию*** между переменными при ***исключении (элиминировании) влияния*** одной или нескольких переменных.

Выборочным частным коэффициентом корреляции между переменными X_i и X_j при ***фиксированных*** значениях остальных (p – 2) переменных называется выражение:

$$r_{ij.1,2\ldots p} = \frac{-q_{ij}}{\sqrt{q_{ii}q_{jj}}},$$

где q_{ii} и q_{jj} – алгебраические дополнения элементов r_{ii} и r_{jj} *матрицы выборочных коэффициентов корреляции*:
r_{ij} – выборочные парные линейные коэффициенты корреляции

$$q_p = \begin{pmatrix} 1 & r_{12} & \ldots & r_{1p} \\ r_{21} & 1 & \ldots & r_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ r_{p1} & r_{p2} & \ldots & 1 \end{pmatrix}$$

6. Частная корреляция

Если *p=3 (3 переменные)*, то выборочный частный коэффициент корреляции:

$$r_{ij.k} = \frac{r_{ij} - r_{ik}r_{jk}}{\sqrt{\left(1 - r_{ik}^2\right)\left(1 - r_{jk}^2\right)}}$$

Диапазон значений выборочного частного коэффициента корреляции: **от +1 до -1**;

Статистическую значимость частного коэффициента корреляции $r_{ij.12\ldots p}$ оценивают *так же, как и обычного коэффициента корреляции r – с использованием критерия Стьюдента*, но при этом используют в качестве *числа наблюдений n'=n–p+2*

Вопросы изученные в Теме 5:

1) **Мультиколлинеарность**
2) **Отбор наиболее существенных объясняющих переменных в регрессионной модели**
3) **Линейные регрессионные модели с переменной структурой. Фиктивные переменные**
4) **Критерий Г. Чоу**
5) **Нелинейные модели регрессии**
6) **Частная корреляция**