ЧЕЧЕНСКИЙ ГОСУДАРСТВЕННЫЙ ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ

Экономико-математическое моделирование

цисов Магомед-Рамзан Бувайсарович,
дидат экономических наук, доцент

Лекция 4. Обобщенная модель множественной регрессии

План

1. Типы данных
2. Классическая модель множественной регрессии: предпосылки
3. Качество подгонки:
   – стандартная ошибка регрессии (SEE),
   – и скорректированный
4. Гипотезы и доверительные интервалы
5. Спецификация уравнения
6. Контрольные переменные и замещающие переменные

1. Типы данных в статистических исследован

- Пространственные выборки
- Временные ряды
- Панельные данные

2. Классическая линейная модель множественной регрессии (КЛММР)

$$y_i = a_0 + a_1 x_i^{(1)} + a_2 x_i^{(2)} + \ldots + a_p x_i^{(p)} + \varepsilon_i$$

y_i – значения зависимой переменной,

$x_i^{(p)}$ – значения независимых переменных (регрессоров),

p – число коэффициентов в модели,

ε_i – случайные ошибки.

Предпосылки КЛММР

1) модель линейна по параметрам и правильно специфицирована

$$y_i = a_0 + a_1 x_i^{(1)} + a_2 x_i^{(2)} + \ldots + a_p x_i^{(p)} + \varepsilon_i$$

2) $x^{(1)} \ldots x^{(p)}$ – детерминированные переменные, которые являются линейно независимыми

3) математическое ожидание случайных ошибок равно нулю: $E(\varepsilon_i) = 0$

Предпосылки КЛММР

4) случайные ошибки имеют постоянную дисперсию: $\boldsymbol{V(\varepsilon_i)=\sigma^2=const}$

5) случайные ошибки, соответствующие разным наблюдениям не зависят друг от друга (не коррелированны) $\boldsymbol{Cov(\varepsilon_i, \varepsilon_j) = 0}$ при $\boldsymbol{i \neq j}$

6)* случайные ошибки $\boldsymbol{\varepsilon_i}$ имеют нормальное распределение

Теорема Гаусса – Маркова

Если выполнены условия (1) – (5), **то** оценка коэффициентов модели, полученная по МНК, является:

(а) несмещенной;

(б) эффективной.

3. Качество подгонки модели

1. Стандартная ошибка регрессии
2. Коэффициент детерминации R^2
3. Скорректированный (нормированный) коэффициент детерминации R^2

Оценка дисперсии случайной ошибки

Если выполнены условия (1)–(6), то несмещенная оценка дисперсии случайных ошибок σ^2 имеет вид:

$$\hat{\delta}^2 = S^2 = \frac{1}{n - p} \sum_{i=1}^{n} \varepsilon_i^2$$

S^2 – используется для расчета стандартных ошибок коэффициентов и стандартной ошибки регрессии.

Стандартная ошибка регрессии

(Standard Error of Estimate – не путать с ESS)

$$SEE = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n} \varepsilon_i^{\,2}}{n - p}}$$

Измеряет среднюю величину ошибки модели.

Используется для оценки качества подгонки модели: чем меньше SEE, тем точнее модель.

Можно использовать для сравнения нескольких однотипных уравнений регрессии.

Коэффициент детерминации R^2

Ранее нами показано, что:

$$y_i = \tilde{y}_i + e_i \qquad Var(y) = Var(\varepsilon) + Var(\tilde{y})$$

Это означает, что мы можем разложить $Var(y)$ на две части: $Var(\tilde{y})$ часть, которая «объясняется» уравнением регрессии и $Var(\varepsilon)$ «необъяснённую» часть.

$$R^2 = \frac{Var(\tilde{y})}{Var(y)} = \frac{RSS}{TSS} = 1 - \frac{Var(e)}{Var(y)}$$

$$0 \le R^2 \le 1$$

Скорректированный (нормированный) R^2

R^2 с учетом штрафа за число переменных:

$$R^2_{adj} = R^2 - \frac{p - 1}{n - p}(1 - R^2)$$

R_{adj}^2 можно использовать для сравнения моделей с одинаковой зависимой переменной, но разным числом независимых переменных.

Использование R^2 и R^2_{adj}

Есть соблазн свести выбор уравнения к задаче максимизации R^2 или R^2_{adj}. Не стоит этого делать:

1. Высокий R^2 (или R^2_{adj}) **говорит** о том, что регрессоры предсказывают большую долю изменений ***y***.

2. Высокий R^2 (или R^2_{adj}) **не говорит** о том, что вы верно выявили причинно-следственную связь между переменными.

Тестирование гипотез

1. Тестирование значимости коэффициента
2. Тестирование гипотезы $a_i = A$
3. Доверительный интервал для коэффициента
4. Тестирование значимости уравнения

Тестирование значимости коэффициента

Рассматриваемая модель:

$$y_i = a_0 + a_1 x_i^{(1)} + a_2 x_i^{(2)} + \ldots + a_p x_i^{(p)} + \varepsilon_i$$

Тестируемая гипотеза H0: $a_p = 0$ «Переменная $x^{(p)}$ не оказывает значимого влияния на переменную y».

Альтернативная гипотеза H1: $a_p \neq 1$ «Переменная $x^{(p)}$ оказывает значимое влияние на переменную y».

Тестирование значимости коэффициента

Шаг 1.Вычисляем расчетное значение *t*-статистики

$$t_{расч} = \frac{\widetilde{a}_p}{SE(\widetilde{a}_p)}$$

Шаг 2. Выбираем **уровень значимости**

Уровень значимости – **вероятность ошибки** первого рода, то есть вероятность отклонить гипотезу H0, если на самом деле гипотеза H0 верна.

В эконометрике обычно используется уровень значимости α=0,01=1% или α = 0,05=5%.

Тестирование значимости коэффициента

Шаг 3. Из таблиц *t*-распределения Стьюдента находим критическое значение *t*-статистики $t_{кр}$. Оно зависит от уровня значимости α и так называемого числа степеней свободы, которое в случае нашего теста равно (*p*–*n*).

Шаг 4. Сравниваем расчетное и критическое значение t- статистик

$$\left|t_{расч}\right| < t_{кр},$$

Если

то гипотеза H0 не отклоняется (принимается), то есть мы делаем вывод о том, что переменная $\boldsymbol{x^{(p)}}$ не оказывает значимого влияния на переменную ***y***. В этом случае коэффициент при переменной $\boldsymbol{x^{(p)}}$ называют незначимым.

Тестирование значимости уравнения

Рассматриваемая модель:

$$y_i = a_0 + a_1 x_i^{(1)} + a_2 x_i^{(2)} + \ldots + a_p x_i^{(p)} + \varepsilon_i$$

Тестируемая гипотеза H0: $a_0 \ldots a_p = 0$ «Все переменные $x^{(1)} \ldots x^{(p)}$ не оказывают значимого влияния на переменную y».

Альтернативная гипотеза H1: «Хотя бы одна из переменных $x^{(1)} \ldots x^{(p)}$ оказывает значимое влияние на переменную y».

Тестирование значимости уравнения

Алгоритм проведения теста:

Шаг 1. Вычисляем расчетное значение *F*-статистики. $F_{расч.} = \frac{R^2}{1 - R^2} \frac{n - p}{p - 1}$

Шаг 2. Выбираем уровень значимости.

Шаг 3. Из таблиц *F*-распределения Фишера находим критическое значение *F*-статистики ($F_{кр}$).

Оно зависит от уровня значимости α и от

Тестирование значимости уравнения

Шаг 4. Сравниваем расчетное и критическое значение
F-статистик.

Если $F_{расч.} < F_{кр.}$,

то гипотеза H0 не отклоняется (принимается), то есть мы делаем вывод о том, что все переменные $x^{(1)}...x^{(p)}$ не оказывает значимого влияния на переменную **y**.

В этом случае уравнение называют незначимым.

В противном случае гипотеза H0 не принимается

Литература к лекции

1. **Магнус** Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс: Учеб. – 6-е изд., перераб. и доп. – М.: Дело, 2004. **Глава 3**
2. **Доугерти** К. Введение в эконометрику: Учебник. 3-е изд. / Пер. с англ. – М.: ИНФРА-М, 2009. **Глава 3**

5. Спецификация уравнения

План

1. Что будет, если специфицировать уравнение неправильно?

2. Контрольные переменные и замещающие переменные

3. Как специфицировать уравнение правильно?

Спецификация уравнения

Под спецификацией уравнения мы будем понимать:

- выбор набора переменных,
- выбор функциональной формы зависимости.

Спецификация уравнения: выбор набора переменных

Последствия ошибок спецификации:

- Последствия **не**включения в модель существенной переменной
- Последствия включения в модель **не**существенной переменной

Продолжение...

Последствия **не**включения в модель существенной переменной

Для краткости будем называть переменную **существенной**, если она должна быть включена в уравнение (согласно правильной теории)

Последствия **НЕ**включения в модель существенной переменной

1. Коэффициенты при оставшихся переменных могут оказаться смещенными — **omitted variable bias**

2. Их стандартные ошибки и другие показатели качества становятся некорректными и не могут быть использованы для суждения о качестве уравнения (неправильный выбор функциональной формы приводит к аналогичным проблемам)

Спецификация уравнения: выбор набора переменных

Последствия включения в модель **не**существенной переменной

Для краткости будем называть переменную **несущественной**, если она не должна быть включена в уравнение (согласно правильной теории)

Последствия включения в модель **НЕ**существенной переменной

1. Не теряется возможность правильной оценки и интерпретации уравнения

2. Коэффициенты при переменных остаются несмещенными

3. Стандартные ошибки растут, следовательно точность оценок падает

4. Увеличивается риск мультиколлинеарности

6. Контрольные переменные и замещающие переменные

Пример: пусть нас интересует ответ на вопрос, есть ли дискриминация на рынке труда?

Для ответа на этот вопрос нельзя ограничиваться такой моделью, так как на зарплату влияют многие факторы, и оценка коэффициента будет смещена из-за пропуска существенных переменных (omitted variable bias)

Контрольные переменные

Пример: пусть нас интересует ответ на вопрос, есть ли дискриминация на рынке труда?

Чтобы избежать смещения оценок из-за пропуска существенных переменных, в модель добавляют остальные факторы, которые влияют на зарплату (образование, возраст и т.п.):

Где:

Пол работника () — переменная, влияние которой нас интересует (**variable of interest**)

Образование () и возраст () — контрольные переменные (**control variables**)

Контрольные переменные

Variable of interest — переменная, влияние которой на зависимую переменную нас интересует.

Контрольные переменные (control variables) — переменные, которые мы включаем в модель для того, чтобы избежать смещения коэффициента при интересующей нас переменной.

Замещающие переменные

Рассмотрим следующую ситуацию.

Пусть истинная модель имеет вид:

Если оценить модель без то оценка коэффициента будет смещена.

Как решить эту проблему?

Продолжение...

Предположим, у нас есть данные о переменной , которая связана с :

Мы не знаем параметров и , но знаем, что связь есть.

Продолжение...

Пример: у нас нет данных о таланте, респондента, но есть данные о результатах его IQ-теста.

Подставим одно уравнение в другое:

Продолжение...

Плохая новость: коэффициент мы оценить не сможем (*понятно ли почему?*)

Хорошая новость: если оценить последнюю модель, то оценка будет **не**смещенной (*докажите это!*)

=> исходная проблема решена

Продолжение...

В рассмотренной ситуации переменная называется **замещающей переменной (proxy variable)** для переменной

Таким образом, удачный выбор замещающей переменной может решить проблему смещения оценки интересующего нас коэффициента

3. Как специфицировать уравнение правильно?

Простых рецептов нет.

Не используйте механические алгоритмы!

Используйте здравый смысл и хорошее понимание содержательной стороны исследуемой проблемы.

Один из подходов к спецификации уравнения (1/2)

(1) Сформулируйте точно, наличие какой причинно-следственной связи (causal effect) вы хотите проверить

Например: *влияет ли пол работника на его заработную плату?*

(2) Подумайте, какие контрольные переменные следует включить в уравнение, чтобы избежать omitted variable bias?

Например: *образование и опыт работы*

Один из подходов к спецификации уравнения (1/2)

(3) На основе пунктов (1) и (2) запишите и оцените «базовую» модель регрессии

(4) Оцените несколько альтернативных моделей

Используйте дополнительные переменные или другую функциональную форму зависимости

(5) Оказались ли добавленные переменные значимыми? Улучшают ли

Некоторые критерии для включения переменной в модель

1. Роль переменной в уравнении опирается на прочные теоретические основания
— Ну или хотя бы на здравый смысл
2. Переменная статистически значима
3. Оценки других коэффициентов сильно меняются при включении новой переменной в модель
— Это значит, что до ее включения они страдали от omitted variable bias
4. Скорректированный существенно увеличивается в результате включения переменной в модель
— Но не пытайтесь механически максимизировать . Ранее мы обсуждали ограниченность этого показателя

Тест Рамсея (RESET – Regression Equation Specification Error Test)

1)H0: верна спецификация модели

$$y_i = \beta_1 + \beta_2 x_i^{(2)} + \ldots + \beta_k x_i^{(k)} + \varepsilon_i$$

2)оцениваем параметры исходной модели =>

3)оцениваем параметры вспомогательной модели:

$$y_i = \beta_1 + \ldots + \beta_k x_i^{(k)} + \alpha_2 (\hat{y}_i)^2 + \alpha_3 (\hat{y}_i)^3 + u_i$$

3) Тестируем гипотезу: $\alpha_2 = \alpha_3 = 0$. Если эта гипотеза принимается, то делаем вывод о том, что спецификация исходного уравнения верна.

* Иногда используют другое число дополнительных переменных (например, одну или три)

Литература к лекции

- **Магнус** Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс: Учеб. — 6-е изд., перераб. и доп. — М.: Дело, 2004. **Глава 4**
- **Доугерти** К. Введение в эконометрику: Учебник. 3-е изд. / Пер. с англ. — М.: ИНФРА-М, 2009. **Глава 6**
- **Анатольев Станислав (2008), «Оформление эконометрических отчетов», Квантиль №4**
  http://quantile.ru/04/04-SA.pdf