«Напоминалка»

Данные. Типы данных для анализа:
Пространственные – cross sectional data – гипотеза: одинаковые условия ля всех объектов наблюдения
Временны’е ряды – time-series data – объект один, наблюдаем как изменяется с течением времени
«Панели» - panel data – объекты разные, наблюдаем состояния множества объектов в разные моменты времени (некая комбинация)

Чаще всего students помнят, что коэффициенты уравнения регрессии оцениваются МНК.
Однако, любой специалист скажет, что строго говоря применять МНК для не «пространственных» данных никак не возможно… Но не только для не пространственных…

О чем мы?

Модель парной линейной регрессии: терминология

Модель парной линейной регрессии имеет вид:

$Y_i = \beta_0 + \beta_1 X_i + u_i$,

где индекс i пробегает по всем наблюдениям, $i = 1,...,n$;

Y_i – *зависимая переменная*, регрессируемая переменная или просто *переменная слева*;

X_i – *независимая переменная*, *объясняющая переменная*, *регрессор* или просто *переменная справа*;

$\beta_0 + \beta_1 X_i$ – *линия теоретической регрессии*, *линия регрессии генеральной совокупности*, или *функция регрессии генеральной совокупности*;

β_0 – *константа* линии теоретической регрессии;

β_1 – *угловой коэффициент* линии теоретической регрессии; и

u_i – *ошибка*.

Предположения модели:

Предположения метода наименьших квадратов

$Y_i = \beta_0 + \beta_1 X_i + u_i$, $i = 1,...,n$,

где

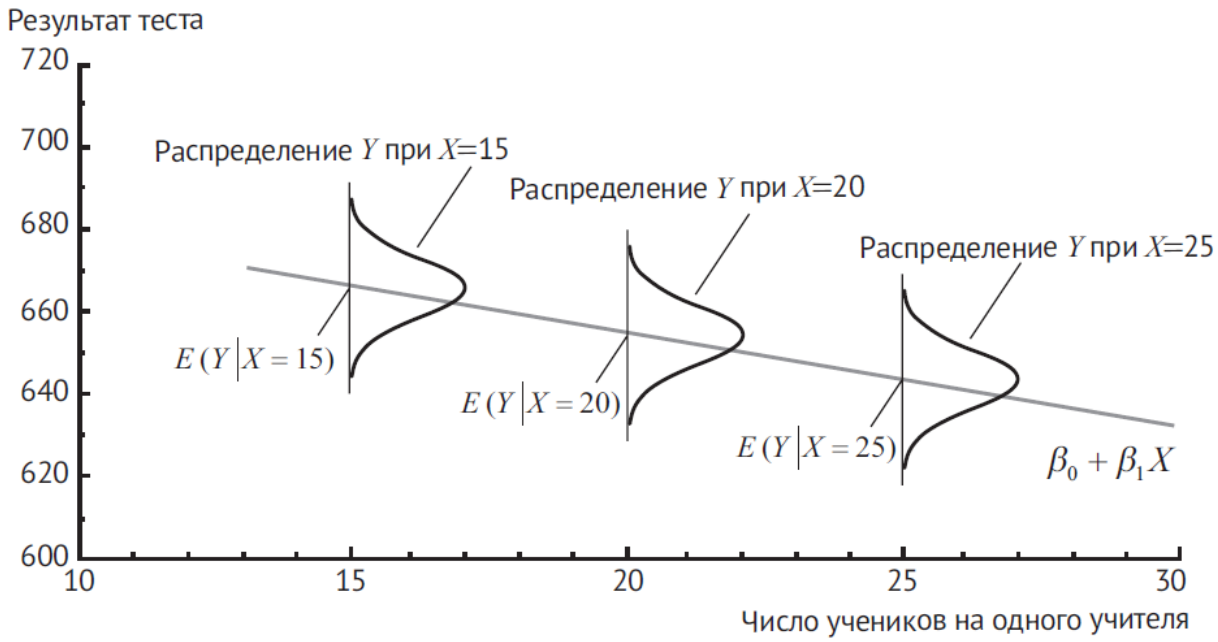
1. Ошибка имеет нулевое условное среднее при заданном X_i: $E(u_i \mid X_i) = 0$.
2. (X_i, Y_i), $i = 1,...,n$ — независимые и одинаково распределенные (i.i.d.), извлеченные из их совместного распределения.
3. Большие выбросы маловероятны: X_i и Y_i имеют ненулевой конечный четвертый момент.

Что это значит?

Предположение № 1: условное распределение u_i относительно X_i имеет нулевое среднее

Первое из трех *предположений метода наименьших квадратов* заключается в том, что условное распределение u_i относительно X_i имеет нулевое среднее. Это предположение является формальным математическим утверждением о других «факторах», содержащихся в u_i, и утверждает, что эти другие факторы не связаны с X_i в том смысле что при заданном значении X_i среднее значение распределения этих других факторов равно нулю.

Ус



Условное распределение вероятности и линия регрессии

Второе предположение метода наименьших квадратов заключается в том, что (X_i, Y_i), $i = 1,\ldots,n$ независимы и одинаково распределены (i.i.d.). Как обсуждалось в разделе 2.5 (см. вставку «Основные понятия 2.5»), это предположение является утверждением о способе формирования выборки. Если наблюдения отобраны простым случайным образом из единственной большой генеральной совокупности, тогда (X_i, Y_i), $i = 1,\ldots,n$ являются i.i.d. Например, пусть X — это возраст работника, а Y — его или ее зарплата, и представим, что мы выбираем человека случайным образом из генеральной совокупности всех работников.

Предположение об i.i.d. является разумным для многих схем сбора данных. Например, данные обследования населения со случайно выбранным подмножеством всей генеральной совокупности (всего населения) обычно можно рассматривать как i.i.d.

Не все методы выбор наблюдений удовлетворяют данному предположению!

В частности – ситуация, когда наблюдения являются временным рядом. А это значит, что для временных рядов разработаны свои методы.

Временные ряды представляют собой данные, собранные для одного показателя в различные моменты времени; они могут использоваться, чтобы ответить на количественные вопросы, для которых межобъекные выборки не являются адекватными. Одним из таких вопросов может быть вопрос о том, какое влияние на интересующую нас переменную Y оказывают изменения во времени, происходящие с другой переменной X? Иными словами, какой динамический причинный эффект оказывают на Y изменения в X?