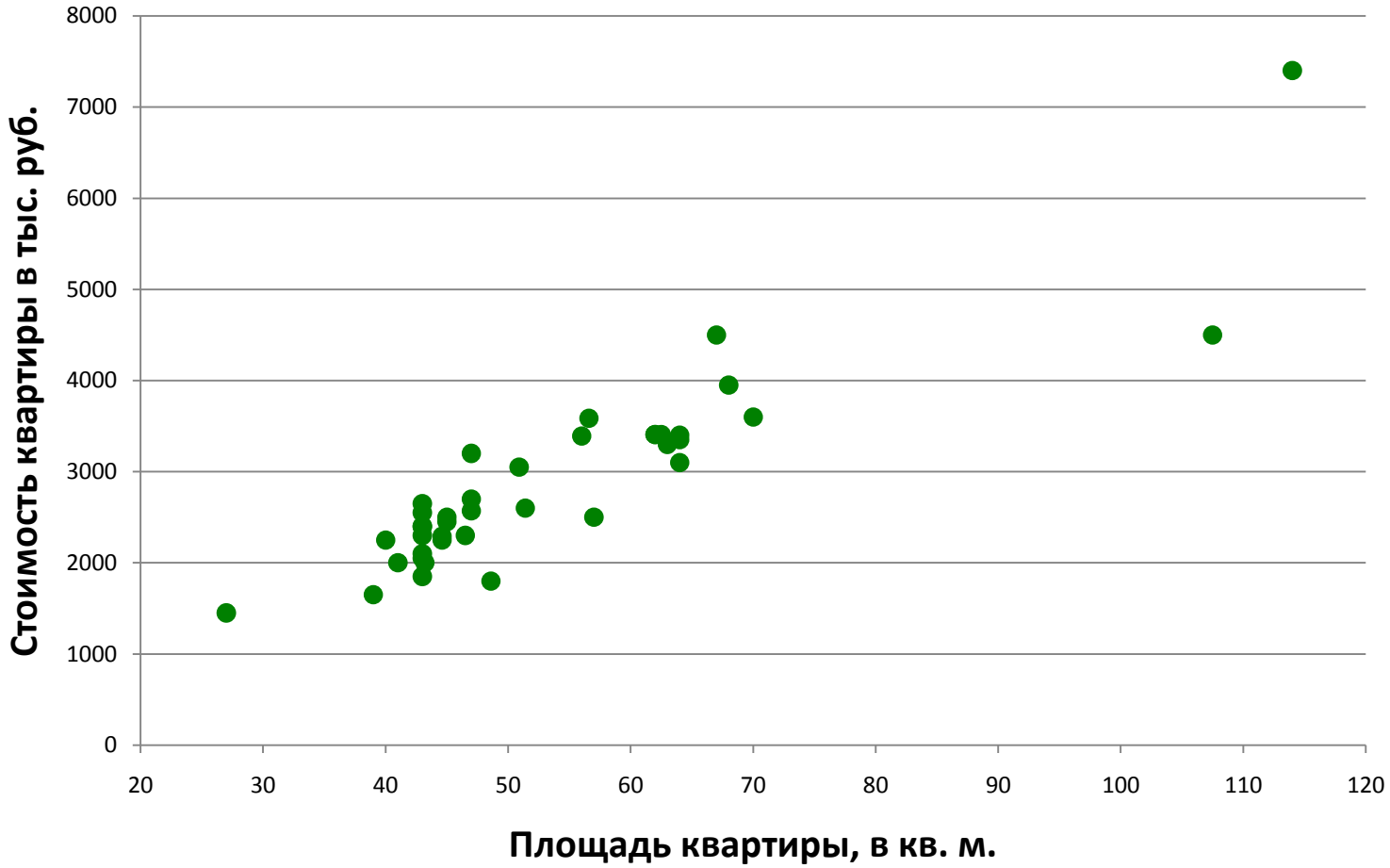
Эконометрическое моделирование

Лекция № 1

Преподаватель – Аристова Елена Владимировна

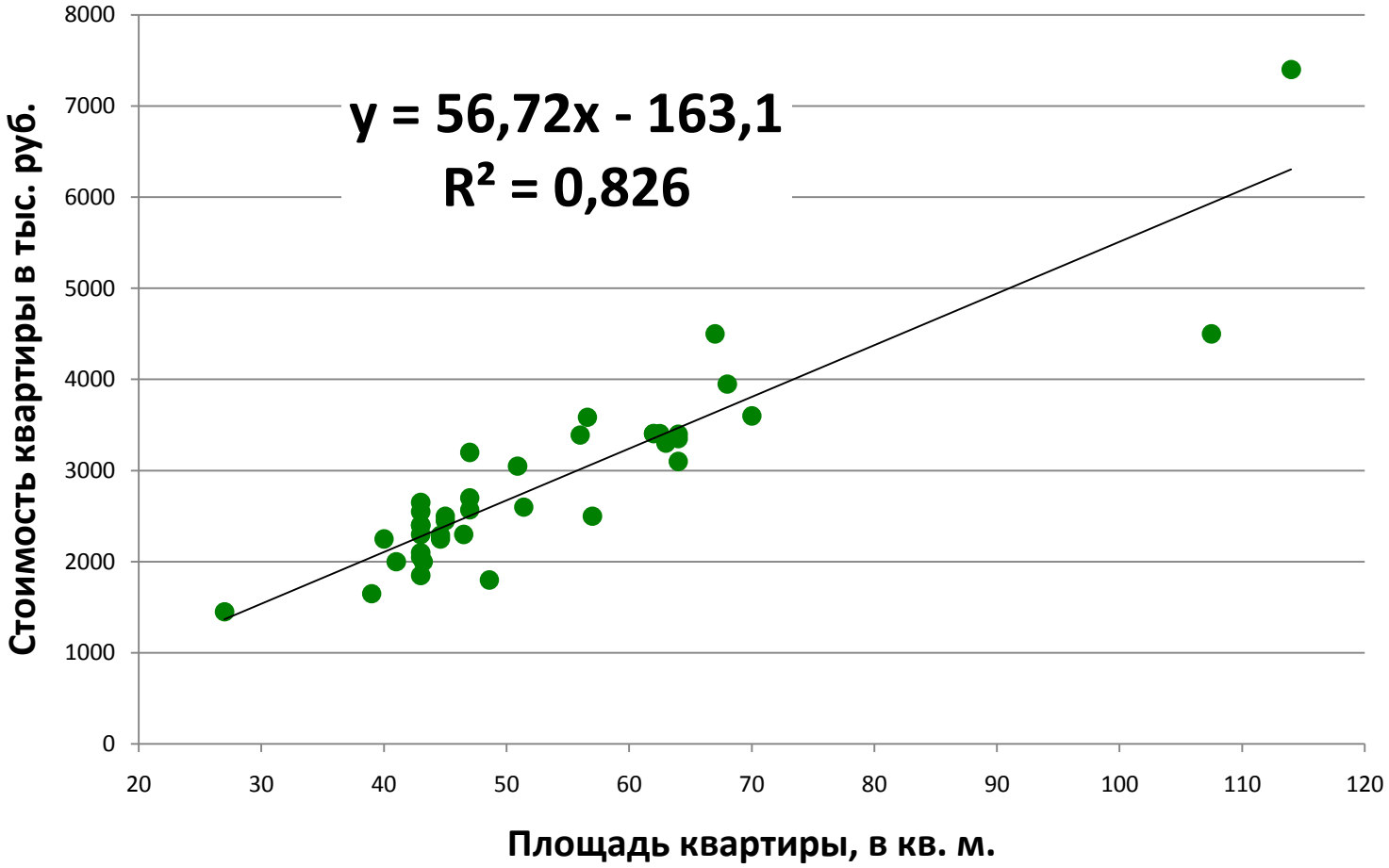
Как связана площадь квартиры и ее цена?

Данные по 2-х комнатным квартирам в кировском р-не г. Томск (2014)



Как связана площадь квартиры и ее цена?

Данные по 2-х комнатным квартирам в кировском р-не г. Томск (2014)



Если польза от подготовительных курсов?

Результаты вступительного экзамена по экономике для абитуриентов магистратуры

	Ходили на курсы	Не ходили на курсы
Средний балл за экзамен	43 балла (50 человек)	48,5 баллов (100 человек)

Негативный эффект от посещения курсов?

Если польза от подготовительных курсов?

С учетом еще одного фактора

	Ходили на курсы	Не ходили на курсы
Выпускники экономического факультета	55 баллов (10 человек)	50 баллов (95 человек)
Другие абитуриенты	40 баллов (40 человек)	20 баллов (5 человек)

Если польза от подготовительных курсов?

	Ходили на курсы	Не ходили на курсы
Выпускники экономического факультета	55 баллов (10 человек)	50 баллов (95 человек)
Другие абитуриенты	40 баллов (40 человек)	20 баллов (5 человек)
ИТОГО	43 балла (50 человек)	48,5 баллов (100 человек)

Если польза от подготовительных курсов?

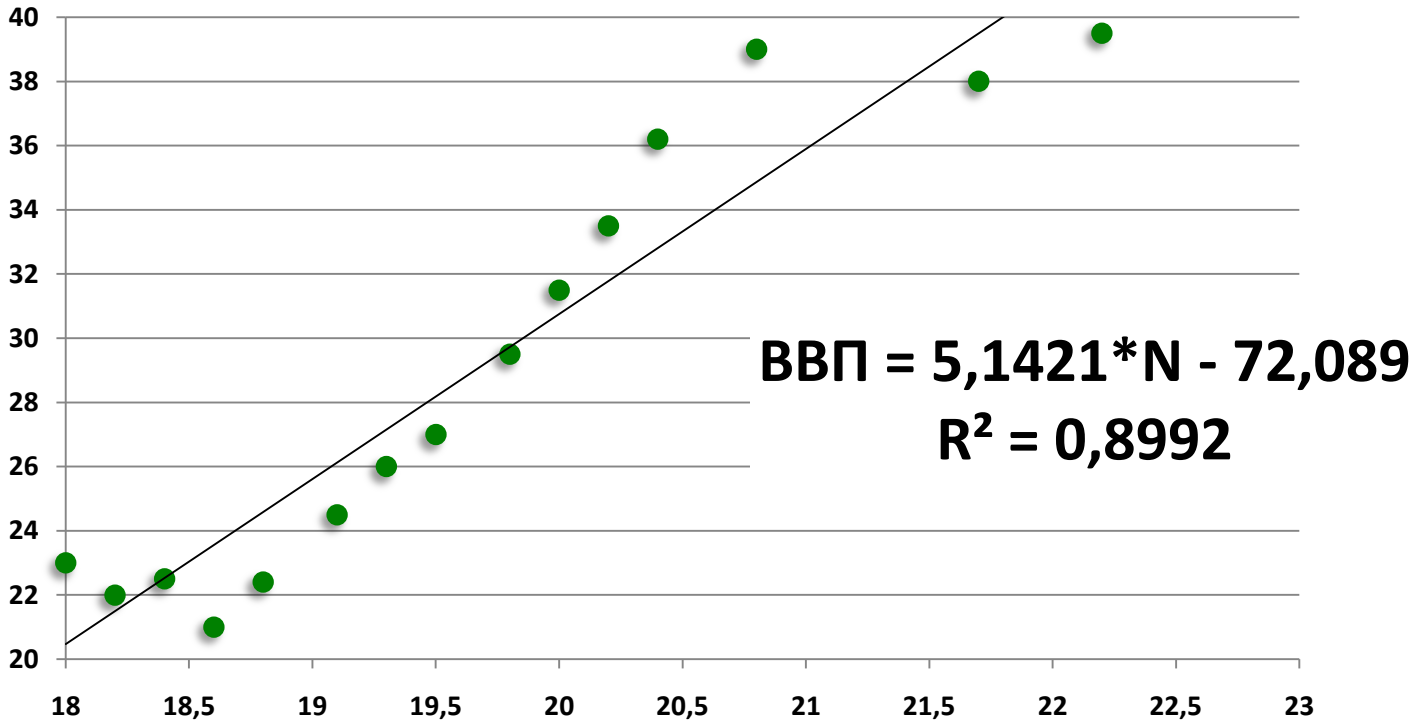
Вывод: если игнорировать существенные переменные, то будут получены смещенные результаты

Ложная регрессия при анализе временных рядов

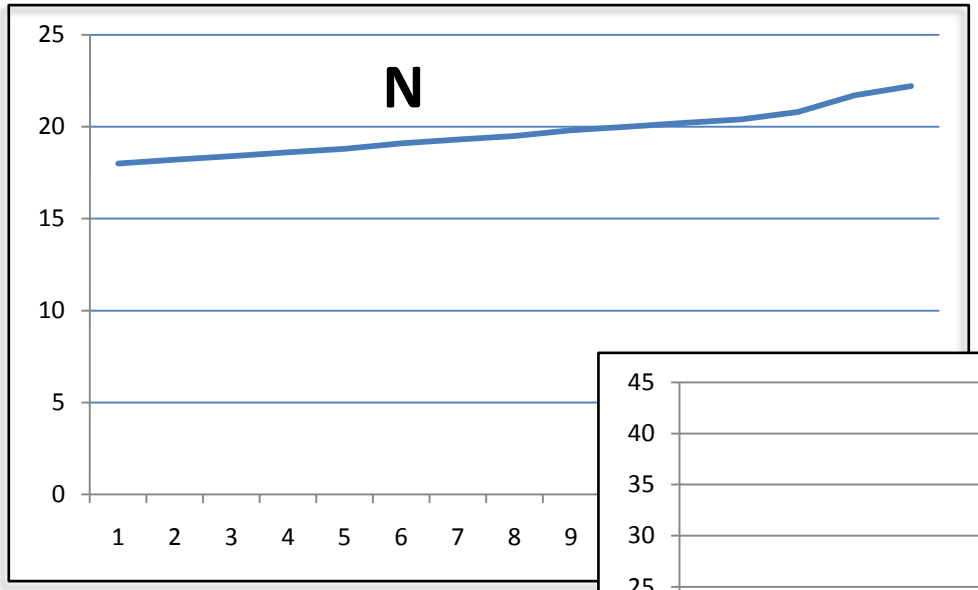
Что определяет темпы экономического роста?

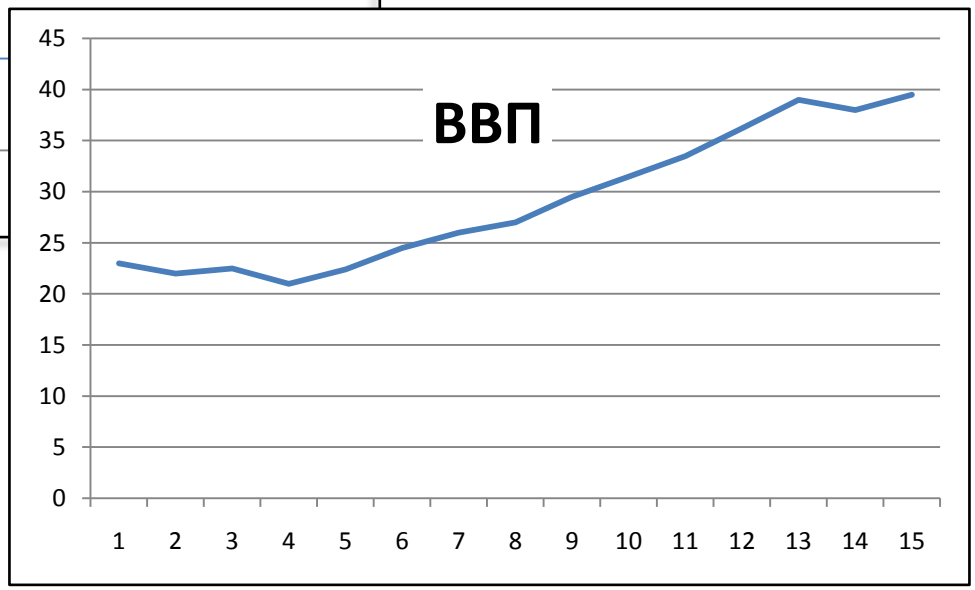
Ложная регрессия при анализе временных рядов

Зависимость ВВП России от некоторого фактора в 1995–2012 гг. (реальный ВВП в ценах 2008 г., трлн руб.)



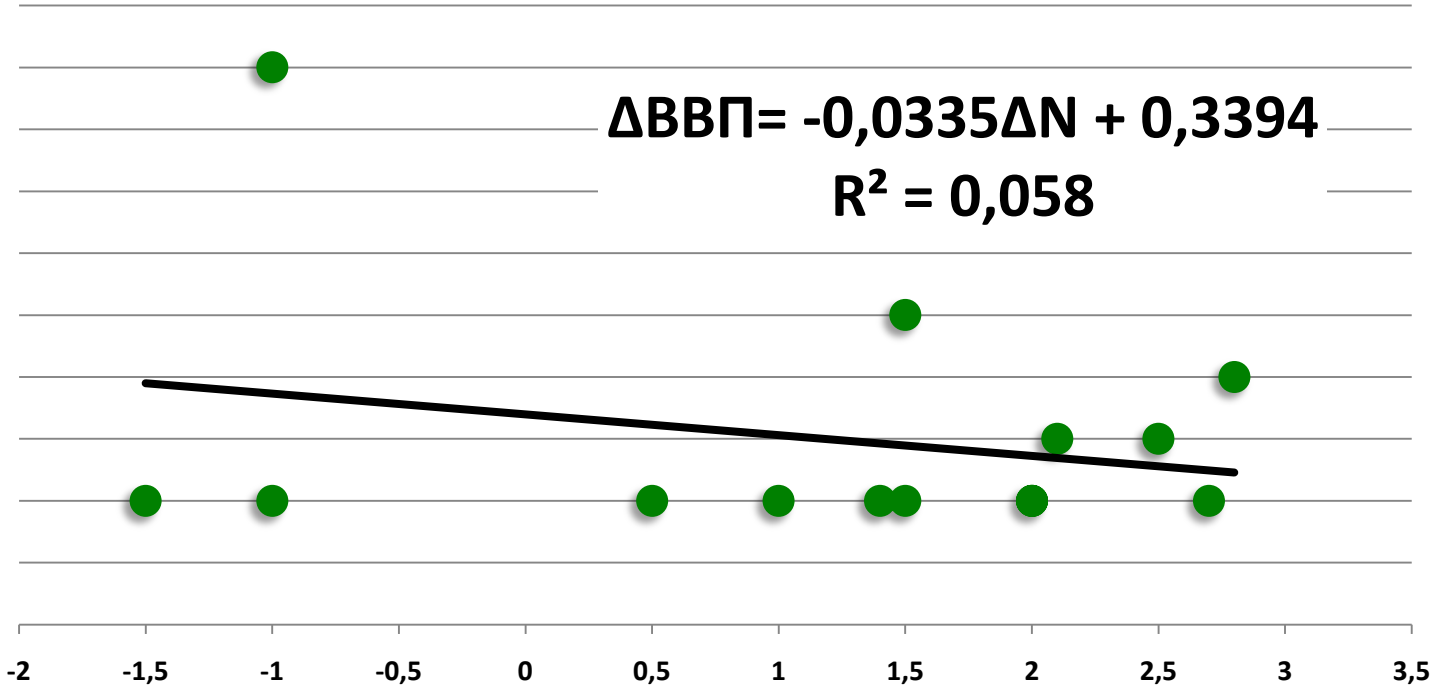
Ложная регрессия при анализе временных рядов





Ложная регрессия при анализе временных рядов

Если оценивать регрессию для разностей, то связь становится незначимой:



Ложная регрессия при анализе временных рядов

Вывод: **если игнорировать свойства временных рядов (например, наличие трендов), будут получены искаженные результаты**

Этапы эконометрического исследования

1. Постановка проблемы (качественный анализ связей экономических переменных – выделение зависимых и независимых переменных);
2. Получение данных, анализ их качества;
3. Спецификация модели (форма связи между переменными);
4. Оценка параметров модели;
5. Интерпретация результатов.

Типы данных

Пространственные данные

Совокупность данных по различным объектам в определенный момент времени

Временные ряды

Данные по одному объекту за ряд последовательных моментов (промежутков) времени

Панельные данные

Сочетают в себе как данные пространственного типа, так и данные типа временных рядов

Типы шкал, по которым производятся измерения в эконометрике

Шкала наименований (номинальная)

Измерением является любая классификация, по которой класс получает числовое наименование

Порядковая шкала (ранговая, бальная)

Шкала, в которой порядок элементов по уровню проявления некоторого свойства существенен, а количественное выражение различия несущественно или плохо осуществимо

Типы шкал, по которым производятся измерения в эконометрике

Интервальная шкала

Шкала разностей

Позволяет рассчитать разность между объектами по определенному свойству
Пример: указание года рождения

Шкала отношений

Шкала на которой можно сделать вывод о том, что одна величина в несколько раз больше другой
Пример: измерения экономических параметров – цена, себестоимость и др.

Пример: определите типы шкал для параметров, описывающих квартиры

Наименование	Тип данных	
Район города	текст	Номинальная
Тип дома (кирпичный, панельный, деревянный)	текст	Номинальная
Состояние квартиры (отличное, хорошее, нормальное)	текст	Порядковая
Количество комнат	число	Интервальная
Площадь квартиры	число	Интервальная
Наличие балкона	текст	Номинальная
Цена	число	Интервальная

Меры изменчивости и связи двух переменных

Среднее значение

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Меры изменчивости и связи двух переменных

Дисперсия (выборочная дисперсия)

$$Var(x) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \overline{x^2} - \bar{x}^2$$

характеризует степень разброса значений вокруг своего среднего

Стандартное отклонение

$$Std.Dev.(x) = \sqrt{Var(x)}$$

Позволяет измерять степень разброса переменных в тех же единицах, к которых измеряется сама переменная

Меры изменчивости и связи двух переменных

Выборочная ковариация

$$Cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \overline{xy} - \bar{x}\cdot\bar{y}$$

Если Cov(x,y) > 0, то между переменными x и y существует положительная связь и наоборот

Коэффициент ковариации может принимать сколь угодно малые и большие значения, так как зависит от единиц измерения

Меры изменчивости и связи двух переменных

Выборочный коэффициент корреляции

$$Corr(x, y) = \frac{Cov(x, y)}{\sqrt{Var(x) \cdot Var(y)}}$$

Характеристика линейной связи между переменными.

Может принимать значения в пределах от -1 до 1.

Если между переменными нет связи, то коэффициент корреляции близок к нулю.

Парная регрессия

Пусть имеется информация о двух переменных: x и y. Выборка из n наблюдений переменной x (x_1, x_2, … x_n) и n наблюдений переменной y (y_1, y_2, … y_n).

Мы предполагаем, что переменные x и y взаимосвязаны и хотим проверить это предположение.

Парный регрессионный анализ заключается в исследовании зависимости между парой переменных.

Линейная парная регрессия

Линейная модель наблюдений

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \qquad i = 1, \ldots n.$$

Линейная модель связи

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$$

Ошибка (отклонение)

$$\varepsilon_i = y_i - (\hat{\beta}_1 + \hat{\beta}_2 x_i) = y_i - \hat{y}_i$$

Причины присутствия случайных ошибок

Ошибки спецификации – неправильный выбор функциональной зависимости между переменными или недоучет в уравнении какого-либо существенного фактора

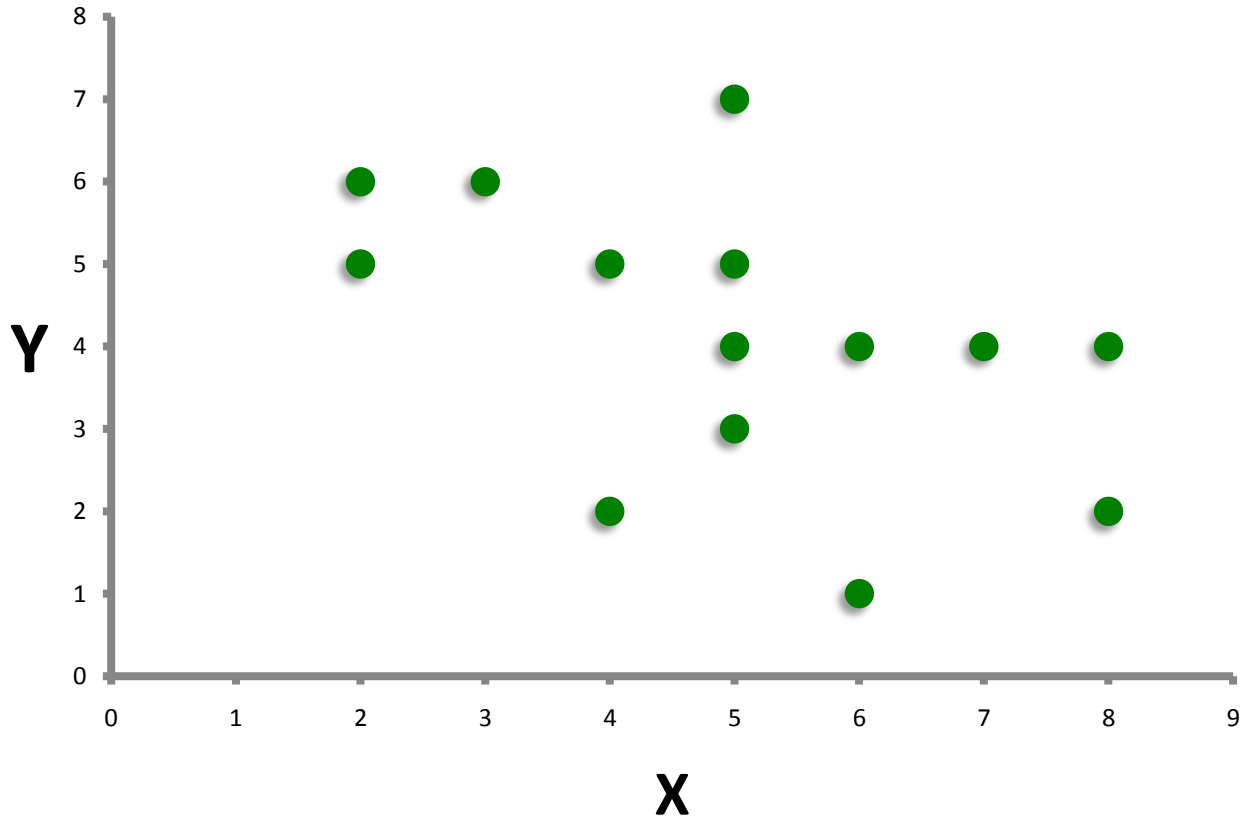
Ошибки выборки – возникают в силу неоднородности данных в исходной статистической совокупности

Ошибки измерения – ошибки возникающие в процессе измерения параметров, описывающих данные

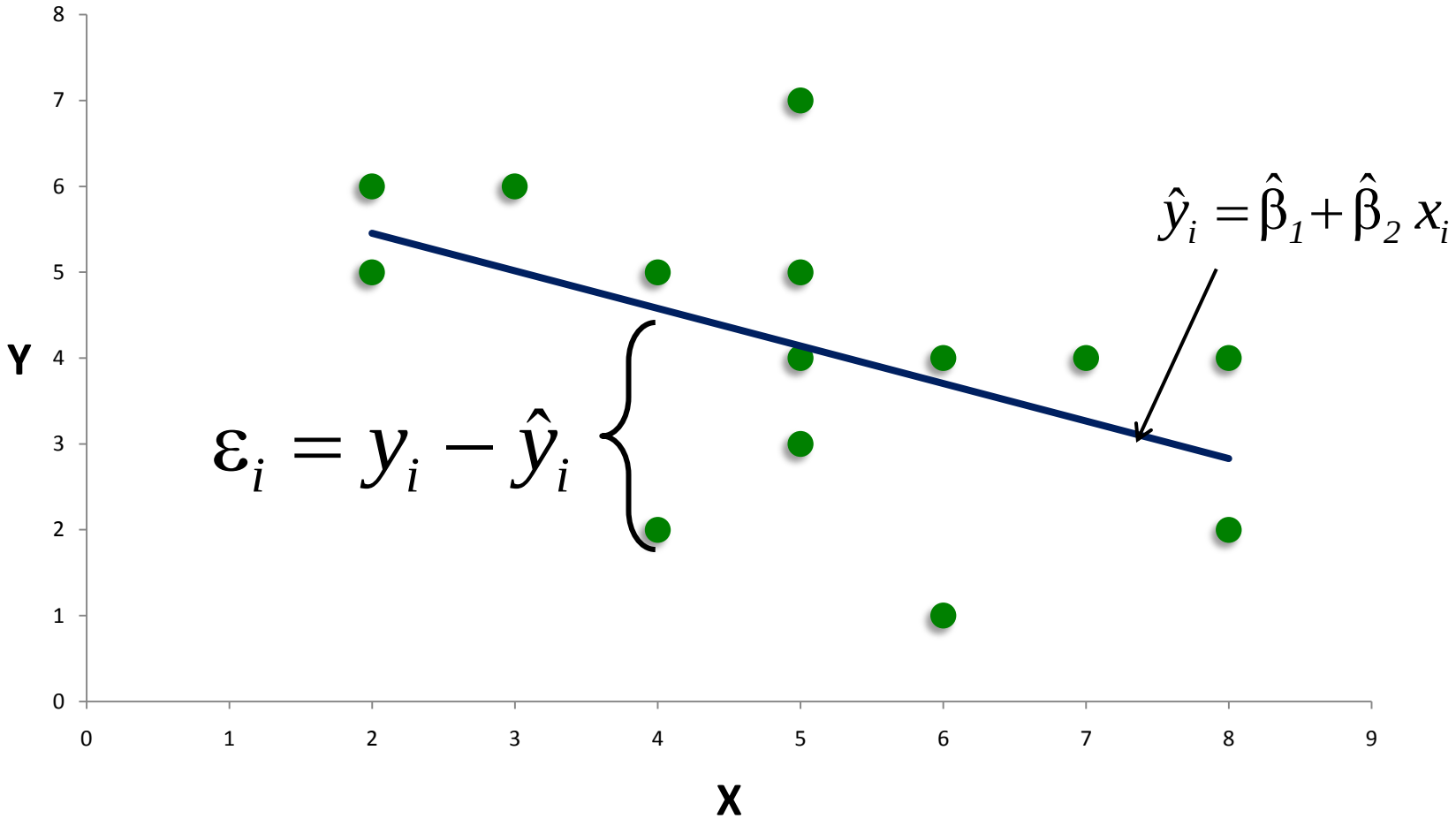
Метод наименьших квадратов

Пусть имеется набор из n пар y и x

Задача: подобрать функцию, которая наилучшим образом описывает зависимость y от x



Метод наименьших квадратов



Метод наименьших квадратов

Идея метода наименьших квадратов: подбор параметров

$\hat{\beta}_1$ **и** $\hat{\beta}_2$

таким образом, чтобы сумма квадратов отклонений

$$\varepsilon_1^2 + \varepsilon_2^2 + \ldots + \varepsilon_n^2 = \sum_{i=1}^{n} \varepsilon_i^2$$

была наименьшей

Метод наименьших квадратов

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_1 - \hat{\beta}_2 \right)^2 \to \min_{\hat{\beta}_1, \hat{\beta}_2}$$

Откуда получим:

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

$$\hat{\beta}_2 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{Cov(x, y)}{Var(x)}$$

Свойства оценок параметров

Существуют следующие критерии оценок параметров:

- **несмещенность,**
- **состоятельность,**
- **эффективность.**

Несмещенность оценки означает, что при ее использовании мы не получаем систематической ошибки, и только при наличии этого свойства оценки могут иметь практическую значимость. Математически несмещенность оценки означает, что математическое ожидание остатков равно 0.

Свойства оценок параметров

Состоятельность оценки гарантирует приближение оценки к истинному значению (т. е. увеличение их точности) при увеличении объема выборки.

Эффективная оценка является наилучшей в смысле минимума среднеквадратичного отклонения. Оценки, полученные методом наименьших квадратов при выполнении всех необходимых предпосылок (гипотез), являются эффективными.

Качество оценивания

Коэффициент детерминации R^2 показывает долю дисперсии зависимой переменной, «объясненной» уравнением регрессии

$$R^2 = \frac{Var(\hat{y})}{Var(y)} = \frac{\sum_{i=1}^{n}(\hat{y} - \bar{y})^2}{\sum_{i=1}^{n}(y - \bar{y})^2}$$

Если между переменными существует точная линейная связь, то $R^2 = 1$

Если в выборке отсутствует видимая линейная связь между переменными, то R^2 близок к 0.

Качество оценивания

Высокий R^2 сам по себе не гарантирует, что модель является хорошей.
Остается риск ложной регрессии.

Низкий R^2 говорит о том, что существуют важные факторы, которые мы не учли в модели

Оценка дисперсии случайной ошибки

Используется для расчета стандартных ошибок коэффициентов

$$\sigma^2 = \frac{1}{n-2}\sum_{i=1}^{n}\varepsilon_i^2$$

Стандартные ошибки коэффициентов

Используются для измерения точности полученных оценок: чем ниже стандартные ошибки, тем точнее оценки коэффициентов

В модели парной регрессии стандартные ошибки (Standard errors) коэффициентов оцениваются по следующим формулам

$$SE(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{Var(x)}\right)}$$

$$SE(\hat{\beta}_2) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Тестирование значимости коэффициента

$\hat{\beta}_1$ и $\hat{\beta}_2$ - оценки, полученные при помощи МНК, на основе случайной выборки. Следовательно, они сами являются случайными величинами

Нужно уметь определять, достаточно ли сильно $\hat{\beta}_2$ отличается от нуля для того, чтобы можно было с уверенностью утверждать, что и истинное значение коэффициента также не равно нулю?

На практике для решения этой задачи используется **тест на значимость коэффициента**

Тестирование значимости коэффициента

Рассматриваемая модель $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$

Тестируемая гипотеза

H0: $\beta_2 = 0$ **«Переменная x не оказывает значимого влияния на переменную y»**

Альтернативная гипотеза

H1: $\beta_2 \neq 0$ **«Переменная x оказывает значимое влияние на переменную y»**

Тестирование значимости коэффициента

Алгоритм проведения теста

Шаг 1

Вычисляем расчетное значение t-статистики

$$t_{расч} = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)}$$

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 или 5%.

Тестирование значимости коэффициента

Шаг 3

Из таблиц t-распределения Стьюдента находим критическое значение t-статистики $t_{кр}$

Оно зависит от уровня значимости (двусторонний тест) α и от так называемого числа степеней свободы, которое в случае нашего теста равно (n – 2)

Тестирование значимости коэффициента

Шаг 4

Сравниваем расчетное и критическое значение t-статистик

Если $\left|t_{расч}\right| < t_{кр}$

то гипотеза H0 не отклоняется (принимается),

то есть мы делаем вывод о том, что переменная x не оказывает значимого влияния на переменную y. В этом случае коэффициент при переменной x называют незначимым.

В противном случае гипотеза H0 не принимается (отклоняется).

Тестирование значимости коэффициента

Исследуется зависимость часового заработка (в $) работника (Z) от числа законченных лет обучения (S):

$$Z = \beta_1 + \beta_2 S + \varepsilon_i$$

На основе данных о 540 работниках было получено следующее уравнение регрессии (в скобках – стандартные ошибки оценок коэффициентов):

$$Z = \underset{(3,2)}{-13,9} + \underset{(0,2)}{2,4}S$$

Можно ли утверждать, что число лет обучения значимо влияет на заработок?

Тестирование значимости коэффициента

H0: $\beta_2 = 0$ **«Переменная S не оказывает значимого влияния на переменную Z»**

$$t_{расч} = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} = \frac{2,4}{0,2} = 12$$

При уровне значимости 1% и числе степеней свободы (n – 2) = 540 – 2 = 538

$$t_{кр} = t(538) = 2,6 \qquad \left|t_{расч}\right| > t_{кр}$$

следовательно гипотеза H0 отвергается и мы делаем вывод том, что число лет обучения значимо влияет на заработок

Тестирование гипотезы H0: $\beta_2 \equiv A$

$$t_{расч} = \frac{\hat{\beta}_2 - A}{SE(\hat{\beta}_2)}$$

Доверительные интервалы

$$(\hat{\beta}_2 - t_{n-2} \cdot SE(\hat{\beta}_2), \hat{\beta}_2 + t_{n-2} \cdot SE(\hat{\beta}_2))$$

Доверительный интервал – это границы в которых с вероятностью (1-α) находятся значения истинных параметров регрессии

Классическая линейная модель множественной регрессии

$$y_i = \beta_1 + \beta_2 x_i^{(2)} + \beta_3 x_i^{(3)} + \ldots + \beta_k x_i^{(k)} + \varepsilon_i$$

y_i — значения зависимой переменной,

$x_i^{(k)}$ — значения независимых переменных (регрессоров),

k — число коэффициентов в модели,

ε_i — случайные ошибки,

Качество подгонки модели

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

Стандартная ошибка регрессии

$$SEE = \sqrt{\frac{\sum \varepsilon_i^2}{n-k}}$$

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

Коэффициент детерминации R^2 показывает долю дисперсии зависимой переменной, «объясненной» уравнением регрессии

$$R^2 = \frac{Var(\hat{y})}{Var(y)} = \frac{\sum_{i=1}^{n}(\hat{y} - \bar{y})^2}{\sum_{i=1}^{n}(y - \bar{y})^2}$$

При добавлении в модель новых переменных R^2 не может уменьшаться, поэтому сравнение моделей с разным числом переменных на основе этого показателя некорректно

Скорректированный (нормированный) R^2

R^2 с учетом штрафа за число переменных

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

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

Использование R^2 и R^2_{adj}

Не следует сводить выбор уравнения к задаче максимизации R^2 или R^2 adj.

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

3. Высокий R^2 (или R^2_{adj}) не гарантирует отсутствия смещения оценок из-за некорректной спецификации

Тестирование значимости коэффициента

$$y_i = \beta_1 + \beta_2 x_i^{(2)} + \beta_3 x_i^{(3)} + \ldots + \beta_k x_i^{(k)} + \varepsilon_i$$

Тестируемая гипотеза

H0: $\beta_k = 0$ «Переменная $x^{(k)}$ не оказывает значимого влияния на переменную y»

Альтернативная гипотеза

H1: $\beta_k \neq 0$ «Переменная $x^{(k)}$ оказывает значимое влияние на переменную y»

Тестирование значимости коэффициента

Алгоритм проведения теста

Шаг 1

Вычисляем расчетное значение t-статистики

$$t_{расч} = \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)}$$

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 или 5%.

Тестирование значимости коэффициента

Шаг 3

Из таблиц t-распределения Стьюдента находим критическое значение t-статистики $t_{кр}$

Оно зависит от уровня значимости (двусторонний тест) α и от так называемого числа степеней свободы, которое в случае нашего теста равно $(n - k)$

Тестирование значимости коэффициента

Шаг 4

Сравниваем расчетное и критическое значение t-статистик

Если $\left|t_{расч}\right| < t_{кр}$

то гипотеза H0 не отклоняется (принимается),

то есть мы делаем вывод о том, что переменная $x^{(k)}$ не оказывает значимого влияния на переменную y. В этом случае коэффициент при переменной $x^{(k)}$ называют незначимым.

В противном случае гипотеза H0 не принимается (отклоняется).

Тестирование коэффициента

Тестирование гипотезы H0: $\beta_k = A$

$$t_{расч} = \frac{\hat{\beta}_k - A}{SE(\hat{\beta}_k)}$$

Доверительные интервалы

$$(\hat{\beta}_k - t_{n-k} \cdot SE(\hat{\beta}_k), \hat{\beta}_k + t_{n-k} \cdot SE(\hat{\beta}_k))$$

Доверительный интервал – это границы в которых с вероятностью (1-α) находятся значения истинных параметров регрессии

Тестирование значимости уравнения

Рассматриваемая модель

$$y_i = \beta_1 + \beta_2 x_i^{(2)} + \beta_3 x_i^{(3)} + \ldots + \beta_k x_i^{(k)} + \varepsilon_i$$

Тестируемая гипотеза
H0: $\beta_2 = \beta_3 = \ldots = \beta_k = 0$ «Все переменные $x^{(2)}$ … $x^{(k)}$ не оказывают значимого влияния на переменную y»

Альтернативная гипотеза
H1: «Хотя бы одна из переменных $x^{(2)}$ … $x^{(k)}$ оказывает значимое влияние на переменную y»

Тестирование значимости уравнения

Алгоритм проведения теста

Шаг 1

Вычисляем расчетное значение F-статистики

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

Шаг 2

Выбираем уровень значимости

Тестирование значимости уравнения

Шаг 3

Из таблиц F-распределения Фишера находим критическое значение F-статистики $F_{кр}$

Оно зависит от уровня значимости α и от числа степеней свободы, которые равны ($k - 1$) и ($n - k$)

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 $x^{(2)}…x^{(k)}$ не оказывает значимого влияния на переменную y

В этом случае уравнение называют незначимым.

В противном случае гипотеза H0 не принимается (отклоняется).

Тестирование значимости уравнения

Пример

Получено следующее уравнение (n = 30):

$$y = 5,34 + 11,2x^{(2)} + 0,3x^{(3)} + 6,1x^{(4)}$$

$$R^2 = 0,75$$

Тестируемая гипотеза

H0: $\beta_2 = \beta_3 = \beta_4 = 0$

Тестирование значимости уравнения

Пример

Шаг 1

Вычисляем расчетное значение F-статистики

$$F_{расч} = \frac{R^2}{1-R^2}\frac{n-k}{k-1} = \frac{0,65}{1-0,65}\frac{30-4}{4-1} = 16,09$$

Шаг 2

Выбираем уровень значимости α = 0,05

Тестирование значимости уравнения

Пример

Шаг 3

Из таблиц F-распределения Фишера находим критическое значение F-статистики $F_{кр}$
Оно зависит от уровня значимости α и от числа степеней свободы, которые равны ($k - 1$) и ($n - k$)

Для α =0,05 Fкр (k-1, n-k) = Fкр (2, 27) = 3,35

Тестирование значимости уравнения

Пример

Шаг 4

Сравниваем расчетное и критическое значение F-статистик

$F_{расч} = 16,09$ **и** Fкр = 3,35

$F_{расч} > F_{кр}$, **делаем вывод о том, что хотя бы одна переменная** $x^{(2)}…x^{(k)}$ **оказывает значимое влияния на переменную y**

В этом случае уравнение называют значимым.

Тест «короткая» регрессия против «длинной»

«Короткая» регрессия:

$$y_i = \beta_1 + \beta_2 x_i^{(2)} + \beta_3 x_i^{(3)} + \ldots + \beta_m x_i^{(m)} + \varepsilon_i$$

«Длинная» регрессия (*m + q = k*):

$$y_i = \beta_1 + \beta_2 x_i^{(2)} + \beta_3 x_i^{(3)} + \ldots + \beta_m x_i^{(m)} +$$

$$\beta_{m+1} x_i^{(m+1)} + \ldots + \beta_{m+q} x_i^{(m+q)} + \varepsilon_i$$

Тестируемая гипотеза

H0: $\beta_{m+1} = \beta_{m+2} = \ldots = \beta_{m+q} = 0$

Тест «короткая» регрессия против «длинной»

Тестируемая гипотеза

H0: $\beta_{m+1} = \beta_{m+2} = \ldots = \beta_{m+q} = 0$

R^2_{UR} (unrestricted) — в «длинной» регрессии

R^2_R (restricted) — в «короткой» регрессии

$$F_{расч} = \frac{\left(R^2_{UR} - R^2_R\right)/q}{\left(1 - R^2_{UR}\right)/(n-k)}$$

$$F_{табл} = F(q, n-k)$$

Тест «короткая» регрессия против «длинной»

Пример

Получены следующие уравнения (n = 30):

$$y = 5{,}34 + 11{,}2x^{(2)} + 0{,}3x^{(3)} + 6{,}1x^{(4)}$$

m = 4

$$R^2 = 0{,}75$$

k = 6

$$y = 6{,}32 + 10{,}8x^{(2)} - 1{,}3x^{(3)} + 6{,}1x^{(4)} + 17{,}5x^{(5)} - 4{,}2x^{(6)}$$

$$R^2 = 0{,}78$$

q = 2

Тестируемая гипотеза

H0: $\beta_5 = \beta_6 = 0$

Тест «короткая» регрессия против «длинной»

Пример

$$F_{расч} = \frac{\left(R^2_{UR} - R^2_R\right)/q}{\left(1 - R^2_{UR}\right)/\left(n-k\right)} = \frac{(0{,}80 - 0{,}75)/2}{(1-0{,}80)/(30-6)} = 3{,}012$$

$$F_{табл} = F\left(q, n-k\right) = F(2{,}26) = 3{,}39$$

$F_{расч} < F_{кр}$, делаем вывод о том, что переменные $x^{(5)}$ и $x^{(6)}$ *не* оказывают значимое влияния на переменную y

В этом случае добавление еще двух переменных не имеет смысла.

Мультиколлинеарность

Строгая мультиколлинеарность – наличие линейной функциональной связи между объясняющими переменными
= Линейная зависимость столбцов матрицы регрессоров

Оценка коэффициентов модели при помощи метода наименьших квадратов невозможна

Возникновение
строгой мультиколлинеарности означает, что на начальном этапе набор независимых переменных был сформирован некорректно

Мультиколлинеарность

Частичная мультиколлинеарность —наличие сильной линейной корреляционной связи между регрессорами

— Основное негативное последствие – стандартные ошибки оценок коэффициентов оказываются высокими. Точность оценивания оказывается низкой

— Та или иная степень корреляции между регрессорами существует всегда. Проблема возникает только когда эта линейная связь проявляется слишком сильно

Выявление м/к на начальном этапе моделирования (до регрессии)

(1) Большие (по абсолютной величине) парные коэффициенты корреляции между независимыми переменными

(2) Высокие (>10) значения коэффициента VIF

Выявление м/к на начальном этапе моделирования (до регрессии)

Коэффициент VIF (variance inflation factor) характеризует силу мультиколлинеарности

Вычисляется на основе значений R^2 во вспомогательных регрессиях одного регрессора на другие:

$$x_i^k = \beta_1 + \beta_2 x_i^{(2)} + \beta_3 x_i^{(3)} + \ldots + \beta_{k-1} x_i^{(k-1)} + u_i$$

$$VIF = \frac{1}{1 - R^2}$$

Выявление м/к в построенной модели («симптомы» м/к)

— Небольшое изменении исходных данных, приводит к существенному изменению оценок коэффициентов

— Каждая переменная в отдельности является незначимой, а уравнение в целом имеет высокий R^2 и является значимым

— Оценки коэффициентов имеют неправильные с точки зрения экономической теории знаки или неоправданно большие значения

Устранение мультиколлинеарности

- **Увеличить число наблюдений**
- **Исключить переменную, с которой связана мультиколлинеарность.**

Следует помнить, что иногда это может привести к более серьезным проблемам

- **Использовать нелинейные формы зависимостей**
- **Использовать агрегаты: линейные комбинации переменных**

Фиктивные переменные

Фиктивные переменные — бинарные переменные (принимают значения 0 или 1)

Используются для моделирования качественных признаков

Фиктивные переменные: пример

Существует ли дискриминация на рынке труда?

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i$$

Y – зарплата, долларов в час

X – стаж работы, лет

$$D_i = \begin{cases} 0, если - респондет - мужчина \\ 1, если - респондет - женщина \end{cases}$$

Фиктивные переменные: пример

$$\hat{Y}_i = 4{,}2 + 2{,}1X_i - 3{,}5D_i$$

(все переменные оказались значимы)

$$Мужчины (D_i = 0): \hat{Y}_i = 4{,}2 + 2{,}1X_i$$

$$Женщины (D_i = 1): \hat{Y}_i = 4{,}2 + 2{,}1X_i - 3{,}5D_i$$

Интерпретация:

При прочих равных условиях (в данном случае при равном стаже работы) женщины получают на 3,5 доллара в час меньше, чем мужчины

Фиктивные переменные: пример

Фиктивные переменные сдвига и наклона

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + \beta_4 X_i D_i + \varepsilon_i$$

Мужчины ($D_i = 0$):

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$$

Женщины ($D_i = 1$):

$$Y_i = (\beta_1 + \beta_3) + (\beta_2 + \beta_4) X_i + \varepsilon_i$$

Тестирование структурного сдвига

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + \beta_4 X_i D_i + \varepsilon_i$$

Как проверить, оправдано ли включение фиктивных переменных в модель?

Надо проверить гипотезу: $H0$: $\beta_3 = \beta_4 = 0$

Можно сделать это при помощи обычного теста для сравнения «короткой» и «длинной» регрессии.

Линейная модель

Пока мы рассматривали линейные модели

$$y_i = \beta_1 + \beta_2 x_i^{(2)} + \beta_3 x_i^{(3)} + \varepsilon_i$$

Какова интерпретация коэффициента β_2 в такой модели? (Будем считать, что $\beta_2 > 0$)

Линейная модель

Пока мы рассматривали линейные модели

$$y_i = \beta_1 + \beta_2 x_i^{(2)} + \beta_3 x_i^{(3)} + \varepsilon_i$$

Какова интерпретация коэффициента β_2 в такой модели? (Будем считать, что $\beta_2 > 0$)

$$\Delta y_i = \beta_2 \Delta x_i^{(2)}$$

При прочих равных условиях (то есть при неизменности других переменных)

увеличение переменной $x^{(2)}$ на единицу вызывает увеличение переменной y на β_2 единиц

(Двойная) логарифмическая модель

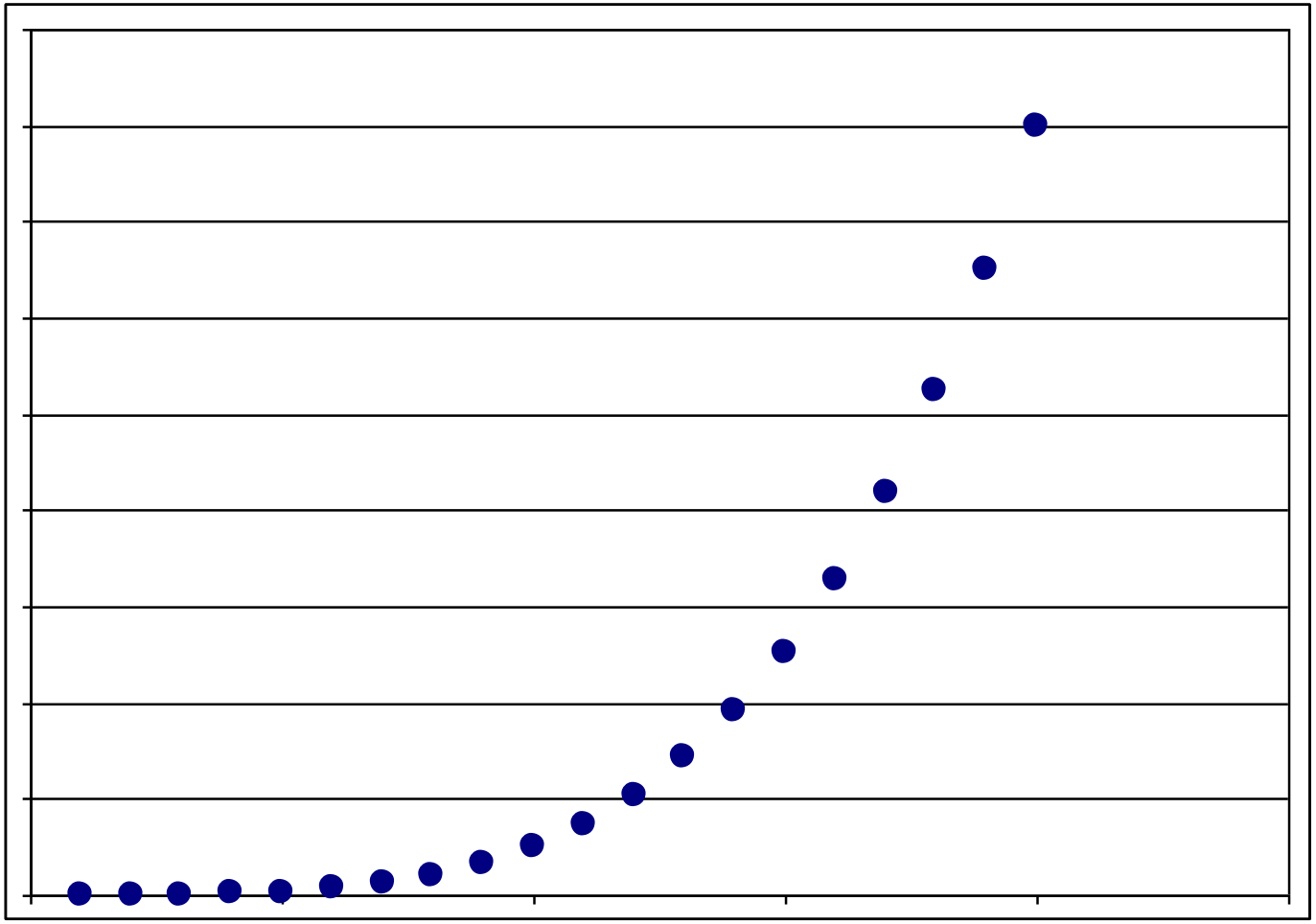
Часто в экономике зависимости носят не линейный, а степенной характер

Например, производственная функция Кобба — Дугласа

$$Q = AK^{\alpha}L^{\beta}$$

(Двойная) логарифмическая модель

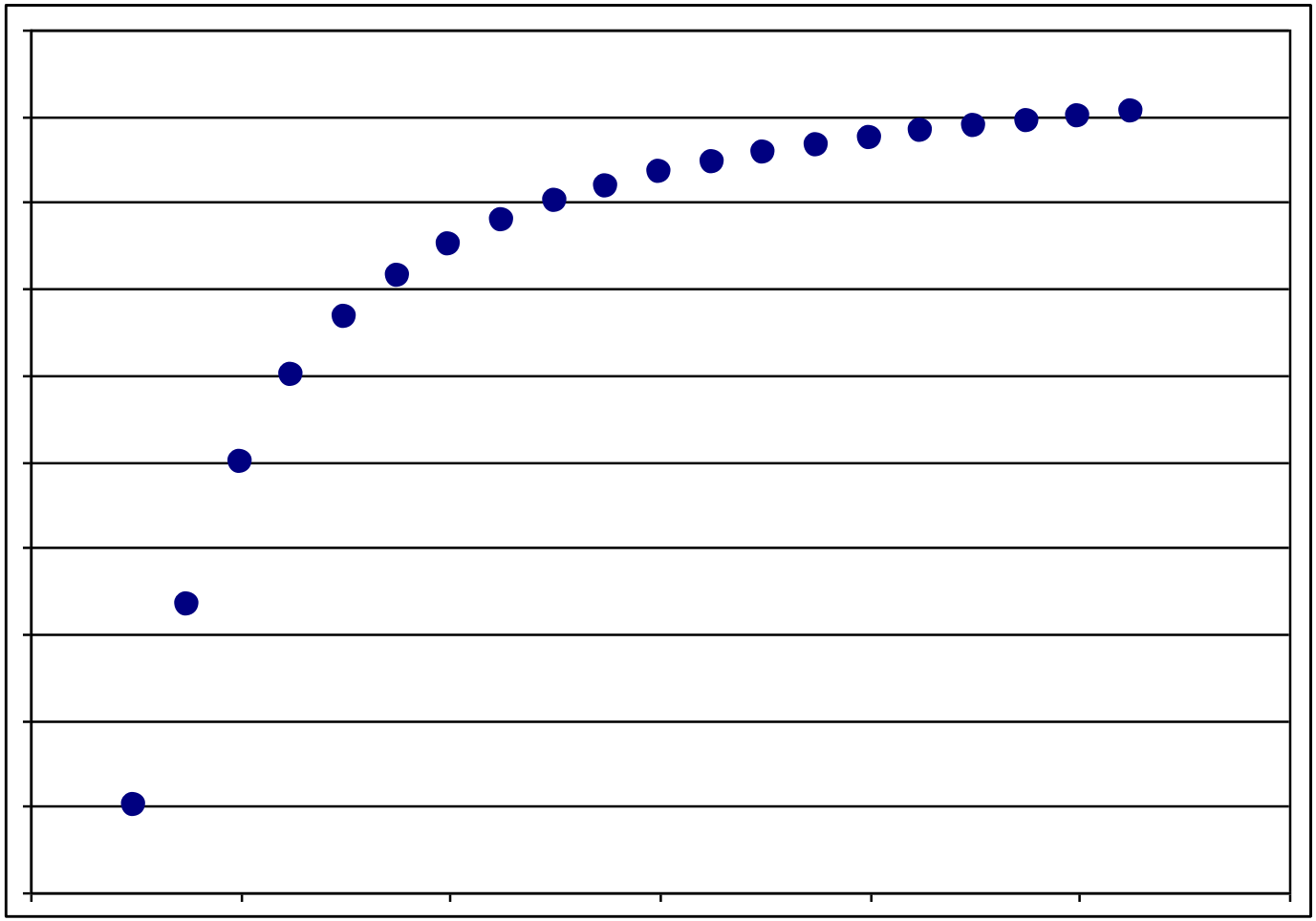
$$y = Ax^{\alpha} \qquad \boxed{\alpha > 1}$$



(Двойная) логарифмическая модель

$$y = Ax^{\alpha}$$

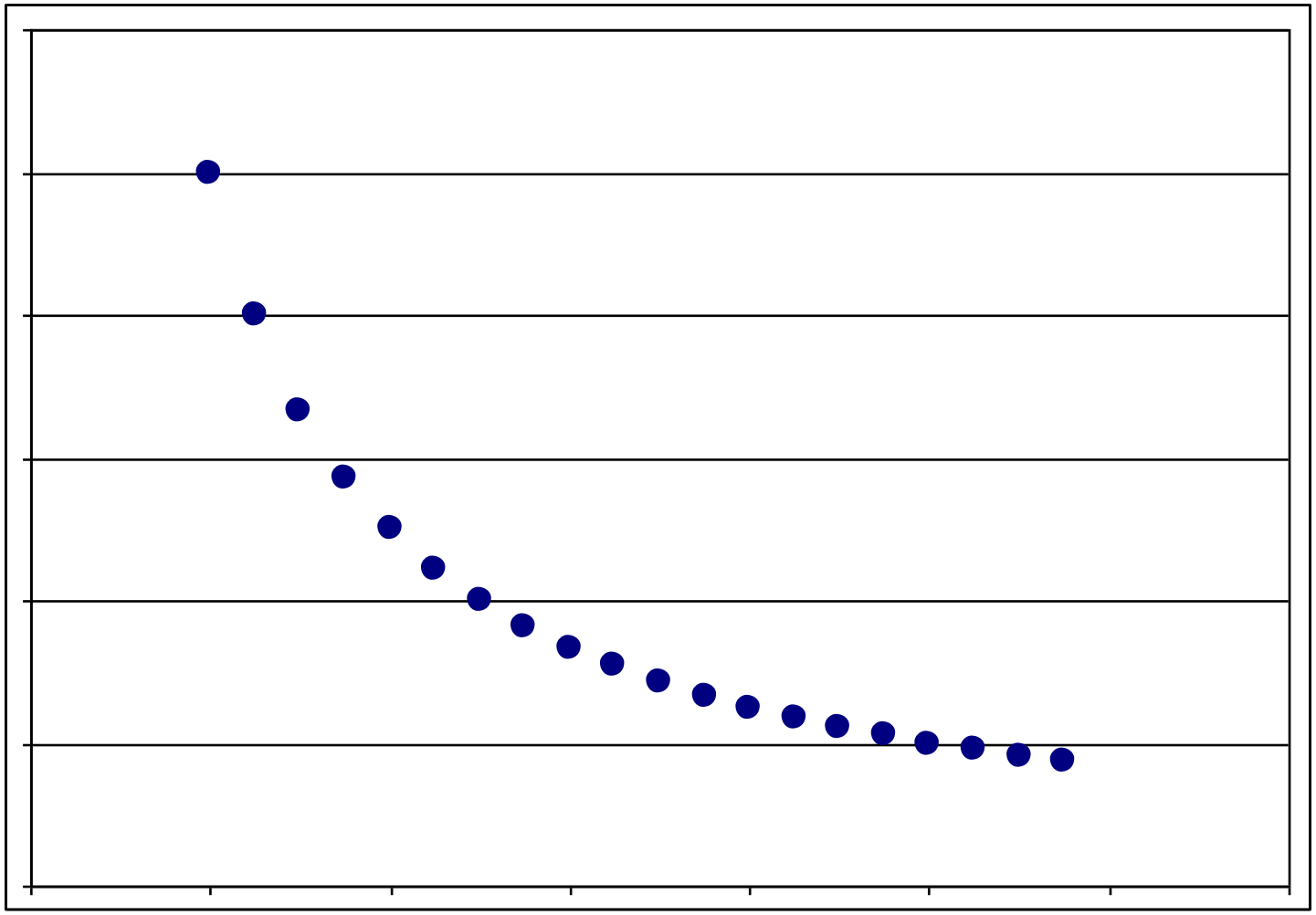
$$0 < \alpha < 1$$



(Двойная) логарифмическая модель

$$y = Ax^{\alpha}$$

$\alpha < 0$



(Двойная) логарифмическая модель

Как, используя метод наименьших квадратов, оценить параметры такой модели?

$$y = Ax^{\alpha}$$

Нужно сделать ее линейной по параметрам:

$$\ln y = \ln A + \alpha \ln x$$

$$\ln y = \beta_1 + \beta_2 \ln x$$

(Двойная) логарифмическая модель

$$\ln y = \beta_1 + \beta_2 \ln x$$

Интерпретация коэффициента

$$\frac{dy}{y} = \beta_2 \frac{dx}{x} \qquad \frac{\Delta y}{y} \approx \beta_2 \frac{\Delta x}{x}$$

Увеличение ***x*** на 1% $\Rightarrow$ увеличение ***y*** на β_2 %

β_2 — это эластичность ***y*** по ***x***

Логарифмически-линейная модель

Экспоненциальная зависимость

$$y = e^{\beta_1 + \beta_2 x}$$

Параметры такой зависимости оцениваются при помощи логарифмически-линейной модели

$$\ln y = \beta_1 + \beta_2 x$$

Логарифмически-линейная модель

$$\ln y = \beta_1 + \beta_2 x$$

Интерпретация коэффициента

$$\frac{dy}{y} = \beta_2 dx \qquad \frac{\Delta y}{y} \approx \beta_2 \Delta x$$

Увеличение ***x*** на единицу $\Rightarrow$
увеличение ***y*** на $(100\beta_2)\%$

Важно: чем больше β_2, тем менее точным является это приближение

Если $\beta_2 > 0{,}1$ то лучше пользоваться не приближенными вычислениями, а точными

Логарифмически-линейная модель

Пример

Моделирование экономического роста

$$\ln \boldsymbol{ВВП}_t = 4{,}2 + 0{,}03t$$

Увеличение ***t*** на единицу $\Rightarrow$
увеличение ***ВВП*** на $(100 * 0{,}03)\%$

Темп прироста ВВП составляет 3% в год

Логарифмически-линейная модель

Вопрос: **какую формулу следует использовать, если β_2 больше 0,1?**

Для ответа на него нужно вспомнить, что логарифмически-линейная модель характеризует экспоненциальную зависимость:

$$y_i = e^{\beta_1 + \beta_2 x_i}$$

Логарифмически-линейная модель

Обозначим прирост зависимой переменной как . То есть

$$\Delta y = \frac{y_1 - y_0}{y_0} = \frac{y_1}{y_0} - 1$$

Где y_1 характеризует y после изменения x, а y_0 – до изменения.

$$\Delta y = \frac{e^{\beta_1 + \beta_2 x_1}}{e^{\beta_1 + \beta_2 x_0}} - 1 = e^{(\beta_1 + \beta_2 x_1) - (\beta_1 + \beta_2 x_0)} - 1 = e^{\beta_2 \Delta x} - 1$$

Пример

Допустим, при оценке какой-нибудь зависимости мы получили следующие результаты:

$$\ln \hat{y}_i = 23 + 0{,}03 x_i^{(2)} + 2{,}4 x_i^{(3)}$$

Как мы можем интерпретировать коэффициенты при иксах?

Логарифмически-линейная модель

$$\ln \hat{y}_i = 23 + 0{,}03x_i^{(2)} + 2{,}4x_i^{(3)}$$

При $x^{(2)}$: при прочих равных условиях при увеличении $x^{(2)}$ на единицу, *y* увеличивается на

$$\left(e^{0{,}03} - 1\right) \cdot 100\% = 3{,}045\%$$

Действительно, 3,045 ≈ 3. То есть приближенная формула показывает верный результат.

Логарифмически-линейная модель

$$\ln \hat{y}_i = 23 + 0{,}03x_i^{(2)} + 2{,}4x_i^{(3)}$$

При $x^{(3)}$: при прочих равных условиях при увеличении $x^{(3)}$ на единицу, *y* увеличивается на

$$\left(e^{2,4} - 1\right) \cdot 100\% = 1002{,}32\%$$

При этом по приближенной формуле мы бы получили всего лишь 240%. То есть мы бы допустили ошибку на 762%.

Логарифмически-линейная модель

Поскольку при интерпретации коэффициентов мы исходим из того, что *x* изменился на единицу, то Δx=1 . То есть:

$$\Delta y = e^{\beta_2} - 1$$

Чтобы получить изменение в процентах, а не в долях (как это сейчас), полученное выражение следует умножить на 100%. Таким образом, при прочих равных условиях увеличение *x* на единицу приведет к изменению *y* на

$$(e^{\beta_2} - 1) * 100\%$$

Линейно-логарифмическая модель

$$y = \beta_1 + \beta_2 \ln x$$

Интерпретация коэффициента

$$dy = \beta_2 \frac{dx}{x} \qquad \Delta y \approx \beta_2 \frac{\Delta x}{x}$$

Увеличение ***x*** на 1% $\Rightarrow$

увеличение ***y*** на $(\beta_2 / 100)$ единиц

Пример

По 1000 наблюдений было оценено следующее уравнение регрессии (в скобках указаны стандартные отклонения оценок коэффициентов):

$$\hat{y}_i = \underset{(19)}{-117} + \underset{(4)}{20}\ln x_i + \underset{(5)}{20} z_i, \qquad R^2 = 0.95.$$

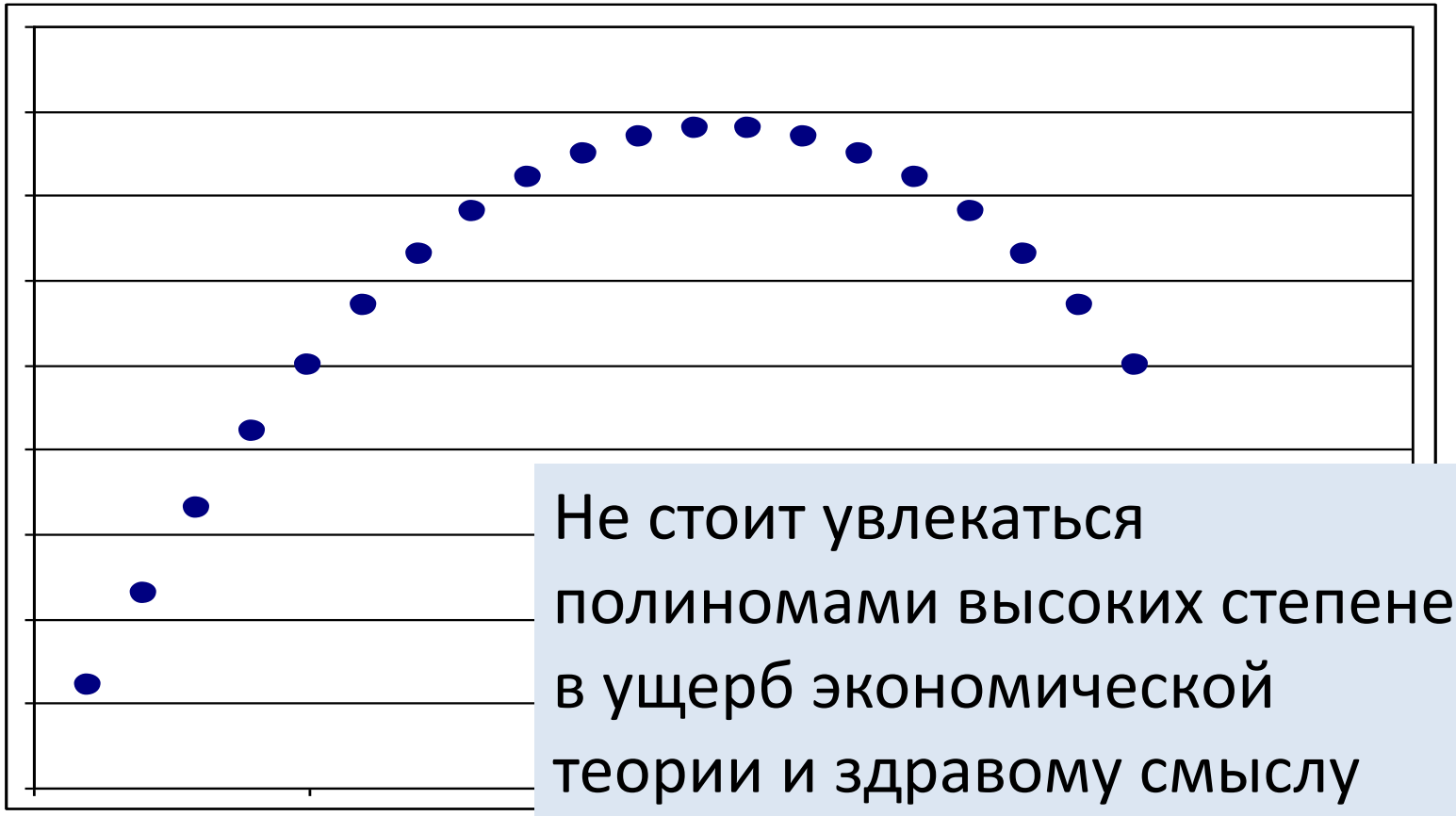
Дайте интерпретацию коэффициента при переменной ln x.

При прочих равных условиях при увеличении переменной *x* на 1% переменная *y* увеличивается на 0,2 единицы

Полиномы относительно регрессоров

Позволяют моделировать немонотонные зависимости

$$y = \beta_1 + \beta_2 x + \beta_3 x^2$$



Не стоит увлекаться полиномами высоких степеней в ущерб экономической теории и здравому смыслу