Спецификация переменных в уравнениях регрессии

1. Ошибки спецификации.
2. Влияние неполноты включения в уравнения переменных.
3. Влияние избыточности факторов.
4. Лаговые переменные.

Моделирование

- Вопросы:
  - К каким результатам приведет включение в уравнение регрессии переменной, которой там недолжно быть;
  - Каковы последствия отсутствия переменной, которая должна присутствовать;
  - Что произойдет, если вместо некоторых исходных данных решим использовать «заменители».

Результаты неправильной спецификации переменных

- Опущена необходимая переменная –
  - Оценки коэффициентов регрессии оказываются смещенными,
  - Стандартные ошибки коэффициентов и t-тесты в целом становятся некорректными
- Включена ненужная переменная –
  - Оценки коэффициентов регрессии оказываются несмещенными, однако неэффективными;
  - Стандартные ошибки в целом корректны, но из-за эффективности будут излишне большими.

Влияние отсутствия необходимой переменной

- Проблема смещения

  истинная модель $y=\alpha + \beta_1 x_1 + \beta_2 x_2$

  строим модель $y=\alpha + \beta_1 x_1$

$$E\left\{\frac{\operatorname{cov}(x_1, y)}{\operatorname{var}(x_1)}\right\} = \beta_1 + \beta_2 \frac{\operatorname{cov}(x_1, x_2)}{\operatorname{var}(x_1)}$$

- Неприменимость статистических тестов

Свойства коэффициентов регрессии

- Интерпретация коэффициентов регрессии
- Несмещенность коэффициентов
- Точность коэффициентов
- Предположения:
  - 1) выполняются 4 условия Гаусса-Маркова
  - 2) имеется достаточное количество данных
  - 3) между независимыми переменными нет строгой линейной зависимости

Интерпретация коэффициентов регрессии

- Утверждение
  - b_i – оценивает влияние x_i на y при неизменности влияния на y остальных переменных
- Для p=2 оценка коэффициента b_1 по МНК

$$b_1 = \frac{Cov(x_1, y)Var(x_2) - Cov(x_2, y)Cov(x_1, x_2)}{Var(x_1)Var(x_2) - [Cov(x_1, x_2)]^2}$$

Доказательство утверждения: см. на доску

Несмещенность

- Случай p=2
- Теорема

$$b_1 = \beta_1 + \frac{1}{\Delta}\left\{Cov(x_1,u)Var(x_2) - Cov(x_2,u)Cov(x_1,x_2)\right\}$$

- где

$$\Delta = Var(x_1)Var(x_2) - \left[Cov(x_1,x_2)\right]^2$$

- Следствие

$$E(b_1) = \beta_1$$

- доказательство

$$E(b_1) = \beta_1 + \frac{1}{\Delta}\left\{Var(x_2)E\left[Cov(x_1,u)\right] - Cov(x_1,x_2)E\left\{Cov(x_2,u)\right\}\right\} = \beta_1$$

Точность

- МНК дает наиболее эффективные линейные оценки (теорема Гаусса-Маркова)
- Факторы, влияющие на точность:
  - ЧИСЛО НАБЛЮДЕНИЙ В ВЫБОРКЕ;
  - ДИСПЕРСИЯ ВЫБОРКИ ОБЪЯСНЯЮЩИХ ПЕРЕМЕННЫХ;
  - ТЕОРЕТИЧЕСКАЯ ДИСПЕРСИЯ СЛУЧАЙНОГО ЧЛЕНА;
  - СВЯЗЬ МЕЖДУ СОБОЙ ОБЪЯСНЯЮЩИХ ПЕРЕМЕННЫХ.
- Доказательство для случая p=2

$$Var(b_i) = \frac{\sigma_u^2}{nVar(x_i)} \cdot \frac{1}{1 - r_{x_1 x_2}^2}$$

Стандартные ошибки коэффициентов регрессии

- «Стандартная ошибка» коэффициента множественной регрессии - оценка стандартного отклонения распределения коэффициента регрессии

Для случая p=2:

$$m_{b_i} = \sqrt{\frac{\sigma_u^2}{nVar(x_i)} \cdot \frac{1}{1-r_{x_1x_2}^2}} =$$

$$= \sqrt{\frac{n/(n-3)Var(e)}{nVar(x_i)} \cdot \frac{1}{1-r_{x_1x_2}^2}} = \sqrt{\frac{Var(e)}{(n-3)Var(x_i)} \cdot \frac{1}{1-r_{x_1x_2}^2}}$$

Мультиколлинеарность

- Мультиколлинеарность – понятие, используемое для описания ситуации, когда нестрогая линейная зависимость приводит к получению ненадежных оценок регрессии
- Замечание 1: если другие факторы благоприятны, то можно получить и хорошие оценки
- Замечание 2: проблема мультиколлинеарности является обычной для временных рядов

Проверка мультиколлинеарности факторов

- Проверяем гипотезу о независимости переменных

H_0: det R=1

Теорема

Величина $$\left\{ n-1-\frac{1}{6}(2m+5)\lg\det R \right\}$$

асимптотически имеет χ^2-распределение с 0,5n(n-1) степенями свободы.

Следствие

если $\chi^2_{факт} > \chi^2_{табл(df,\alpha)}$, то гипотеза H_0 отклоняется

Методы смягчения мультиколлинеарности

- А) Попытки повысить степень выполнения четырех параметров:
  - число наблюдений;
  - выборочные дисперсии объясняющих переменных;
  - дисперсия случайного члена.
- Б) использование внешней информации:
  - теоретические ограничения;
  - внешние эмпирические оценки.

F-тест

- F-статистика

$$F = \frac{D_{факт}}{D_{ост}} = \frac{\sum (\hat{y}_i - \bar{y})^2 / k}{\sum (y - \hat{y}_x)^2 / (n-k-1)} =$$

$$= \frac{R^2}{1-R^2} \cdot \frac{k}{n-k-1}$$

- F–тест оценивает значимость уравнения в целом:

$$\beta_1 = \beta_2 = \ldots = \beta_k = 0$$

- проверяется гипотеза H_0:

Качество оценивания: коэффициент R^2

- R^2 – один из ряда диагностических показателей (причем не самый важный)
- Скорректированный R^2

$$\overline{R}^2 = 1 - (1 - R^2) \cdot \frac{n-1}{n-k-1} = R^2 - \frac{k}{n-k-1}(1 - R^2)$$

Дальнейший анализ дисперсии

- ESS – объясненная сумма квадратов
- RSS – остаточная сумма квадратов
- 2 этапа оценивания:
  - оцениванием регрессию с k независимыми переменными
  - оцениванием регрессию с m>k независимыми переменными
- Гипотеза H_0: дополнительные переменные не увеличивают объяснение регрессией

$$\Downarrow$$

F-статистика:

$$F = \frac{(RSS_k - RSS_m)/(m-k)}{RSS_m/(n-m-1)}$$

Зависимость между F- и t-статистиками

- t-тест обеспечивает проверку **предельного вклада** каждой переменной при допущении, что все другие переменные уже включены в уравнение
- t-тест эквивалентен F-тесту для предельного вклада переменной, которая была отброшена
- **Замечание:** возможна ситуация, когда t-тест для каждой переменной незначим, а F-тест для уравнения в целом значим.
  - Объяснение: если объясняющие способности независимых переменных перекрываются, т.е. имеется мультиколлинеарность.

Поведение R^2 при невключении объясняющей переменной

- Значение R^2 может быть смещено вверх (при положительной корреляции объясняющих переменных) или вниз (при отрицательной корреляции)

Замещающие переменные

- Вместо отсутствующей переменной используем ***заменитель*** (proxy)
- Пример.
- модель $$\ln y = \alpha + \beta_1 \ln x + \beta_2 \ln p + u$$
- y – расходы потребителя на питание
- x – располагаемый личный доход
- p – относительная цена продовольствия
- Пусть lnx имеет явно выраженный временной тренд, тогда время t можно использовать как заменитель x $$\ln y = a + b_2 \ln p + b_3 t$$

Результаты моделирования

Объясняющая переменная	Оценки коэффициентов			R^2
	b_1	b_2	b_3	
lnx, lnp	0,64 (0,03)	-0,48 (0,12)		0,99
lnp		2,04 (0,33)		0,63
lnp, t		-0,47 (0,13)	0,023 (0,001)	0,98

Непреднамеренное использование замещающих переменных

- Если корреляция между z и x незначительна, то результаты будут плохими
- Если корреляция между z и x тесная, то результаты будут удовлетворительными
- Если цель регрессии – предсказание значений y, то использование замещающих переменных целесообразно
- Если цель регрессии – научное любопытство, то использование замещающих переменных обычно нецелесообразно
- Если хотим использовать объясняющую переменную как инструмент экономической политики, то последствия использования замещающей переменной могут быть

Анализ остатков

- Взгляд пессимиста:
  - свидетельство неудачи
- Взгляд оптимиста:
  - источник новых идей
  - основа для постановки новых задач
  - конструктивная критика

Пример: продажа предметов длительного

ЛАГОВЫЕ ПЕРЕМЕННЫЕ

- *лаговые переменные – это экзогенные или эндогенные переменные, которые относятся к предыдущим моментам времени и находятся в эконометрической модели одновременно с переменными, относящимися к текущему моменту времени.*

  - *Например, xt-1 – это лаговая экзогенная переменная, а yt-1 – это лаговая эндогенная переменная*