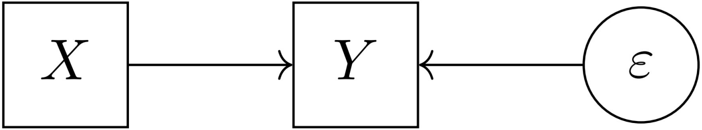
Занятие 2. Линейная регрессия: основы

30 января 2024

Путевая диаграмма: регрессия



Y – зависимая переменная (отклик);
X – независимая переменная (объясняющая переменная / предиктор);
ε – ошибка

Классическая линейная регрессия

Вопрос

Запишем спецификацию парной регрессии в общем виде.

Классическая линейная регрессия

Вопрос

Запишем спецификацию парной регрессии в общем виде.

Ответ

$y_i = b_0 + b_1 x_i + e_i$,
где y_i – зависимая переменная (отклик),
b_0 – константа (intercept),
b_1 – коэффициент при предикторе (slope coefficient),
x_i – независимая переменная (предиктор),
e_i – ошибка.

Классическая линейная регрессия

Вопрос

Запишем спецификацию парной регрессии в общем виде.

Ответ

$y_i = b_0 + b_1 x_i + e_i$,
где y_i – зависимая переменная (отклик),
b_0 – константа (intercept),
b_1 – коэффициент при предикторе (slope coefficient),
x_i – независимая переменная (предиктор),
e_i – ошибка.

$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$ – это предсказанное значение зависимой переменной;
$\hat{e}_i = y_i - \hat{y}_i$, где $\hat{e}_i$ – это остаток (оценка ошибки).

Классическая линейная регрессия

Вопрос

Метод наименьших квадратов (МНК) – один из методов оценивания параметров в регрессии. Покажем основной принцип этого метода.

Классическая линейная регрессия

Вопрос

Метод наименьших квадратов (МНК) – один из методов оценивания параметров в регрессии. Покажем основной принцип этого метода.

Ответ

В соответствии с МНК выбираем такие оценки коэффициентов, при которых линия предсказания наиболее близка к наблюдениям. Математически происходит минимизация суммы квадратов остатков:

$$\min \sum_{i=1}^{n}(y_i - (\hat{b}_0 + \hat{b}_1 x_i))^2.$$

Классическая линейная регрессия

Вопрос

Метод наименьших квадратов (МНК) – один из методов оценивания параметров в регрессии. Покажем основной принцип этого метода.

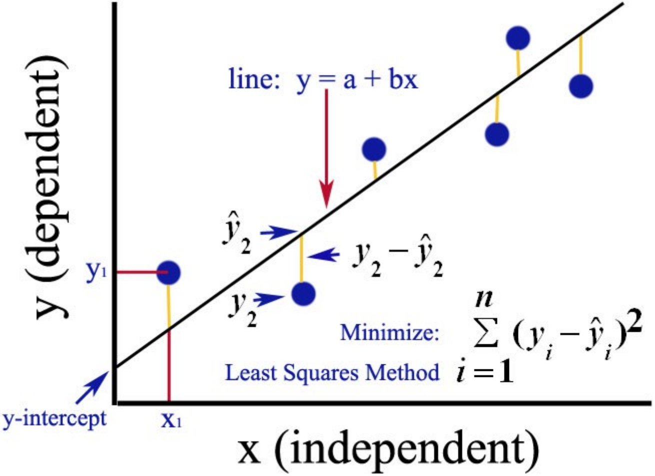
Ответ

В соответствии с МНК выбираем такие оценки коэффициентов, при которых линия предсказания наиболее близка к наблюдениям. Математически происходит минимизация суммы квадратов остатков:

$$\min \sum_{i=1}^{n} (y_i - (\hat{b}_0 + \hat{b}_1 x_i))^2.$$

Или можем переписать это в таком виде: $\min \sum^{n} (y_i - \hat{y}_i)^2$

Иллюстрация принципа МНК



Источник картинки: ссылка

Выведение МНК-оценки в модели на константу

Запишем исходную спецификацию:

$$y_i = \beta_0 + \varepsilon_i$$

Выведение МНК-оценки в модели на константу

Запишем исходную спецификацию:

$$y_i = \beta_0 + \varepsilon_i$$

Перепишем в терминах модельных (предсказанных) значений, то есть, отклик (зависимая переменная) в среднем равна константе (некоторому постоянному значению):

$$\hat{y}_i = \hat{\beta}_0$$

Выведение МНК-оценки в модели на константу

Руководствуясь принципом МНК, минимизируем сумму квадратов остатков:

$$\frac{\partial \sum_{i=1}^{n}(y_i - \hat{\beta}_0)^2}{\partial \hat{\beta}_0} = 0$$

Выведение МНК-оценки в модели на константу

Руководствуясь принципом МНК, минимизируем сумму квадратов остатков:

$$\frac{\partial \sum_{i=1}^{n} (y_i - \hat{\beta}_0)^2}{\partial \hat{\beta}_0} = 0$$

$$\frac{\partial \sum_{i=1}^{n} (y_i^2 - 2y_i\hat{\beta}_0 + \hat{\beta}_0^2)}{\partial \hat{\beta}_0} = 0$$

Выведение МНК-оценки в модели на константу

Руководствуясь принципом МНК, минимизируем сумму квадратов остатков:

$$\frac{\partial \sum_{i=1}^{n}(y_i - \hat{\beta}_0)^2}{\partial \hat{\beta}_0} = 0$$

$$\frac{\partial \sum_{i=1}^{n}(y_i^2 - 2y_i\hat{\beta}_0 + \hat{\beta}_0^2)}{\partial \hat{\beta}_0} = 0$$

$$\frac{\partial (\sum_{i=1}^{n} y_i^2 - 2\hat{\beta}_0 \sum_{i=1}^{n} y_i + n\hat{\beta}_0^2)}{\partial \hat{\beta}_0} = 0$$

Выведение МНК-оценки в модели на константу

Руководствуясь принципом МНК, минимизируем сумму квадратов остатков:

$$\frac{\partial \sum_{i=1}^{n}(y_i - \hat{\beta}_0)^2}{\partial \hat{\beta}_0} = 0$$

$$\frac{\partial \sum_{i=1}^{n}(y_i^2 - 2y_i\hat{\beta}_0 + \hat{\beta}_0^2)}{\partial \hat{\beta}_0} = 0$$

$$\frac{\partial (\sum_{i=1}^{n} y_i^2 - 2\hat{\beta}_0 \sum_{i=1}^{n} y_i + n\hat{\beta}_0^2)}{\partial \hat{\beta}_0} = 0$$

$$-2\sum_{i=1}^{n} y_i + 2n\hat{\beta}_0 = 0$$

Выведение МНК-оценки в модели на константу

Руководствуясь принципом МНК, минимизируем сумму квадратов остатков:

$$\frac{\partial \sum_{i=1}^{n}(y_i - \hat{\beta}_0)^2}{\partial \hat{\beta}_0} = 0$$

$$\frac{\partial \sum_{i=1}^{n}(y_i^2 - 2y_i\hat{\beta}_0 + \hat{\beta}_0^2)}{\partial \hat{\beta}_0} = 0$$

$$\frac{\partial (\sum_{i=1}^{n} y_i^2 - 2\hat{\beta}_0 \sum_{i=1}^{n} y_i + n\hat{\beta}_0^2)}{\partial \hat{\beta}_0} = 0$$

$$-2\sum_{i=1}^{n} y_i + 2n\hat{\beta}_0 = 0 => \hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}$$

Выведение МНК-оценок в парной регрессии

Найдем оптимальную оценку константы ($\hat{\beta}_0$) в парной линейной регрессии, при которой сумма квадратов остатков будет минимальна.
Рассмотрим частную производную по $\hat{\beta}_0$:

$$\frac{\partial \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_0} = 0$$

Выведение МНК-оценок в парной регрессии

Найдем оптимальную оценку константы ($\hat{\beta}_0$) в парной линейной регрессии, при которой сумма квадратов остатков будет минимальна.
Рассмотрим частную производную по $\hat{\beta}_0$:

$$\frac{\partial \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_0} = 0$$

$$(-2)\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

Выведение МНК-оценок в парной регрессии

Найдем оптимальную оценку константы ($\hat{\beta}_0$) в парной линейной регрессии, при которой сумма квадратов остатков будет минимальна.
Рассмотрим частную производную по $\hat{\beta}_0$:

$$\frac{\partial \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_0} = 0$$

$$(-2)\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

Выведение МНК-оценок в парной регрессии

Найдем оптимальную оценку константы ($\hat{\beta}_0$) в парной линейной регрессии, при которой сумма квадратов остатков будет минимальна.
Рассмотрим частную производную по $\hat{\beta}_0$:

$$\frac{\partial \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_0} = 0$$

$$(-2) \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$

Выведение МНК-оценок в парной регрессии

Рассмотрим частную производную по $\hat{\beta}_1$:

$$\frac{\partial \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_1} = 0$$

Выведение МНК-оценок в парной регрессии

Рассмотрим частную производную по $\hat{\beta}_1$:

$$\frac{\partial \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_1} = 0$$

$$(-2)\sum_{i=1}^{n}(x_i)(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

Выведение МНК-оценок в парной регрессии

Рассмотрим частную производную по $\hat{\beta}_1$:

$$\frac{\partial \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_1} = 0$$

$$(-2)\sum_{i=1}^{n}(x_i)(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \hat{\beta}_0 x_i - \sum_{i=1}^{n} \hat{\beta}_1 x_i^2 = 0$$

Выведение МНК-оценок в парной регрессии

Рассмотрим частную производную по $\hat{\beta}_1$:

$$\frac{\partial \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_1} = 0$$

$$(-2)\sum_{i=1}^{n}(x_i)(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \hat{\beta}_0 x_i - \sum_{i=1}^{n} \hat{\beta}_1 x_i^2 = 0$$

Выведение МНК-оценок в парной регрессии

Вспомним, что ранее мы уже получили оценку константы, подставим ее в уравнение:

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} (\bar{y} - \hat{\beta}_1 \bar{x}) x_i - \sum_{i=1}^{n} \hat{\beta}_1 x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \bar{y} x_i + \sum_{i=1}^{n} \hat{\beta}_1 \bar{x} x_i - \sum_{i=1}^{n} \hat{\beta}_1 x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \bar{y} + \hat{\beta}_1 \sum_{i=1}^{n} x_i \bar{x} - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\widehat{Cov}(x, y)}{\widehat{Var}(x)}$$

Оценки в соответствии с МНК

Модель на константу

$$y_i = \beta_0 + e_i$$

$$\hat{\beta}_0 = \bar{y}$$

Оценки в соответствии с МНК

Модель на константу

$$y_i = \beta_0 + e_i$$

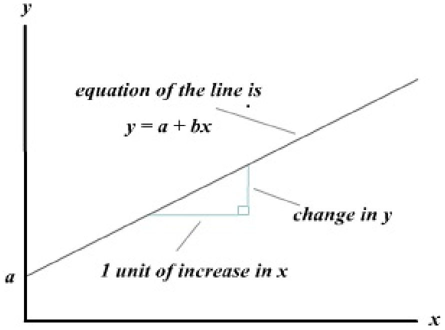
$$\hat{\beta}_0 = \bar{y}$$

Модель парной регрессии

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

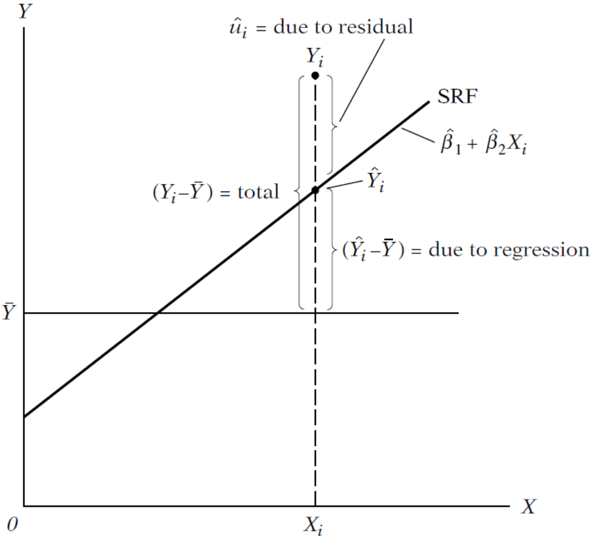
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\widehat{Cov}(x, y)}{\widehat{Var}(x)}$$

Интерпретация оценок коэффициентов



$\hat{b}_0$ (также обозначается как a) – среднее значение отклика при условии равенства предикторов 0.
$\hat{b}_1$ – на сколько в среднем изменяется отклик при увеличении предиктора на единицу измерения при прочих равных.

Разложение вариации зависимой переменной



Разложение вариации зависимой переменной

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$TSS = ESS + RSS$$

Разложение вариации зависимой переменной

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$TSS = ESS + RSS$$

$$\frac{TSS}{TSS} = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

(коэффициент детерминации)

Задачка для практики

Рассчитайте коэффициент детерминации

Построена регрессия индекса потребительских цен на уровень безработицы на основе данных 50 стран. Несмещенная выборочная оценка дисперсии индекса потребительских цен равна 800, а сумма квадратов остатков регрессии равна 25000.

Задачка для практики

Рассчитайте коэффициент детерминации

Построена регрессия индекса потребительских цен на уровень безработицы на основе данных 50 стран. Несмещенная выборочная оценка дисперсии индекса потребительских цен равна 800, а сумма квадратов остатков регрессии равна 25000.

Решение

$$\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = 800$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 49 \times 800 = 39200$$

$$R^2 = 1 - \frac{25000}{39200} \approx 0.36$$

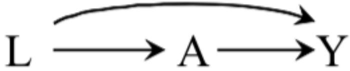
Спецификация

Объясняющие переменные

В соответствии с гипотезами включаем набор ключевых предикторов. Второй тип объясняющих переменных – контрольные переменные. Включаются для уменьшения omitted variable bias.

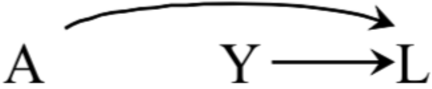
$y_i = b_0 + b_1 x_{1i} + ... + b_k x_{ki} + e_i$,
где y_i – зависимая переменная (отклик),
b_0 – константа (intercept),
$b_1, ..., b_k$ – коэффициенты при соответствующих предикторах (slope coefficients),
$x_{1i}, ... x_{ki}$ – предикторы,
e_i – ошибка.

Подходящая контрольная переменная – переменная-confounder



Переменная L влияет и на A, и на Y. НУжно включить L как контрольную.

Переменную-collider включать как контрольную нельзя



На переменную L влияет A и Y. Включение L в качестве контрольной привнесет только лишнее смещение в оценку коэффициента при A.

Условия для получения идентифицируемой модели линейной регрессии и BLUE-оценок

Для того, чтобы модель была идентифицируемая,

1. наблюдений должно быть больше, чем количество оцениваемых параметров
2. не должно быть строгой мультиколлинеарности (то есть, нет линейно зависимых предикторов)

Для получения BLUE-оценок (то есть, наиболее эффективных среди класса всех линейных несмещенных оценок) ошибки в модели должны удовлетворять ряду свойств:

- $Cov(e_i, x) = 0$ – экзогенность
- $Var(e_i|x) = const$ – гомоскедастичность
- $Cov(e_i, e_j|x) = 0$ – отсутствие автокорреляции

Мультиколлинеарность

Виды

1. Строгая (линейная зависимость предикторов)
2. Сильная (предикторы довольно сильно связаны, негативно отражающаяся на оценках в модели)
3. Слабая (слабая связь между предикторами допустима)

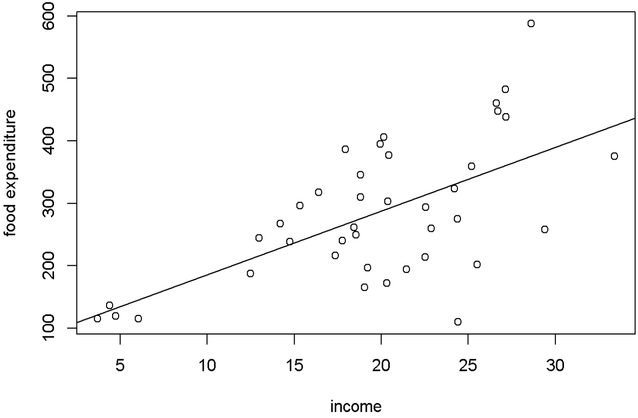
Последствия мультиколлинеарности

1. В случае строгой мультиколлинеарности невозможность получить оценки
2. В случае сильной мультиколлинеарности незначимые оценки при высоком R^2
3. В случае сильной мультиколлинеарности неустойчивые результаты

Диагностика мультиколлинеарности

1. корреляционная матрица
2. визуализация
3. VIF (Variance Inflation Factor)
   Переоцениваем набор вспомогательных регрессий. К примеру, одна из таких моделей:
   $x_{1i} = a_0 + a_1 x_{2i} + ... a_{k-1} x_{k-1i} + u_i$
   Рассчитываем R^2 из вспомогательной модели и считаем $VIF_j = \frac{1}{1 - R^2}$. Показатель более 10 сигнализирует о довольно сильной мультиколлинеарности.

Иллюстрация гетероскедастичности



Why should we care?

Последствия гетероскедастичности

1. неэффективность оценок, при этом остаются состоятельными и несмещенными
2. распределение статистик уже другое

Why should we care?

Последствия гетероскедастичности

1. неэффективность оценок, при этом остаются состоятельными и несмещенными
2. распределение статистик уже другое

Итог: главная проблема

Эти последствия делают проверку гипотез о незначимости коэффициентов проблематичной.

Откуда берется гетероскедастичность?

Источники гетероскедастичности

Откуда берется гетероскедастичность?

Источники гетероскедастичности

1. работаем с объектами разного «размера»
2. нетипичные наблюдения
3. неверно определена функциональная форма взаимосвязи
4. пропущены важные факторы
5. разные методики сбора данных

Как выявить гетероскедастичность?

Диагностики

Как выявить гетероскедастичность?

Диагностики

1. еще до диагностик важно обратиться к Вашим теоретическим предпосылкам, они и будут самыми важными для того, чтобы принять решение о том, как работать далее с оценками модели
2. визуализация
3. формальные тесты

Диагностики, основанные на визуализации

Графики

- OY – зависимая переменная, OX – предиктор
- OY – зависимая переменная, OX – предсказанное значение ($\hat{y}$)
- OY – остатки в квадрате, OX – предиктор
- OY – остатки в квадрате, OX – предсказанное значение ($\hat{y}$)

Изменяется ли вариация при разных значениях X?

Диагностики: тест Уайта

Предпосылки

- большая выборка
- отсутствуют требования о нормальности распределения ошибок

Шаги реализации:

- оцениваем модель и сохраняем остатки $(\hat{e})$
- строим дополнительную модель остатков в квадрате (в качестве зависимой переменной) на все исходные предикторы, их квадраты и попарные произведения
- сохраняем из дополнительной модели R^2
- считаем статистику критерия: $nR^2 \sim \chi^2_{k-1}$, где k – количество параметров в дополнительной модели