

Спецификация уравнения регрессии

Спецификация уравнения регрессии

Под спецификацией уравнения мы будем понимать:

- выбор набора переменных,
- выбор функциональной формы зависимости



Спецификация уравнения регрессии

1. Что будет, если специфицировать уравнение неправильно?
2. Контрольные переменные и замещающие переменные
3. Как специфицировать уравнение правильно?

Если специфицировать уравнение неправильно

Последствия ошибок спецификации:

- Последствия **невключения** в модель существенной переменной
- Последствия включения в модель **несущественной** переменной

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

Последствия НЕвключения 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)

Контрольные переменные

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

$$Wage_i = \beta_1 + \beta_2 * D_i + \varepsilon_i$$

$$D_i = \begin{cases} 0, & \text{если } i_{\text{ый}} \text{ респондент мужчина} \\ 1, & \text{если } i_{\text{ая}} \text{ респондентка женщина} \end{cases}$$

Для ответа на этот вопрос нельзя ограничиваться такой моделью, так как на зарплату влияют многие факторы, и оценка коэффициента β_2 будет смещена из-за пропуска существенных переменных (omitted variable bias)

Контрольные переменные

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

$$\begin{aligned} Wage_i &= \\ &= \beta_1 + \beta_2 * D_i + \beta_3 * Education_i + \beta_4 * Age_i + \epsilon_i \end{aligned}$$

В нашем примере:

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

Образование ($Education$) и возраст (Age) — 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

9.1. Смещение в оценках коэффициентов, вызванное невключением существенных переменных

Как уже отмечалось при формулировке теоремы Гаусса — Маркова, эта теорема справедлива, если модель правильно специфицирована, т.е. имеет правильную функциональную форму (ее выбор обсуждался в предыдущей главе), все необходимые факторы в модель включены, а «лишних» факторов нет. Но как определить, какие факторы являются необходимыми, а какие — «лишними»? Каковы последствия неправильного выбора факторов? Обо всем этом пойдет речь в текущей главе.

Для того чтобы понять последствия невключения в модель необходимых факторов, рассмотрим следующий пример.

Предположим, что данные $Y_i, X_{1i}, X_{2i}, i = 1, \dots, n$, были порождены с помощью процесса $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.

Однако исследователь об этом не подозревает и оценивает модель, в которой переменная X_2 пропущена:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon.$$

Тогда МНК-оценка коэффициента β_1 будет иметь вид (см. задачу 3.2)

$$\hat{\beta}_1 = \frac{\widehat{\text{cov}}(X_1, Y)}{\widehat{\text{var}}(X_1)}.$$

Покажем, что эта оценка будет смещенной. Действительно,

$$\begin{aligned}\hat{\beta}_1 &= \frac{\widehat{\text{cov}}(X_1, Y)}{\widehat{\text{var}}(X_1)} = \frac{\widehat{\text{cov}}(X_1, \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon)}{\widehat{\text{var}}(X_1)} = \\ &= \frac{\widehat{\text{cov}}(X_1, \beta_0)}{\widehat{\text{var}}(X_1)} + \beta_1 \frac{\widehat{\text{cov}}(X_1, X_1)}{\widehat{\text{var}}(X_1)} + \beta_2 \frac{\widehat{\text{cov}}(X_1, X_2)}{\widehat{\text{var}}(X_1)} + \frac{\widehat{\text{cov}}(X_1, \varepsilon)}{\widehat{\text{var}}(X_1)} = \\ &= 0 + \beta_1 + \beta_2 \frac{\widehat{\text{cov}}(X_1, X_2)}{\widehat{\text{var}}(X_1)} + \frac{\widehat{\text{cov}}(X_1, \varepsilon)}{\widehat{\text{var}}(X_1)}.\end{aligned}$$

Для проверки, является ли эта оценка несмещенной, найдем ее математическое ожидание:

$$\begin{aligned}E(\hat{\beta}_1) &= E\left(\beta_1 + \beta_2 \frac{\widehat{\text{cov}}(X_1, X_2)}{\widehat{\text{var}}(X_1)} + \frac{\widehat{\text{cov}}(X_1, \varepsilon)}{\widehat{\text{var}}(X_1)}\right) = \\ &= \beta_1 + \beta_2 \frac{\widehat{\text{cov}}(X_1, X_2)}{\widehat{\text{var}}(X_1)} + \frac{E(\widehat{\text{cov}}(X_1, \varepsilon))}{\widehat{\text{var}}(X_1)} = \\ &= \beta_1 + \beta_2 \frac{\widehat{\text{cov}}(X_1, X_2)}{\widehat{\text{var}}(X_1)} + \frac{1}{\widehat{\text{var}}(X_1)} \frac{1}{(n-1)} \sum_{i=1}^n (X_{1i} - \bar{X}_1) E(\varepsilon_i - \bar{\varepsilon}) = \\ &= \beta_1 + \beta_2 \frac{\widehat{\text{cov}}(X_1, X_2)}{\widehat{\text{var}}(X_1)},\end{aligned}$$

Гаусса — Маркова.

Мы показали, что оценка коэффициента β_1 будет смещенной во всех случаях, кроме следующих:

- 1) $\beta_2 = 0$, что тривиально;
- 2) $\widehat{\text{cov}}(X_1, X_2) = 0$, т.е. когда пропущенный и включенный фактор не коррелируют.

Во всех остальных случаях оценка коэффициента β_1 является смещенной и несостоятельной, причем величина смещения равна

$$\beta_2 \frac{\widehat{\text{cov}}(X_1, X_2)}{\widehat{\text{var}}(X_1)}.$$

Как следует из последней формулы, знак смещения зависит от знаков двух множителей: коэффициента при невключенном факторе и выборочной ковариации включенного и невключенного факторов. Аналогично можно показать, что оценка свободного члена также будет смещенной.

Если модель, «порождающая данные», часто именуемая в учебниках «истинной», содержит более двух переменных, а оценивается модель, из которой часть факторов исключена, то оценки коэффициентов при включенных факторах будут смещенными, как и в рассмотренном выше примере.

Пример

- «Введение в эконометрику», Дж.Сток, М. Уотсон. Стр. 341 – 343.

Замещающие переменные

Рассмотрим следующую ситуацию.

Пусть истинная модель имеет вид

$$y = \beta_1 + \beta_2 x^{(2)} + \beta_3 x^{(3)} + \varepsilon$$

При этом переменная $x^{(3)}$ ненаблюдаема:
у нас нет данных о ней.

Пример: на уровень дохода работника
влияет его талант, но у нас нет
статистических данных об уровне таланта

Замещающие переменные

Рассмотрим следующую ситуацию.

Пусть истинная модель имеет вид

$$y = \beta_1 + \beta_2 x^{(2)} + \beta_3 x^{(3)} + \varepsilon$$

Если оценить модель без $x^{(3)}$ то оценка коэффициента β_2 будет смещена.

Как решить эту проблему?

Замещающие переменные

$$y = \beta_1 + \beta_2 x^{(2)} + \beta_3 x^{(3)} + \varepsilon$$

Предположим, у нас есть данные о переменной z , которая связана с $x^{(3)}$:

$$x^{(3)} = \theta + \mu * z$$

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

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

Замещающие переменные

$$y = \beta_1 + \beta_2 x^{(2)} + \beta_3 x^{(3)} + \varepsilon$$

$$x^{(3)} = \theta + \mu * z$$

Подставим одно уравнение в другое:

$$y = \beta_1 + \beta_2 x^{(2)} + \beta_3 (\theta + \mu * z) + \varepsilon$$

$$y = (\beta_1 + \beta_3 \theta) + \beta_2 x^{(2)} + (\beta_3 \mu) * z + \varepsilon$$

$$y = \alpha_1 + \beta_2 x^{(2)} + \alpha_3 * z + \varepsilon$$

Замещающие переменные

$$y = (\beta_1 + \beta_3\theta) + \beta_2x^{(2)} + (\beta_3\mu) * z + \varepsilon$$

$$y = \alpha_1 + \beta_2x^{(2)} + \alpha_3 * z + \varepsilon$$

Плохая новость: коэффициент β_3 мы оценить не сможем (*понятно ли почему?*)

Хорошая новость: если оценить последнюю модель, то оценка β_2 будет **несмещенной** (*докажите это!*)

=> исходная проблема решена

Замещающие переменные

В рассмотренной ситуации переменная z называется **замещающей переменной (proxy variable)** для переменной $x^{(3)}$

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

Последствия включения в модель НЕсущественной переменной

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

Последствия включения в модель НЕсущественной 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

Какие переменные включать в уравнение?

Простых рецептов нет.

Не используйте механические алгоритмы.

Используйте здравый смысл и хорошее понимание содержательной стороны исследуемой проблемы.

На следующих слайдах — один из возможных подходов.



Подход к спецификации уравнения

(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

Подход к спецификации уравнения

(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 изменения модель?

Некоторые критерии для включения переменных 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 R^2 существенно увеличивается в результате включения переменной в модель
 - Но не пытайтесь механически максимизировать R^2 . Ранее мы обсуждали ограниченность этого показателя