Федеральное агентство по образованию
Иркутский государственный университет

А.Ю. Филатов

Конспект лекций по эконометрике

учебное пособие

Иркутск 2006

Печатается по решению редакционно-издательского совета
Иркутского государственного университета

УДК 330.43
ББК 22.172я73

Рецензенты: д.т.н., проф. Зоркальцев В.И.
(зав.кафедрой математической экономики ИМЭИ ИГУ),
к.ф.-м.н. Тюрнева Т.Г.
(доцент кафедры теории вероятностей и дискретной математики ИМЭИ ИГУ),
к.ф.-м.н. Солодуша С.В.
(доцент кафедры математики в экономике Института экономики ИрГТУ)

Филатов А.Ю. Конспект лекций по эконометрике: учеб. пособие. – Иркутск: Иркутский ун-т, 2006. – 35с.

Охватывает основные разделы регрессионного анализа. В частности, дается информация о классической и обобщенной линейной модели множественной регрессии, методе наименьших квадратов, методах обработки неоднородных данных, нелинейных регрессионных моделях. Изложение сопровождается задачным материалом, приводятся модели из области микро- и макроэкономики. Издание содержит типовые задания для контрольных работ и вопросы к экзамену.

Пособие предназначено для студентов, изучающих эконометрику, в качестве дополнения к лекционному курсу и рекомендуемой литературе.

От автора

Несмотря на то, что курс эконометрики читается во многих вузах, до сих пор является незаполненной столь востребованная студентами ниша «конспектов лекций», выполненных преподавателями. Есть много хороших как российских, так и переводных учебников (среди них особо отметим двухтомник С. А. Айвазяна и В. С. Мхитаряна «Прикладная статистика и основы эконометрики», на следование которому, в первую очередь, и ориентирован данный курс). Есть совсем небольшие «шпаргалки», содержащие минимум информации и основные формулы. В то же время имеется насущная необходимость в чем-то среднем – а именно, в пособии, где студент, прослушавший курс лекций, быстро найдет необходимую формулу, свойство или разобранный пример, не утонув при этом в потоке излишней информации. Данное пособие может быть полезно для подготовки к экзаменам, для быстрого восстановления в памяти нужного материала, при решении практических задач.

При написании автор руководствовался принципом максимального приближения стиля изложения к конспекту лекций: тезисно, иногда обрывочными фразами (а именно так пишутся конспекты!) дать основную информацию; по возможности, разложить ее по пунктам; выделить самое главное (для этого использованы жирный и курсивный шрифты, шрифты большего размера, обрамление); привести основные формулы, а также соответствующие команды из Excel (базовые навыки работы в Excel необходимы для проведения расчетов). Поскольку численные примеры существенно облегчают восприятие, учебное пособие содержит примеры и задачи, основанные на моделях микро- и макроэкономики. Все примеры обозначаются знаком «##».

Материал, содержащийся в пособии, разбит на 14 лекций, каждая из которых занимает две страницы на одном развороте. Сделана попытка полностью соответствовать изложению в рамках курса лекций, который автор читает в ИМЭИ ИГУ. Для этого произведен отход от некоторых принятых стандартов. В частности, список литературы не вынесен на отдельную страницу, а дан (как и при чтении лекций студентам) в начале курса. В содержании указаны не названия лекций, а приведена краткая информация о темах, в них содержащихся. В заключительной части пособия приведены типовые задания для контрольных работ и вопросы к экзамену.

Издаваемый «конспект лекций» является первым, но не последним в серии предполагаемых пособий. В частности, планируется издание конспектов по двум другим эконометрическим курсам: «Многомерным статистическим методам» и «Эконометрическому моделированию».

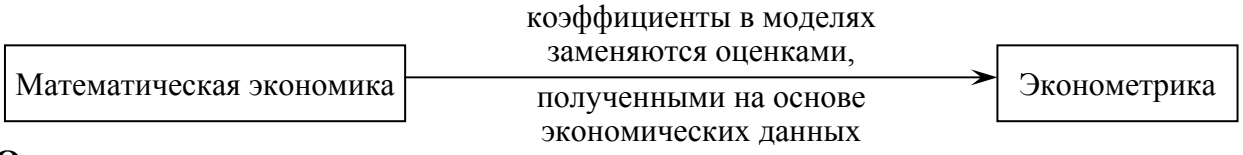
Автор выражает благодарность В.И. Зоркальцеву, Т.Г. Тюрневой и С.В. Солодуша за рецензирование учебного пособия и ценные идеи и замечания, высказанные в процессе его обсуждения.

Литература

1. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. М.: Юнити, 2002.
2. Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс. М.: Дело, 1997.
3. Суслов В.И., Ибрагимов Н.М., Талышева Л.П., Цыплаков А.А. Эконометрия. Новосибирск, 2003.
4. Green W. Econometric Analysis, 3-d edition, Prentice Hall, 1997.

Введение в эконометрику

Эконометрика – «измерения в экономике» (Рагнар Фриш, норвежский экономист, 1926);
– придает количественное выражение качественным закономерностям, вводимым экономической теорией.



Основа эконометрики

1. Экономические законы (микроэкономика, макроэкономика, математическая экономика).
2. Информационное обеспечение (экономическая статистика).
3. Методы (математико-статистический инструментарий).

Используемые методы

1. Корреляционный анализ.
2. Регрессионный анализ.
3. Анализ временных рядов.
4. Системы одноврем. уравн.
5. Методы классификации.
6. Методы снижения размерности.

Конечные прикладные цели эконометрики

1. Мониторинг. 2. Прогноз. + 3. Управление. + 4. Устойчивое развитие.

Уровни иерархии

1. Макроуровень (страны, мир).
2. Мезоуровень (регионы, отрасли).
3. Микроуровень (домашние хозяйства, фирмы).

Принципиальная идея – наличие взаимосвязей между переменными.

спрос ← цена, доход, цены на другие товары;
затраты ← объем производства, его динамика, цены на факторы производства;
потребительские расходы ← доход, ликвидные активы, предельный уровень потребления.

Даже при фиксации объясняющих переменных на едином уровне есть варьирование результируюшей переменной – имеется случайная составляющая!

Аддитивная линейная форма (наиболее распространена):

$y_t = \theta_0 + \theta_1 x_t^{(1)} + ... + \theta_p x_t^{(p)} + \varepsilon_t$, $\theta_0, ..., \theta_p$ – параметры (обычно неизвестные);

$\hat{y}_t = \theta_0 + \theta_1 x_t^{(1)} + ... + \theta_p x_t^{(p)}$ – модельное значение, y_t – наблюдаемое значение;

$y_t - \hat{y}_t = \varepsilon_t$ – случайная ошибка прогноза, $M\varepsilon_t = 0$, чтобы исключить систематич. ошибку.

Увеличение числа объясняющих переменных сокращает случайную составляющую!

Проблемы построения модели

1. Выбор зависимости (линейные, полиномиальные, логарифмические,...).
2. Учет запаздывания во времени (зависимость от текущего периода, предыдущего, динамики).
3. Появление «внешних» переменных (модель становится неполной).

Типы переменных

1. Объясняющие (экзогенные) – в определенной степени управляемы, задаются извне.
2. Результирующие (эндогенные) – являются предметом объяснения, формируются внутри.
3. Предопределенные – измерены в прошлом.

Системы одновременных уравнений (СОУ)

СОУ – системы, в которых переменные являются объясняющими в одних уравнениях и результирующими в других.

В эконометрических исследованиях наиболее распространены системы линейных одновременных уравнений и нелинейные модели, поддающиеся непосредственной линеаризации!

m – число результирующих переменных, $(p+1)$ – число объясняющих переменных, $x_t^{(0)} \equiv 1$.

m_1 – число уравнений со случайными компонентами, m_2 – тождеств, $m = m_1 + m_2$.

Структурная форма

$$Y_t^{(1)} = \begin{pmatrix} y_t^{(1)} \\ \dots \\ y_t^{(m_1)} \end{pmatrix}, \quad Y_t^{(2)} = \begin{pmatrix} y_t^{(m_1+1)} \\ \dots \\ y_t^{(m_1+m_2)} \end{pmatrix}, \quad X_t = \begin{pmatrix} x_t^{(0)} \\ \dots \\ x_t^{(p)} \end{pmatrix}, \quad \Delta_t = \begin{pmatrix} \delta_t^{(1)} \\ \dots \\ \delta_t^{(m_1)} \end{pmatrix}, \quad t = 1,\dots,n.$$

$$\begin{cases} B_1 Y_t^{(1)} + B_2 Y_t^{(2)} + C_1 X_t = \Delta_t, & B_1 \in R^{m_1 \times m_1}, \quad B_2 \in R^{m_1 \times m_2}, \quad C_1 \in R^{m_1 \times (p+1)}, \\ B_3 Y_t^{(1)} + B_4 Y_t^{(2)} + C_2 X_t = 0. & B_3 \in R^{m_2 \times m_1}, \quad B_4 \in R^{m_2 \times m_2}, \quad C_2 \in R^{m_2 \times (p+1)}. \end{cases}$$

Известными являются коэффициенты матриц B_3, B_4, C_2.

$$BY_t + CX_t = \overline{\Delta}_t, \qquad B = \begin{pmatrix} B_1 & B_2 \\ B_3 & B_4 \end{pmatrix}, \; C = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}, \; Y_t = \begin{pmatrix} Y_t^{(1)} \\ Y_t^{(2)} \end{pmatrix}, \; \overline{\Delta}_t = \begin{pmatrix} \Delta_t \\ 0 \end{pmatrix}.$$

Нормировка системы: $\beta_{ii} = 1, \; i = 1,\dots,m$.

Структурная форма с исключенными тождествами

Если B_4 – невырожденная матрица, то $Y_t^{(2)}$ можно выразить через $Y_t^{(1)}$ и X_t и исключить.

$$B_4 Y_t^{(2)} = -B_3 Y_t^{(1)} - C_2 X_t, \quad Y_t^{(2)} = -B_4^{-1} B_3 Y_t^{(1)} - B_4^{-1} C_2 X_t,$$

$$B_1 Y_t^{(1)} - B_2 B_4^{-1} B_3 Y_t^{(1)} - B_2 B_4^{-1} C_2 X_t + C_1 X_t = \Delta_t.$$

Осталось m_1 уравнений, тождества отсутствуют:

$$B^* Y_t^{(1)} + C^* X_t = \Delta_t, \quad B^* = B_1 - B_2 B_4^{-1} B_3, \quad C^* = C_1 - B_2 B_4^{-1} C_2.$$

Если B^* – невырожденная матрица, то, явно выразив все результирующие переменные, можно перейти к приведенной форме системы линейных одновременных уравнений.

Приведенная форма

$$Y_t^{(1)} = -\left(B^*\right)^{-1} C^* X_t + \left(B^*\right)^{-1} \Delta_t,$$

$$Y_t^{(1)} = \Pi^* X_t + \varepsilon_t^*, \qquad \Pi^* = -\left(B^*\right)^{-1} C^*, \quad \varepsilon_t^* = \left(B^*\right)^{-1} \Delta_t.$$

Возможно каждое из уравнений приведенной формы идентифицировать отдельно, однако во многих моделях сложно по идентифицированным коэффициентам приведенной формы отыскать коэффициенты структурной формы!

Потребление – функция располагаемого дохода, возрастающая медленнее, чем сам доход.
Инвестиции – возрастающая функция ВВП и убывающая функция процентной ставки.
ВНП = потребительские расходы + инвестиционные расходы + государственные расходы.

$y^{(1)}$ – потребительские расходы $x^{(1)}$ – налоги
$y^{(2)}$ – инвестиционные расходы $x^{(2)}$ – процентная ставка
$y^{(3)}$ – ВНП $x^{(3)}$ – государственные расходы

$$\begin{cases} y_t^{(1)} = \alpha_0 + \alpha_1\left(y_t^{(3)} - x_t^{(1)}\right) + \delta_t^{(1)}, \\ y_t^{(2)} = \beta_1 y_{t-1}^{(3)} + \beta_2 x_t^{(2)} + \delta_t^{(2)}, \\ y_t^{(3)} = y_t^{(1)} + y_t^{(2)} + x_t^{(3)}. \end{cases} \qquad 0 < \alpha_1 < 1,\ \beta_1 > 0,\ \beta_2 < 0.$$

1. Структурная форма

$$\begin{cases} y_t^{(1)} - \alpha_0 - \alpha_1\left(y_t^{(3)} - x_t^{(1)}\right) = \delta_t^{(1)}, \\ y_t^{(2)} - \beta_1 x_t^{(4)} - \beta_2 x_t^{(2)} = \delta_t^{(2)}, \\ y_t^{(3)} - y_t^{(1)} - y_t^{(2)} - x_t^{(3)} = 0. \end{cases} \qquad m=3,\ m_1=2,\ m_2=1,\ p=4\left(x_t^{(0)} \equiv 1,\ x_t^{(4)} \equiv y_{t-1}^{(3)}\right).$$

$$B_1 = \begin{matrix} y_t^{(1)} & y_t^{(2)} \end{matrix}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad B_2 = \begin{matrix} y_t^{(3)} \end{matrix}\begin{pmatrix} -\alpha_1 \\ 0 \end{pmatrix} \quad C_1 = \begin{matrix} x_t^{(0)} & x_t^{(1)} & x_t^{(2)} & x_t^{(3)} & x_t^{(4)} \end{matrix}\begin{pmatrix} -\alpha_0 & \alpha_1 & 0 & 0 & 0 \\ 0 & 0 & -\beta_2 & 0 & -\beta_1 \end{pmatrix}$$

$$B_3 = \begin{pmatrix} -1 & -1 \end{pmatrix} \quad B_4 = \begin{pmatrix} 1 \end{pmatrix} \quad C_2 = \begin{pmatrix} 0 & 0 & 0 & -1 & 0 \end{pmatrix}$$

Статистическому оцениванию подлежат 4 коэффициента из 24: $\alpha_0, \alpha_1, \beta_1, \beta_2$.

2. Структурная форма с исключенными тождествами (исключаем $y_t^{(3)} = y_t^{(1)} + y_t^{(2)} + x_t^{(3)}$):

$$\begin{cases} \left(1-\alpha_1\right) y_t^{(1)} - \alpha_1 y_t^{(2)} - \alpha_0 + \alpha_1 x_t^{(1)} - \alpha_1 x_t^{(3)} = \delta_t^{(1)}, \\ y_t^{(2)} - \beta_2 x_t^{(2)} - \beta_1 x_t^{(4)} = \delta_t^{(2)}. \end{cases}$$

$$B^* = \begin{matrix} y_t^{(1)} & y_t^{(2)} \end{matrix}\begin{pmatrix} 1-\alpha_1 & -\alpha_1 \\ 0 & 1 \end{pmatrix} \quad C^* = \begin{matrix} x_t^{(0)} & x_t^{(1)} & x_t^{(2)} & x_t^{(3)} & x_t^{(4)} \end{matrix}\begin{pmatrix} -\alpha_0 & \alpha_1 & 0 & -\alpha_1 & 0 \\ 0 & 0 & -\beta_2 & 0 & -\beta_1 \end{pmatrix}$$

3. Приведенная форма (выражаем $y_t^{(1)}$ и $y_t^{(2)}$ через $x_t^{(0)}, x_t^{(1)}, \ldots, x_t^{(4)}$):

$$\begin{cases} y_t^{(1)} = \dfrac{1}{1-\alpha_1}\left(\alpha_0 - \alpha_1 x_t^{(1)} + \alpha_1\beta_2 x_t^{(2)} + \alpha_1 x_t^{(3)} + \alpha_1\beta_1 x_t^{(4)}\right) + \dfrac{1}{1-\alpha_1}\left(\delta_t^{(1)} + \alpha_1\delta_t^{(2)}\right), \\ y_t^{(2)} = \beta_2 x_t^{(2)} + \beta_1 x_t^{(4)} + \delta_t^{(2)}. \end{cases}$$

$$\Pi^* = \begin{matrix} x_t^{(0)} & x_t^{(1)} & x_t^{(2)} & x_t^{(3)} & x_t^{(4)} \end{matrix}\begin{pmatrix} \dfrac{\alpha_0}{1-\alpha_1} & \dfrac{-\alpha_1}{1-\alpha_1} & \dfrac{\alpha_1\beta_2}{1-\alpha_1} & \dfrac{\alpha_1}{1-\alpha_1} & \dfrac{\alpha_1\beta_1}{1-\alpha_1} \\ 0 & 0 & \beta_2 & 0 & \beta_1 \end{pmatrix} \qquad \varepsilon_t^* = \begin{pmatrix} \dfrac{1}{1-\alpha_1}\delta_t^{(1)} + \dfrac{\alpha_1}{1-\alpha_1}\delta_t^{(2)} \\ \delta_t^{(2)} \end{pmatrix}$$

Этапы эконометрического исследования

1. *Постановочный* (цели исследования, выбор показателей).
2. *Априорный* (предмодельный анализ сущности явления).
3. *Параметризация* (выбор модели, формы связей).
4. *Информационный* (сбор статистической информации).
5. *Идентификация модели* (статистическое оценивание параметров).
6. *Верификация модели* (сопоставление реальных и модельных данных, оценка точности).
7. *Прогнозирование и управление*.

Спецификация модели (этапы 1–3)

На этапах спецификации модели часто формулируются априорные ограничения относительно значений отдельных элементов матриц B и C, а также стохастической природы остатков Δ_t!

1. $M\delta_t^{(j)} \equiv 0$ – остатки имеют нулевые средние значения.
2. $M\left(\delta_t^{(i)}, \delta_t^{(j)}\right) = 0,\ i \neq j$ – остатки не коррелируют друг с другом.
3. $M\left(\delta_{t_1}^{(j)}, \delta_{t_2}^{(j)}\right) = 0,\ t_1 \neq t_2$ – остатки не имеют автокорреляций.
4. $D\delta_t^{(j)} = const$ – гомоскедастичность (постоянство дисперсии) остатков.

Часто нельзя сказать что-то определенное о свойствах остатков ε_t^*. Это создает существенные трудности в статистическом анализе уравнений приведенной формы.

Идентификация модели

Уравнение структурной формы называется:

1. *Точно идентифицируемым* – если все участвующие в нем неизвестные коэффициенты однозначно восстанавливаются по коэффициентам приведенной формы.
2. *Сверхидентифицируемым* – если все участвующие в нем неизвестные коэффициенты восстанавливаются по коэффициентам приведенной формы, причем некоторые из них могут принимать одновременно несколько значений.
3. *Неидентифицируемым* – если хотя бы один из участвующих в нем неизвестных коэффициентов не может быть восстановлен по коэффициентам приведенной формы.

Эконометрическая модель называется *точно идентифицируемой*, если все уравнения ее структурной формы являются точно идентифицируемыми.

Эконометрическая модель называется *неидентифицируемой*, если хотя бы одно уравнение ее структурной формы является неидентифицируемым.

Методы идентификации

1. *Метод наименьших квадратов* (линейная модель множественной регрессии с гомоскедастичными неавтокоррелированными остатками – классическая модель).
2. *Взвешенный метод наименьших квадратов* (линейная модель множественной регрессии с гетероскедастичными остатками).
3. *Обобщенный метод наименьших квадратов* (линейная модель множественной регрессии с автокоррелированными остатками; с более сложной взаимозависимостью остатков).

Методы верификации

Наиболее распространен *метод ретроспективных расчетов*:

Обучающая выборка: $\left\{\begin{pmatrix} Y_1^T \\ \dots \\ Y_{n-k}^T \end{pmatrix} \begin{pmatrix} X_1^T \\ \dots \\ X_{n-k}^T \end{pmatrix}\right\}$.

Экзаменующая выбора: $\left\{\begin{pmatrix} Y_{n-k+1}^T \\ \dots \\ Y_n^T \end{pmatrix} \begin{pmatrix} X_{n-k+1}^T \\ \dots \\ X_n^T \end{pmatrix}\right\}$.

По обучающей выборке оцениваем коэффициенты модели, делаем прогноз для экзаменующей выборки, сравниваем с наблюдаемыми значениями.

Регрессионный анализ

Этимология: «регрессия» – отступление, возврат

x – отклонение от среднего роста отца
y – отклонение от среднего роста сына
$\Leftarrow$ Положительная связь, но тенденция возврата (отклонение у сына < отклонения у отца)

Закономерность – функция регрессии.

Классическая линейная модель множественной регрессии (КЛММР)

$y_i = \theta_0 + \theta_1 x_i^{(1)} + ... + \theta_p x_i^{(p)} + \varepsilon_i,\ i = 1,...,n.$

1) $M\varepsilon_i = 0,\ i = 1,...,n;$

2) $M(\varepsilon_i \varepsilon_j) = \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases}$ – гомоскедастичность; – взаимная некоррелированность;

3) $rank\ X = p+1 < n$ – одна объясняющая переменная не выражается через другую, существует $(X^T X)^{-1}$, если $p+1 \geq n$, то для выводов недостаточно данных.

Матричная форма:

$$Y = \begin{pmatrix} y_1 \\ ... \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} x_1^{(0)} = 1 & x_1^{(1)} & ... & x_1^{(p)} \\ & ... & & \\ x_n^{(0)} = 1 & x_n^{(1)} & ... & x_n^{(p)} \end{pmatrix} \quad \Theta = \begin{pmatrix} \theta_0 \\ ... \\ \theta_p \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ ... \\ \varepsilon_n \end{pmatrix} \quad \sum_\varepsilon = \begin{pmatrix} M(\varepsilon_1^2) & ... & M(\varepsilon_1 \varepsilon_n) \\ & ... & \\ M(\varepsilon_n \varepsilon_1) & ... & M(\varepsilon_n^2) \end{pmatrix}.$$

$Y = X\Theta + \varepsilon, \quad M\varepsilon = \mathbf{0_n}, \quad \sum_\varepsilon = \sigma^2 E_n, \quad rank\ X = p+1 < n.$

↑ ковариационная матрица остатков

Если $\varepsilon \sim N(0; \sigma^2 E_n)$, то нормальная КЛММР.

Оценивание неизвестных параметров. Метод наименьших квадратов (МНК)

Модельные значения $\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_i^{(1)} + ... + \hat{\theta}_p x_i^{(p)}$ минимально отличаются от наблюдаемых y_i:

$$\sum_{i=1}^n \varepsilon_i^2 \to \min_{\theta_0,...,\theta_p}, \quad \varepsilon_i = y_i - \theta_0 - \theta_1 x_i^{(1)} - ... - \theta_p x_i^{(p)}.$$

Матричная форма:

$\varepsilon = Y - X\Theta, \quad (Y - X\Theta)^T (Y - X\Theta) \to \min_\Theta,$ $\boxed{(AB)^T = B^T A^T}$

$Y^T Y - 2\Theta^T X^T Y + \Theta^T X^T X\Theta \to \min_\Theta, \quad -2X^T Y + 2X^T X\Theta = 0,$

$\boxed{\hat{\Theta} = (X^T X)^{-1} X^T Y}$

↑ невырожденная

Случай парной регрессии:

$$X = \begin{pmatrix} 1 & x_1 \\ & ... \\ 1 & x_p \end{pmatrix} \quad X^T X = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \quad X^T Y = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}.$$

$$\begin{cases} n\theta_0 + \theta_1 \sum x_i = \sum y_i, \\ \theta_0 \sum x_i + \theta_1 \sum x_i^2 = \sum x_i y_i. \end{cases} \qquad \theta_0 = \frac{\sum y_i - \theta_1 \sum x_i}{n},$$

$\sum x_i \sum y_i - \theta_1 (\sum x_i)^2 + \theta_1 n \sum x_i^2 = n \sum x_i y_i,$ $\quad \boxed{\hat{\theta}_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}}, \quad \boxed{\hat{\theta}_0 = \bar{y} - \theta_1 \bar{x}}$

Помесячные данные о печати фотографий в некоторой фирме

	месяц	***y*, кол-во, шт.**	$z^{(1)}$**, цена, руб.**	$z^{(2)}$**, рекл., руб.**	$z^{(3)}$**, праздники**	$z^{(i)}$**, индекс цен**
1	январь, 2003	12 500	2,5	0	3	1
2	февраль	7 600	3	0	1	0,99
3	март	6 900	3	0	1	1,01
4	апрель	13 500	3	5 000	0	1,01
5	май	9 700	3	0	3	1,03
6	июнь	10 700	3	2 000	1	1,04
7	июль	12 100	3	2 000	0	1,05
8	август	9 700	3,5	2 000	0	1,03
9	сентябрь	7 000	4	2 000	0	1,05
10	октябрь	7 200	4	2 000	0	1,05
11	ноябрь	8 200	4	2 000	1	1,06
12	декабрь	8 400	4	2 000	1	1,1
13	январь, 2004	13 100	4	2 000	3	1,11
14	февраль	8 700	4	0	1	1,12
15	март	12 200	4	5 000	1	1,14
16	апрель	6 900	4	0	0	1,16
17	май	6 200	4	0	3	1,17
18	июнь	9 600	4	0	1	1,19
19	июль	8 700	4	0	0	1,18
20	август	11 900	4	4 000	0	1,18
21	сентябрь	12 600	4	6 000	0	1,2
22	октябрь	7 900	4	1 000	0	1,22
23	ноябрь	9 300	4	2 000	1	1,24
24	декабрь	11 800	4	2 000	1	1,27

	y	$x^{(1)} = z^{(1)} / z^{(i)}$	$x^{(2)} = z^{(2)} / z^{(i)}$	$x^{(3)} = z^{(3)}$	$\hat{y}$**, прогноз**	ε**, остатки**
1	12 500	2,50	0	3	11 552	948
2	7 600	3,03	0	1	8 792	–1 192
3	6 900	2,97	0	1	8 968	–2 068
4	13 500	2,97	4 950	0	13 975	–475
5	9 700	2,91	0	3	10 339	–639
6	10 700	2,88	1 923	1	11 398	–698
7	12 100	2,86	1 905	0	10 857	1 243
8	9 700	3,40	1 942	0	9 310	390
9	7 000	3,81	1 905	0	8 059	–1 059
10	7 200	3,81	1 905	0	8 059	–859
11	8 200	3,77	1 887	1	8 745	–545
12	8 400	3,64	1 818	1	9 071	–671
13	13 100	3,60	1 802	3	10 350	2 750
14	8 700	3,57	0	1	7 202	1 498
15	12 200	3,51	4 386	1	12 354	–154
16	6 900	3,45	0	0	6 963	–63
17	6 200	3,42	0	3	8 852	–2 652
18	9 600	3,36	0	1	7 819	1 781
19	8 700	3,39	0	0	7 135	1 565
20	11 900	3,39	3 390	0	10 974	926
21	12 600	3,33	5 000	0	12 964	–364
22	7 900	3,28	820	0	8 390	–490
23	9 300	3,23	1 613	1	10 044	–744
24	11 800	3,15	1 575	1	10 225	1 575

$$X^T X = \begin{pmatrix} 24 & 79{,}23 & 36820 & 22 \\ 79{,}23 & 264{,}18 & 123426 & 70{,}42 \\ 36820 & 123426 & 114453587 & 18607 \\ 22 & 70{,}42 & 18607 & 46 \end{pmatrix}, \quad \left(X^T X\right)^{-1} = \begin{pmatrix} 4{,}83 & -1{,}39 & -0{,}00003 & -0{,}17 \\ -1{,}39 & 0{,}414 & -0{,}000005 & 0{,}0326 \\ -0{,}00003 & -0{,}000005 & 0{,}00000002 & 0{,}00001 \\ -0{,}17 & 0{,}0326 & 0{,}00001 & 0{,}0483 \end{pmatrix},$$

$$X^T Y = \begin{pmatrix} 232400 \\ 760351 \\ 407592667 \\ 217900 \end{pmatrix}, \quad \hat{\Theta} = \begin{pmatrix} 17094 \\ -2938 \\ 1{,}1327 \\ 600{,}86 \end{pmatrix}, \qquad \hat{y}_i = 17094 - 2938 x_i^{(1)} + 1{,}1327 x_i^{(2)} + 600{,}86 x_i^{(3)}.$$

Свойства оценок

На разных выборках – разные оценки (за счет случайного характера остатков)

1. **Состоятельность** (при росте выборки оценка $\hat{\theta}$ стремится к истинному значению θ) (асимптотическое свойство, часто проявляется при огромных размерах выборок).

   ## Симметрично распределенная случайная величина ξ, истинное среднее $\theta = M\xi$:

   $\hat{\theta}_1 = \bar{x}$, $\hat{\theta}_2 = \left(x_{\min} + x_{\max}\right)/2$ – состоятельные оценки.

   Состоятельная оценка может быть сколь угодно далекой от истинного значения

   ## Средняя зарплата в отрасли, где работает N человек (заинтересованы завысить):

   $\hat{\theta} = \begin{cases} \theta_0, & n < N \\ \bar{x}, & n = N \end{cases}$ – при любом объеме выборки n, кроме сплошного обследования, получаем сколь угодно завышенный результат θ_0.

2. **Несмещенность** ($M\hat{\theta} = \theta$ при любом объеме выборки)

   усреднение по всем выборкам данного объема

   (характеристика «хороших свойств» оценки при каждом конечном объеме выборки).

$$M\hat{\Theta} = M\left(\left(X^TX\right)^{-1}X^TY\right) = M\left(\left(X^TX\right)^{-1}X^T(X\Theta+\varepsilon)\right) = M\left(\Theta + \left(X^TX\right)^{-1}X^T\varepsilon\right) =$$

$$= M\Theta + \left(X^TX\right)^{-1}X^TM\varepsilon = M\Theta = \Theta \text{ – несмещенная.}$$

$$M\hat{\sigma}^2 = M\left(\frac{1}{n}\left(Y-\hat{Y}\right)^T\left(Y-\hat{Y}\right)\right)$$

$$Y-\hat{Y} = X\Theta+\varepsilon-X\hat{\Theta} = X\Theta+\varepsilon-X\left(X^TX\right)^{-1}X^T(X\Theta+\varepsilon) = X\Theta+\varepsilon-X\Theta-X\left(X^TX\right)^{-1}X^T\varepsilon =$$

$$= \left(E_n - X\left(X^TX\right)^{-1}X^T\right)\varepsilon = Z\varepsilon.$$

$Z^T = Z$ – симметричная, $Z^2 = ZZ = Z$ – идемпотентная, $M\left(\varepsilon^TZ\varepsilon\right) = \sigma^2 tr\,Z$

след матрицы Z (сумма ее диагональных элементов)

$$M\hat{\sigma}^2 = M\left(\frac{1}{n}\left(Y-\hat{Y}\right)^T\left(Y-\hat{Y}\right)\right) = M\left(\frac{1}{n}(Z\varepsilon)^TZ\varepsilon\right) = \frac{1}{n}M\left(\varepsilon^TZ^TZ\varepsilon\right) = \frac{1}{n}M\left(\varepsilon^TZ\varepsilon\right) = \frac{1}{n}\sigma^2 tr\,Z =$$

$$= \frac{\sigma^2}{n}tr\left(E_n - X\left(X^TX\right)^{-1}X^T\right) = \frac{\sigma^2}{n}\left(tr\,E_n - tr\left(\left(X^TX\right)^{-1}X^TX\right)\right) = \frac{\sigma^2}{n}\left(n-(p+1)\right) \text{ – смещенная.}$$

Несмещенная оценка дисперсии остатков:

$$\hat{\sigma}^2 = \frac{n}{n-p-1}\hat{\sigma}^2_{ММП} = \frac{1}{n-p-1}\sum_{i=1}^{n}\left(y_i - \hat{\theta}_0 - \hat{\theta}_1x_i^{(1)} - \ldots - \hat{\theta}_px_i^{(p)}\right)^2 = \frac{1}{n-p-1}\left(Y-X\hat{\Theta}\right)^T\left(Y-X\hat{\Theta}\right).$$

$\hat{\sigma}^2 = \frac{1}{24-3-1}\left(948^2 + (-1192)^2 + \ldots + 1575^2\right) = 1938734$, $\hat{\sigma} = \sqrt{1938734} = 1392{,}4$.

Ковариационная матрица оценок параметров:

$$\hat{\Sigma}_{\hat{\Theta}} = M\left(\left(\hat{\Theta}-\Theta\right)\left(\hat{\Theta}-\Theta\right)^T\right) = \hat{\sigma}^2\left(X^TX\right)^{-1}.$$

$$\hat{\Sigma}_{\hat{\Theta}} = 1938734\begin{pmatrix} 4{,}83 & -1{,}39 & -0{,}00003 & -0{,}17 \\ -1{,}39 & 0{,}414 & -0{,}000005 & 0{,}0326 \\ -0{,}00003 & -0{,}000005 & 0{,}00000002 & 0{,}00001 \\ -0{,}17 & 0{,}0326 & 0{,}00001 & 0{,}0483 \end{pmatrix} = \begin{pmatrix} 9355484 & -2694758 & -50{,}15 & -328915 \\ -2694758 & 803096 & -9{,}44 & 63228 \\ -50{,}15 & -9{,}44 & 0{,}0396 & 22{,}41 \\ -328915 & 63228 & 22{,}41 & 93598 \end{pmatrix}.$$

Наиболее важны диагональные элементы – квадраты среднеквадратических ошибок s_l оценок коэффициентов θ_l, $l = 0,1,\ldots,p$.

$s_0 = \sqrt{9335484} = 3059$, $s_1 = \sqrt{803096} = 896$, $s_2 = \sqrt{0{,}0396} = 0{,}1990$, $s_3 = \sqrt{93598} = 305{,}94$.

$$\hat{y}_i = \underset{(3059)}{17094} - \underset{(896)}{2938}x_i^{(1)} + \underset{(0{,}1990)}{1{,}1327}x_i^{(2)} + \underset{(305{,}94)}{600{,}86}x_i^{(3)}.$$

3. **Эффективность** (оценка обладает наименьшим случайным разбросом).

$M\left(\hat{\theta}_{eff}-\theta\right)^2=\min\limits_{\hat{\theta}\in M}\left(\hat{\theta}-\theta\right)^2$ – оценка, эффективная в классе M.

Свойства оценок, справедливые для нормальной КЛММР

$t_l=\dfrac{\hat{\theta}_l-\theta_l}{s_l}\sim t\left(n-p-1\right)$ – распределена по закону Стьюдента.

Проверка гипотезы H_0: $\theta_l=\theta_l^0$, в т.ч. о значимости регрессора (при $\theta_l^0=0$):

1) Задаем уровень значимости α.
2) Находим эмпирическую точку $t_l=\left(\hat{\theta}_l-\theta_l^0\right)\big/s_l$.
3) Находим критическую точку $t_{\text{крит}}=t_\alpha\left(n-p-1\right)=СТЬЮДРАСПОБР\left(\alpha;\ n-p-1\right)$.
4) Если $\left|t_l\right|>t_{\text{крит}}$, то H_0 – отвергается, иначе – принимается.

$\theta_0=0$: $t_0=\dfrac{17094}{3059}=5{,}59$; $\theta_1=0$: $t_1=\dfrac{-2938}{896}=-3{,}28$; $\theta_2=0$: $t_2=\dfrac{1{,}1327}{0{,}1990}=5{,}69$; $\theta_3=0$: $t_3=\dfrac{600{,}86}{305{,}94}=1{,}96$

$t_{\text{крит}}=t_{0{,}05}(20)=2{,}09$

Гипотеза $H_0:\theta_l=0$ отвергается для $\theta_0,\theta_1,\theta_2$ и принимается для θ_3 при $\boxed{\alpha=0{,}05}$,
регрессор $x^{(3)}$ – незначим

При $\boxed{\alpha=0{,}1}$ $t_{\text{крит}}=1{,}72$, $H_0:\theta_l=0$ отвергается для всех θ_l, все регрессоры значимы.

При $\boxed{n=1000}$ $t_{\text{крит}}=1{,}96$, $H_0:\theta_l=0$ отвергается для всех θ_l, все регрессоры значимы.

Построение доверительного интервала для θ_l:

1. Задаем доверительную вероятность γ.
2. $\theta_l\in\left[\hat{\theta}_l-t_{1-\gamma}\left(n-p-1\right)s_l;\ \hat{\theta}_l+t_{1-\gamma}\left(n-p-1\right)s_l\right]$.

$\theta_0\in\left[10714;23474\right]$, $\theta_1\in\left[-4807;-1069\right]$, $\theta_2\in\left[0{,}7176;1{,}5478\right]$, $\theta_3\in\left[-37;1239\right]$ при $\gamma=0{,}95$.

Проверка гипотезы $H_0:\ R^2=0$ о значимости модели:

1. Задаем уровень значимости α.
2. Находим эмпирическую точку $F_{\text{эмп}}=\dfrac{\hat{R}^2_{y.X}}{1-\hat{R}^2_{y.X}}\dfrac{n-p-1}{p}$.
3. Находим критическую точку $F_{\text{крит}}=F_\alpha\left(p;n-p-1\right)=FРАСПОБР\left(\alpha;\ p;\ n-p-1\right)$.
4. Если $F_{\text{эмп}}>F_{\text{крит}}$, то H_0 – отвергается, иначе – принимается, линейная модель неадекватна.

$\hat{R}^2_{y.X}$ – квадрат множественного коэффициента корреляции, равен коэффициенту детерминации.

Способы расчета:

1. $\hat{R}^2_{y.X}=1-\dfrac{\left|\hat{R}\right|}{\left|\hat{R}\right|_{00}}$, $\hat{R}=\begin{pmatrix}1 & -0{,}389 & 0{,}613 & 0{,}088\\ -0{,}389 & 1 & 0{,}152 & -0{,}269\\ 0{,}613 & 0{,}152 & 1 & -0{,}391\\ 0{,}088 & -0{,}269 & -0{,}391 & 1\end{pmatrix}$, $\hat{R}^2_{y.X}=1-\dfrac{0{,}2536}{0{,}7832}=0{,}6762$.

2. $\hat{R}^2_{y.X}=\hat{K}_d\left(y.X\right)=1-\dfrac{S^2_\varepsilon}{S^2_y}=1-\dfrac{\frac{1}{24-3-1}\left(948^2+(-1192)^2+...+1575^2\right)}{\frac{1}{24-3-1}\left(2817^2+(-2083)^2+...+2117^2\right)}=1-\dfrac{1938734}{5986667}=0{,}6762$.

$F_{\text{эмп}}=\dfrac{0{,}6762}{1-0{,}6762}\dfrac{24-3-1}{3}=13{,}92$, $F_{\text{крит}}=F_{0{,}05}\left(3;20\right)=3{,}10$, H_0 отвергается, линейная модель значима при $\alpha=0{,}05$.

Ошибки спецификации модели

1. **Неправомерно исключены некоторые объясняющие переменные:**
   1) смещены оценки коэффициентов регрессии
   2) еще сильнее смещена оценка дисперсии остатков σ^2

   $\Rightarrow$ неверные выводы

В пример введена дополнительная переменная $z^{(4)}$ – цена конкурента.

$z^{(4)} = z^{(1)}$ во все периоды кроме 4 месяцев:

$z_2^{(4)} = z_3^{(4)} = 2{,}5$ (февраль-март, 2003), $z_7^{(4)} = 3{,}5$ (июль, 2003), $z_{17}^{(4)} = 3$ (май, 2004).

конкурент позже поднял цены раньше поднял скидки

Новая модель:

$$\hat{y}_i = \underset{(2292)}{15987} - \underset{(1118)}{6648}\, x_i^{(1)} + \underset{(0{,}1522)}{0{,}9885}\, x_i^{(2)} + \underset{(239{,}98)}{914{,}19}\, x_i^{(3)} + \underset{(990)}{4096}\, x_i^{(4)}, \qquad \hat{R}^2_{y.X} = 0{,}8296.$$

Старая модель:

$$\hat{y}_i = \underset{(3059)}{17094} - \underset{(896)}{2938}\, x_i^{(1)} + \underset{(0{,}1990)}{1{,}1327}\, x_i^{(2)} + \underset{(305{,}94)}{600{,}86}\, x_i^{(3)}, \qquad \hat{R}^2_{y.X} = 0{,}6762.$$

2. **В модель введены лишние несущественные переменные:**

   Меньшее из зол, но при увеличении числа переменных:
   1) ослабевает точность выводов, зависящая от $n/(p+1)$;
   2) возможно появление мультиколлинеарности
      – взаимозависимости объясняющих переменных.

Мультиколлинеарность

1. **Полная мультиколлинеарность** – линейная функциональная связь между объясняющими переменными, хотя бы одна из них линейно выражается через остальные:

   $rank\, X < p+1$, $X^T X$ – вырожденная, $\left(X^T X\right)^{-1}$ – не существует.

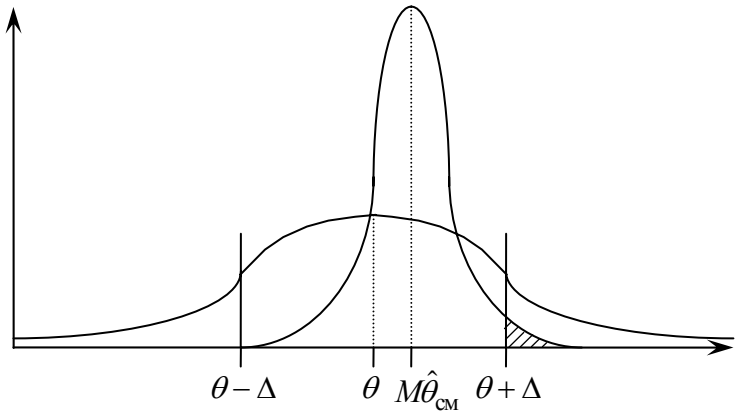
   Избежать легко – на этапе отбора объясняющих переменных.

2. **Частичная мультиколлинеарность** – тесная, но не функциональная связь между объясняющими переменными.

   Эвристические рекомендации для выявления частичной мультиколлинеарности:
   1) *Анализ корреляционной матрицы R*; $\left|r_{jk}\right| > 0{,}8$.
   2) *Анализ обусловленности матрицы* $X^T X$; $\det\left(X^T X\right) \approx 0$.
   3) *Анализ собственных чисел матрицы* $X^T X$; $\lambda_{\min} \approx 0$.
   4) *Анализ коэффициентов детерминации каждой переменной* $x^{(j)}$ по всем остальным объясняющим переменным; $\left|\hat{R}^2_{j(-j)}\right| > 0{,}9$.
   5) *Анализ экономической сущности модели*; некоторые оценки $\hat{\theta}_j$ имеют неверные с точки зрения экономической теории значения (неверные знаки, слишком большие или слишком малые значения).
   6) *Анализ чувствительности модели*; небольшое изменение данных (добавление или изъятие небольшой порции наблюдений) существенно изменяет оценки $\hat{\theta}_j$ коэффициентов модели (вплоть до изменения знаков).
   7) *Анализ значимости модели*; большинство (или даже все) оценки $\hat{\theta}_j$ коэффициентов модели статистически неотличимы от 0, в то время как модель в целом является значимой.

Методы устранения мультиколлинеарности

1. Переход к смещенным методам оценивания



смещенная оценка может быть более точной, чем несмещенная!

возможные значения оценок $\hat{\theta}$ на разных выборках

Один из методов – «*ридж-регрессия*» (ridge – гребень):

$\hat{\Theta}=\left(X^TX+\tau E_{p+1}\right)^{-1}X^TY$ – добавляем к диагональным элементам матрицы X^TX «гребень»

$\tau\in\left(0{,}1;\,0{,}4\right)$, матрица становится хорошо обусловленной.

2. Метод главных компонент – переход к новым объясняющим переменным, линейным комбинациям старых:

1) Центрирование переменных $X_ц = X-\overline{X}$, $Y_ц = Y-\overline{Y}$;
2) Решение характеристического уравнения $\left|\sum-\lambda E\right|=0$:
   a) Нахождение собственных чисел $\lambda_1\ge\lambda_2\ge...\ge\lambda_p>0$,
   b) Нахождение для каждого собственного числа λ_j собственного вектора $l^{(j)}$;
3) Переход к новым переменным $Z=X_ц L$, $X_ц = ZL^{-1}=ZL^T$;
4) Построение линейной регрессии $Y_ц = ZC$, вычисление оценок с помощью МНК $\hat{C}=\left(Z^TZ\right)^{-1}Z^TY=diag\left\{1/\lambda_j\right\}Z^TY$;
5) Проверка гипотез H_{0j}: $c_j=0$, $j=1,...,p$, исключение несущественных переменных;
6) При необходимости переход к исходной модели $\hat{\theta}_j=\sum\limits_{k\in K_{сущ}}\hat{c}_k l_k^{(j)}$, $\hat{\theta}_0=\overline{y}-\sum\limits_{j=1}^{p}\hat{\theta}_j\overline{x}^{(j)}$.

3. Отбор наиболее существенных объясняющих переменных

1) *Версия «всех возможных регрессий»*:

   Для заданного $k=1,2,...,p-1$ находится набор переменных $x^{(j_1)},...,x^{(j_k)}$, дающий максимальное значение коэффициента детерминации $\hat{R}^2(k)$.

   Увеличиваем число переменных *k*, пока возрастает нижняя граница ~95%-доверительного интервала для коэффициента детерминации.

   $$R^2_{\min}(k)=\hat{R}^2_{несм}(k)-2\sqrt{\frac{2k(n-k-1)}{(n-1)\left(n^2-1\right)}}\left(1-\hat{R}^2(k)\right),\quad \hat{R}^2_{несм}(k)=1-\left(1-\hat{R}^2(k)\right)\frac{n-k}{n-k-1}.$$

   Проблема: огромное количество переборов (для 20 переменных – более 1млн).

2) *Версия «пошагового отбора переменных»*:

   При переходе от *k* переменных к $(k+1)$ переменной учитываются результаты предыдущего шага – все отобранные переменные остаются.

   Проблема: нет гарантии получения оптимума.

Обобщенная линейная модель множественной регрессии (ОЛММР)

$Y = X\Theta + \varepsilon, \quad M\varepsilon = \mathbf{0_n}, \quad \sum_{\varepsilon} = \sigma^2\Sigma_0, \quad rank\,X = p+1 < n.$

σ^2 – неизвестная положительная константа, Σ_0 – известная матрица, $M(\varepsilon_i\varepsilon_j) = \sigma_{ij}$.

Частные случаи

1. Линейная модель с гетероскедастичными остатками:

Модель «доход–сбережения».

Постоянство относительного, а не абсолютного разброса регрессионных остатков $\varepsilon(x)$.

$\sum_{\varepsilon}$ – диагональная матрица, дисперсия остатков зависит от значений регрессоров.

$$\sum_{\varepsilon} = \sigma^2 \begin{pmatrix} 1/\lambda_1 & 0 & \dots & 0 \\ 0 & 1/\lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1/\lambda_n \end{pmatrix}$$

λ_i – оцениваются, исходя из вида зависимости $D(\varepsilon : X)$ от X

2. Линейная модель с автокоррелированными остатками:

Данные регистрируются во времени

$\sum_{\varepsilon}$ – не диагональная матрица, регрессионные остатки взаимозависимы

$r(\varepsilon_i, \varepsilon_j) = \rho^{|i-j|}, \quad |\rho| < 1$ – коэффициент корреляции между соседними остатками

1) Корреляционная связь зависит только от разнесенности во времени, $r(\varepsilon_1, \varepsilon_5) = r(\varepsilon_3, \varepsilon_7)$.

2) Корреляционная связь исчезает при $|i-j| \to \infty$.

3) $$\sum_{\varepsilon} = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix}.$$

Обобщенный метод наименьших квадратов (ОМНК)

МНК-оценки – состоятельные и несмещенные, но теряют свойство эффективности!

$\boxed{\hat{\Theta}_{ОМНК} = \left(X^T\Sigma_0^{-1}X\right)^{-1}\left(X^T\Sigma_0^{-1}Y\right)}$ – состоятельные, несмещенные и эффективные

$\Sigma_0 = CC^T, \quad \Sigma_0^{-1} = \left(CC^T\right)^{-1} = \left(C^{-1}\right)^T C^{-1},$

Переходим к преобразованной модели:

$C^{-1}Y = C^{-1}X\Theta + C^{-1}\varepsilon, \quad \Rightarrow \quad \widetilde{Y} = C^{-1}Y, \ \widetilde{X} = C^{-1}X, \ \widetilde{\varepsilon} = C^{-1}\varepsilon.$

$\boxed{\widetilde{Y} = \widetilde{X}\Theta + \widetilde{\varepsilon}}$ – классическая модель, так как

$$\sum_{\widetilde{\varepsilon}} = M\left(\widetilde{\varepsilon}\widetilde{\varepsilon}^T\right) = M\left(C^{-1}\varepsilon\varepsilon^T\left(C^{-1}\right)^T\right) = C^{-1}\sigma^2\Sigma_0\left(C^{-1}\right)^T = \sigma^2 C^{-1}\Sigma_0\left(C^{-1}\right)^T = \sigma^2 E_n.$$

Оценки преобразованной модели:

$$\hat{\widetilde{\Theta}} = \left(\widetilde{X}^T\widetilde{X}\right)^{-1}\widetilde{X}^T\widetilde{Y} = \left(X^T\left(C^{-1}\right)^T C^{-1}X\right)^{-1} X^T\left(C^{-1}\right)^T C^{-1}Y = \left(X^T\Sigma_0^{-1}X\right)^{-1}X^T\Sigma_0^{-1}Y = \hat{\Theta}_{ОМНК}.$$

Ковариационная матрица оценок:

$$\Sigma_{\hat{\Theta}} = \sigma^2\left(\widetilde{X}^T\widetilde{X}\right)^{-1} = \sigma^2\left(X^T\left(C^{-1}\right)^T C^{-1}X\right)^{-1} = \sigma^2\left(X^T\Sigma_0^{-1}X\right)^{-1}.$$

Критерий обобщенного метода наименьших квадратов:

$$\left(\tilde{Y}-\tilde{X}\Theta\right)^T\left(\tilde{Y}-\tilde{X}\Theta\right)\to\min_{\Theta},$$

$$\left(Y-X\Theta\right)^T\left(C^{-1}\right)^T C^{-1}\left(Y-X\Theta\right)=\left(Y-X\Theta\right)^T\Sigma_0^{-1}\left(Y-X\Theta\right)\to\min_{\Theta}.$$

Дисперсия регрессионных остатков:

σ^2 – уже не является, как в классической модели, дисперсией регрессионных остатков.

Можно умножить Σ_0 на любую константу, тогда σ^2 разделится на нее.

σ^2 – является неизвестным параметром модели, который нужно оценить:

$$\hat{\sigma}^2=\frac{1}{n-p-1}\left(Y-X\hat{\Theta}_{ОМНК}\right)^T\Sigma_0^{-1}\left(Y-X\hat{\Theta}_{ОМНК}\right).$$

Проблема практической реализации:

Матрица Σ_0 – неизвестна в подавляющем большинстве случаев

включить ее элементы в число параметров нельзя, так как их число $n(n+1)/2$ превышает объем данных np	$\Rightarrow$	необходимо наложить ограничения на вид Σ_0

Обобщенная модель с гетероскедастичными остатками
Взвешенный МНК (ВМНК)

$$\sum\nolimits_0=\begin{pmatrix}1/\lambda_1 & 0 & \dots & 0\\ 0 & 1/\lambda_2 & \dots & 0\\ \dots & \dots & \dots & \dots\\ 0 & 0 & \dots & 1/\lambda_n\end{pmatrix}$$

1) Остатки взаимно некоррелированы, $M\left(\varepsilon_i,\varepsilon_j\right)=0,\ i\neq j$.
2) Остатки не обладают постоянной дисперсией, $D\varepsilon_i\neq D\varepsilon_j, i\neq j$.
3) По диагонали матрицы Σ_0 стоят дисперсии, $1/\lambda_i=D\varepsilon_i$.

Гетероскедастичные остатки – естественная ситуация для пространственных выборок!

Часто выполняется гипотеза постоянства относительного разброса остатков в вариантах:

1) $D(\varepsilon:X_i)=\sigma^2\left(M(y:X_i)\right)^2$ – от значения результирующей переменной;
2) $D(\varepsilon:X_i)=\sigma^2\left(a+bx_i^{(j)}\right)^2$ – от значения j-объясняющей переменной.

Критерий взвешенного метода наименьших квадратов:

$$\sum_{i=1}^{n}\lambda_i\left(y_i-\theta_0-\theta_1x_i^{(1)}-\dots-\theta_px_i^{(p)}\right)^2\to\min_{\theta_0,\theta_1,\dots,\theta_p}$$ – чем больше разброс, тем меньше вес.

Дисперсия регрессионных остатков:

$$\hat{\sigma}^2=\frac{1}{n-p-1}\sum_{i=1}^{n}\lambda_i\left(y_i-\hat{\theta}_0-\hat{\theta}_1x_i^{(1)}-\dots-\hat{\theta}_px_i^{(p)}\right)^2.$$

Сравнение ВМНК- и МНК-оценок по эффективности:

Доход-сбережения (40 человек, 4 группы с доходами 4, 6, 8 и 10 тыс. руб./мес.)

Несмещенные оценки среднеквадратических отклонений сбережений по каждой группе:

$s_1=633$, $s_2=1602$, $s_3=2245$, $s_4=2890$;

$1/\lambda_1=\dots=1/\lambda_{10}=633^2$, $1/\lambda_{11}=\dots=1/\lambda_{20}=1602^2$, $1/\lambda_{21}=\dots=1/\lambda_{30}=2245^2$, $1/\lambda_{31}=\dots=1/\lambda_{40}=2890^2$.

МНК: $\hat{y}_i=\underset{(1023,1)}{-2614,7}+\underset{(0,139)}{1,0991}x_i$,

$$\frac{D\hat{\theta}_{0.ВМНК}}{D\hat{\theta}_{0.МНК}}=\frac{560,5^2}{1023,1^2}=0,30=30\%.$$

ВМНК: $\hat{y}_i=\underset{(560,5)}{-2586,7}+\underset{(0,113)}{1,0949}x_i$,

$$\frac{D\hat{\theta}_{1.ВМНК}}{D\hat{\theta}_{1.МНК}}=\frac{0,113^2}{0,139^2}=0,66=66\%.$$

Проверка гипотезы о гомоскедастичности

1. Случай сгруппированных данных и больших выборок

$n = n_1 + n_2 + ... + n_k$, $n_j > 3$ – внутри каждой группы значения объясняющих переменных совпадают или принадлежат одному интервалу.

$H_0:\ D\left(\varepsilon : X_1^0\right) = D\left(\varepsilon : X_2^0\right) = ... = D\left(\varepsilon : X_k^0\right)$.

Критерий Бартлетта:

5) Задаем уровень значимости α;

6) Находим несмещенные оценки дисперсий для каждой группы $s_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}\left(y_{ji} - \bar{y}_j\right)^2$;

7) Находим эмпирическую точку

$$X^2_{\text{эмп}} = \frac{1}{q}\sum_{j=1}^{k}\left(n_j - 1\right)\ln\left(s^2 / s_j^2\right),\ q = 1 + \frac{1}{3(k-1)}\left(\sum_{j=1}^{k}\frac{1}{n_j - 1} - \frac{1}{n-k}\right),\ s^2 = \frac{1}{n-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(y_{ji} - \bar{y}_j\right)^2;$$

8) Находим критическую точку

$X^2_{\text{крит}} = X^2_{\alpha}(k-1) = ХИ2ОБР(\alpha; k-1)$;

9) Если $X^2_{\text{эмп}} > X^2_{\text{крит}}$, то H_0 – отвергается (гетероскедастичность), иначе – принимается.

2. Критерий Глейсера:

Регрессия абсолютных величин остатков $\left|\hat{\varepsilon}_i\right| = \left| y_i - \hat{\theta}_{0.МНК} - \hat{\theta}_{1.МНК} x_i^{(1)} - ... - \hat{\theta}_{p.МНК} x_i^{(p)} \right|$ по некоторой функции от объясняющей переменной $x^{(j)}$.

Наиболее распространенные функции:

1) $M\left(\left|\hat{\varepsilon}\right| : x^{(j)}\right) = a_0 + a_1 x^{(j)}$

2) $M\left(\left|\hat{\varepsilon}\right| : x^{(j)}\right) = a_0 + \frac{a_1}{x^{(j)}}$

3) $M\left(\left|\hat{\varepsilon}\right| : x^{(j)}\right) = a_0 \left(x^{(j)}\right)^{a_1}$

$H_0:\ a_1 = 0$, если отвергается, то гетероскедастичность.

Также возможны варианты регрессии по нескольким объясняющим переменным или по результирующей переменной.

Случай $M\left(\left|\hat{\varepsilon}\right| : x^{(j)}\right) = a_0 + a_1 x^{(j)}$**:**

$$\varepsilon_i = \delta_i\left(\hat{a}_{0.МНК} + \hat{a}_{1.МНК} x_i^{(j)}\right),\quad M\delta_i = 0,\quad D\delta_i = \sigma^2,\quad 1/\lambda_i = \left(\hat{a}_{0.МНК} + \hat{a}_{1.МНК} x_i^{(j)}\right)^2$$

$$\sum\nolimits_0 = \begin{pmatrix} \left(\hat{a}_0 + \hat{a}_1 x_1^{(j)}\right)^2 & 0 & ... & 0 \\ 0 & \left(\hat{a}_0 + \hat{a}_1 x_2^{(j)}\right)^2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \left(\hat{a}_0 + \hat{a}_1 x_n^{(j)}\right)^2 \end{pmatrix}$$

Практическое оценивание модели с гетероскедастичными остатками

1. Проверка гипотезы о гетероскедастичности.
2. От модели $\boxed{y_i = \theta_0 x_i^{(0)} + \theta_1 x_i^{(1)} + ... + \theta_p x_i^{(p)} + \varepsilon_i}$ переходим к $\boxed{\widetilde{y}_i = \theta_0 \widetilde{x}_i^{(0)} + \theta_1 \widetilde{x}_i^{(1)} + ... + \theta_p \widetilde{x}_i^{(p)} + \widetilde{\varepsilon}_i}$.

$$M\left(\left|\hat{\varepsilon}\right| : x^{(j)}\right) = a_0 + a_1 x^{(j)} \Rightarrow \boxed{\widetilde{y}_i = \frac{y_i}{a_0 + a_1 x_i^{(j)}},\ \widetilde{x}_i^{(0)} = \frac{1}{a_0 + a_1 x_i^{(j)}},\ \widetilde{x}_i^{(1)} = \frac{x_i^{(1)}}{a_0 + a_1 x_i^{(j)}},\ ...,\ \widetilde{x}_i^{(p)} = \frac{x_i^{(p)}}{a_0 + a_1 x_i^{(j)}}}.$$

3. Оценивание коэффициентов $\hat{\theta}_0$, $\hat{\theta}_1$, …, $\hat{\theta}_p$ с помощью МНК, проверка значимости регрессоров и значимости преобразованной модели.

Оценивание коэффициентов $\hat{\theta}_0$, $\hat{\theta}_1$, …, $\hat{\theta}_p$ в Excel происходит с учетом отсутствия свободного коэффициента, так как он уже включен в модель!

Обобщенная модель с автокоррелированными остатками

Статистические данные регистрируются во времени. Зависимость остатков неограниченно ослабевает по мере удаления их друг от друга

$$\boxed{\varepsilon_i = \rho\varepsilon_{i-1} + \delta_i},\ |\rho|<1,\ M\delta_i \equiv 0,\ M(\delta_i\delta_j) = \begin{cases}\sigma_0^2, i=j\\ 0,\ i\neq j\end{cases} -\ \text{автокорреляционная зависимость 1-порядка}$$

$$\varepsilon_i = \rho\varepsilon_{i-1}+\delta_i = \rho(\rho\varepsilon_{i-2}+\delta_{i-1})+\delta_i = \rho^2\varepsilon_{i-2}+\rho\delta_{i-1}+\delta_i = \rho^2(\rho\varepsilon_{i-3}+\delta_{i-2})+\rho\delta_{i-1}+\delta_i =$$

$$= \rho^3\varepsilon_{i-3}+\rho^2\delta_{i-2}+\rho\delta_{i-1}+\delta_i = \ldots = \delta_i+\rho\delta_{i-1}+\rho^2\delta_{i-2}+\ldots = \sum_{k=0}^{\infty}\rho^k\delta_{i-k},\ \boxed{M\varepsilon_i = \sum_{k=0}^{\infty}\rho^k M\delta_{i-k}\equiv 0}.$$

$$D\varepsilon_i = M\varepsilon_i^2 = M(\delta_i+\rho\delta_{i-1}+\rho^2\delta_{i-2}+\ldots)^2 = \sigma_0^2+\rho^2\sigma_0^2+\rho^4\sigma_0^2+\ldots = \frac{\sigma_0^2}{1-\rho^2} = \sigma^2.$$

$$M(\varepsilon_i\,\varepsilon_{i-k}) = M\big((\delta_i+\rho\delta_{i-1}+\rho^2\delta_{i-2}+\ldots)(\delta_{i-k}+\rho\delta_{i-k-1}+\rho^2\delta_{i-k-2}+\ldots)\big) =$$

$$= M\big((\rho^k\delta_{i-k}+\rho^{k+1}\delta_{i-k-1}+\rho^{k+2}\delta_{i-k-2}+\ldots)(\delta_{i-k}+\rho\delta_{i-k-1}+\rho^2\delta_{i-k-2}+\ldots)\big) = \frac{\rho^k\sigma_0^2}{1-\rho^2} = \rho^k\sigma^2.$$

$$\sum_0 = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \ldots & \rho^{n-2} & \rho^{n-1}\\ \rho & 1 & \rho & \rho^2 & \ldots & \rho^{n-3} & \rho^{n-2}\\ \rho^2 & \rho & 1 & \rho & \ldots & \rho^{n-4} & \rho^{n-3}\\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\ \rho^{n-2} & \rho^{n-3} & \rho^{n-4} & \rho^{n-5} & \ldots & 1 & \rho\\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \rho^{n-4} & \ldots & \rho & 1 \end{pmatrix},\ \sum_0^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho & 0 & 0 & \ldots & 0 & 0 & 0\\ -\rho & 1+\rho^2 & -\rho & 0 & \ldots & 0 & 0 & 0\\ 0 & -\rho & 1+\rho^2 & -\rho & \ldots & 0 & 0 & 0\\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\ 0 & 0 & 0 & 0 & \ldots & -\rho & 1+\rho^2 & -\rho\\ 0 & 0 & 0 & 0 & \ldots & 0 & -\rho & 1 \end{pmatrix},$$

$$C^{-1} = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & 0 & \ldots & 0 & 0\\ -\rho & 1 & 0 & \ldots & 0 & 0\\ 0 & -\rho & 1 & \ldots & 0 & 0\\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\ 0 & 0 & 0 & \ldots & -\rho & 1 \end{pmatrix},\ \left(C^{-1}\right)^T C^{-1} = \begin{pmatrix} 1 & -\rho & 0 & \ldots & 0 & 0\\ -\rho & 1+\rho^2 & -\rho & \ldots & 0 & 0\\ 0 & -\rho & 1+\rho^2 & \ldots & 0 & 0\\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots\\ 0 & 0 & 0 & \ldots & -\rho & 1 \end{pmatrix} = \left(1-\rho^2\right)\Sigma_0^{-1}.$$

Проверка гипотезы об автокорреляции. Критерий Дарбина-Уотсона

1. $d = \dfrac{\sum_{i=2}^{n}(\hat\varepsilon_i-\hat\varepsilon_{i-1})^2}{\sum_{i=1}^{n}\hat\varepsilon_i^2}$; $\hat\varepsilon_i$ – остатки, вычисленные с помощью обычного МНК, если $d\approx 2$, то автокорреляции нет.
2. Задаем уровень значимости α;
3. Вычисляем критические точки $d_h(\alpha;n)$ и $d_u(\alpha;n)$, $d_h(\alpha;n)<d_u(\alpha;n)$;
4. Проверяется гипотеза о положительной или отрицательной автокорреляции
   1) Случай $d<2$ (существование положительной автокорреляции):
   $d<d_h(\alpha;n)\ \Rightarrow$ есть положительная автокорреляция,
   $d\in[d_h(\alpha;n);\ d_u(\alpha;n)]\ \Rightarrow$ неизвестно, есть ли положительная автокорреляция,
   $d>d_u(\alpha;n)\ \Rightarrow$ положительной автокорреляции нет;
   2) Случай $d>2$ (существование отрицательной автокорреляции):
   $4-d<d_h(\alpha;n)\ \Rightarrow$ есть отрицательная автокорреляция,
   $4-d\in[d_h(\alpha;n);\ d_u(\alpha;n)]\ \Rightarrow$ неизвестно, есть ли отрицательная автокорреляция,
   $4-d>d_u(\alpha;n)\ \Rightarrow$ отрицательной автокорреляции нет.

Практические рекомендации по построению регрессионной модели

Обобщенная линейная модель множественной регрессии:

$$Y = X\Theta + \varepsilon, \quad M\varepsilon = \mathbf{0_n}, \quad \sum_{\varepsilon} = M\left(\varepsilon\varepsilon^T\right) = \sigma^2\Sigma_0, \quad rank\,X = p+1 < n.$$

Обобщенный метод наименьших квадратов:

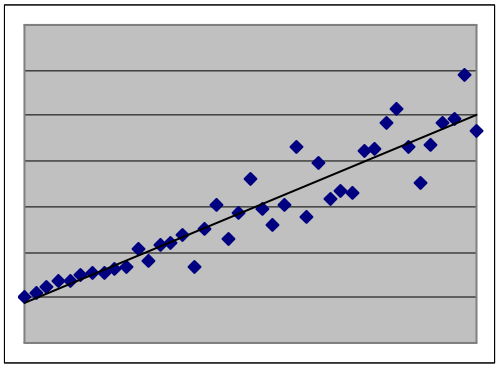
$$\hat{\Theta}_{ОМНК} = \left(X^T\Sigma_0^{-1}X\right)^{-1}\left(X^T\Sigma_0^{-1}Y\right),$$

$$\Sigma_{\hat{\Theta}} = \sigma^2\left(X^T\Sigma_0^{-1}X\right)^{-1},$$

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\left(Y - X\hat{\Theta}_{ОМНК}\right)^T\Sigma_0^{-1}\left(Y - X\hat{\Theta}_{ОМНК}\right).$$

Главная проблема – поиск матрицы Σ_0, необходимо задать ее структуру как функциональную зависимость от m параметров $a_1, a_2, ..., a_m$:

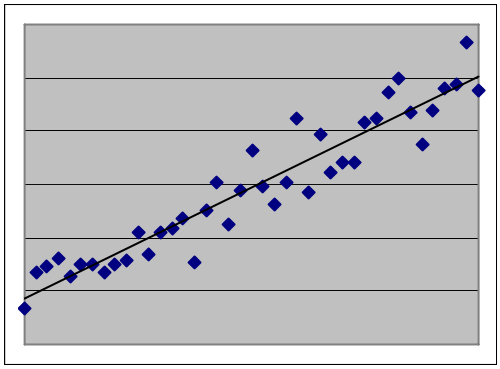
1. Частный вид гетероскедастичных остатков по некоторой переменной: постоянство относительного разброса остатков $|\hat{\varepsilon}| = \sigma x^{(j)}$, $\boxed{m=0}$.



$$\Sigma_0 = \begin{pmatrix} \left(x_1^{(j)}\right)^2 & 0 & \dots & 0 \\ 0 & \left(x_2^{(j)}\right)^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \left(x_n^{(j)}\right)^2 \end{pmatrix}$$

$$\tilde{y}_i = \frac{y_i}{x_i^{(j)}}, \quad \tilde{x}_i^{(k)} = \frac{x_i^{(k)}}{x_i^{(j)}}, \ k = 0,...,p.$$

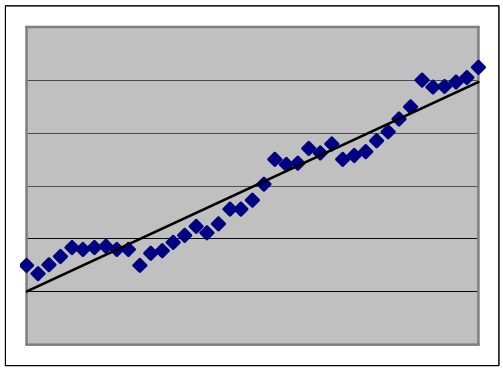
2. Общий вид гетероскедастичных остатков по некоторой переменной $|\hat{\varepsilon}| = a_0 + a_1 x^{(j)}$, $\boxed{m=2}$.



$$\Sigma_0 = \begin{pmatrix} \left(a_0 + a_1x_1^{(j)}\right)^2 & 0 & \dots & 0 \\ 0 & \left(a_0 + a_1x_2^{(j)}\right)^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \left(a_0 + a_1x_n^{(j)}\right)^2 \end{pmatrix}$$

$$\tilde{y}_i = \frac{y_i}{a_0 + a_1x_i^{(j)}}, \quad \tilde{x}_i^{(k)} = \frac{x_i^{(k)}}{a_0 + a_1x_i^{(j)}}, \ k = 0,...,p.$$

3. Автокоррелированные остатки $\varepsilon_i = \rho\varepsilon_{i-1} + \delta_i$, $\boxed{m=1}$.



$$\Sigma_0 = \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix}$$

$$\tilde{y}_1 = \left(\sqrt{1-\rho^2}\right)y_1, \quad \tilde{x}_1^{(k)} = \left(\sqrt{1-\rho^2}\right)x_1^{(k)}, \ k = 0,...,p.$$

$$\tilde{y}_i = y_i - \rho y_{i-1}, \quad \tilde{x}_i^{(k)} = x_i^{(k)} - \rho x_i^{(k)}, \ k = 0,...,p, \ i = 2,...,n.$$

По исходным наблюдениям строятся оценки $\hat{a}_1,...,\hat{a}_m$, затем вычисляется матрица $\hat{\Sigma}_0$, происходит переход к новым переменным $\tilde{y}, \tilde{x}^{(0)}, \tilde{x}^{(1)}, ..., \tilde{x}^{(p)}$, ищутся оценки $\hat{\theta}_0, \hat{\theta}_1, ..., \hat{\theta}_p$. Возможна итеративная процедура (практически реализуемый МНК)!

Итеративная процедура Кохрейна-Оркатта

1. Вычисляются МНК-оценки 1-итерации $\hat\theta_0^{(1)}, \hat\theta_1^{(1)},..., \hat\theta_p^{(1)}$.
2. Подсчитываются невязки 1-итерации $\varepsilon_i^{(1)} = y_i - \hat\theta_0^{(1)} - \hat\theta_1^{(1)} x_i^{(1)} - \hat\theta_p^{(1)} x_i^{(p)}$.
3. С помощью МНК оцениваются параметры $a_1, a_2,..., a_m$ на 1-итерации:
   1) $\left|\hat\varepsilon_i^{(1)}\right| = a_0 + a_1 x^{(j)} + \delta_i \;\Rightarrow\; \hat a_0^{(1)}, \hat a_1^{(1)}$,
   2) $\varepsilon_i^{(1)} = \rho\varepsilon_{i-1}^{(1)} + \delta_i \;\Rightarrow\; \rho^{(1)}$.
4. Осуществляется переход к переменным $\tilde y_i$, $\tilde x_i^{(0)}, \tilde x_i^{(1)},... \tilde x_i^{(p)}$.
5. Для модели с новыми переменными вычисляются МНК-оценки 2-итерации $\hat\theta_0^{(2)}, \hat\theta_1^{(2)},..., \hat\theta_p^{(2)}$.
6. Подсчитываются невязки 2-итерации $\varepsilon_i^{(2)} = y_i - \hat\theta_0^{(2)} - \hat\theta_1^{(2)} x_i^{(1)} - \hat\theta_p^{(2)} x_i^{(p)}$.
7. С помощью МНК оцениваются параметры $a_1, a_2,..., a_m$ на 2-итерации:
   1) $\left|\hat\varepsilon_i^{(2)}\right| = a_0 + a_1 x^{(j)} + \delta_i \;\Rightarrow\; \hat a_0^{(2)}, \hat a_1^{(2)}$,
   2) $\varepsilon_i^{(2)} = \rho\varepsilon_{i-1}^{(2)} + \delta_i \;\Rightarrow\; \rho^{(2)}$.
8. Осуществляется переход к переменным $\tilde y_i$, $\tilde x_i^{(0)}, \tilde x_i^{(1)},... \tilde x_i^{(p)}$.

Есть вероятность скатиться в локальный оптимум, но обычно получаем верные оценки!

y – цена квартиры
$x^{(1)}$ – общая площадь
$x^{(2)}$ – площадь кухни

y	$x^{(0)}$	$x^{(1)}$	$x^{(2)}$
630	1	30	7
640	1	31,5	6,2
610	1	31,8	5,6
980	1	48	7
950	1	46	6
1020	1	48,8	7,9
920	1	45	5,6
1050	1	52	7,2
1280	1	63	6
1310	1	66	6,8
1360	1	68	6,5
1650	1	72	8

$$\hat y_i = \underset{(125)}{-278} + \underset{(1,11)}{20{,}9}\, x_i^{(1)} + \underset{(20,1)}{39{,}5}\, x_i^{(2)} \quad \text{– 1-итерация, значим только } \theta_1$$

$$\left|\varepsilon_i\right| = \underset{(49,7)}{-115} + \underset{(0,44)}{1{,}12}\, x_i^{(1)} + \underset{(7,98)}{13{,}7}\, x_i^{(2)} - \boxed{\text{значимы } \theta_0, \theta_1} \Rightarrow$$

$$\left|\varepsilon_i\right| = -37{,}3 + 1{,}40 x_i^{(1)},\ \tilde y_i = \frac{y_i}{-37{,}3 + 1{,}40 x_i^{(1)}},\ \tilde x_i^{(j)} = \frac{x_i^{(j)}}{-37{,}3 + 1{,}40 x_i^{(1)}}$$

$$\hat y_i = \underset{(47,3)}{-230} + \underset{(0,65)}{20{,}5}\, x_i^{(1)} + \underset{(6,41)}{34{,}8}\, x_i^{(2)} \quad \text{– 2-итерация, значимы } \theta_0, \theta_1, \theta_2$$

$$\left|\varepsilon_i\right| = -46{,}0 + 1{,}54 x_i^{(1)},\ \tilde y_i = \frac{y_i}{-46{,}0 + 1{,}54 x_i^{(1)}},\ \tilde x_i^{(j)} = \frac{x_i^{(j)}}{-46{,}0 + 1{,}54 x_i^{(1)}}$$

$$\hat y_i = \underset{(42,4)}{-242} + \underset{(0,91)}{20{,}7}\, x_i^{(1)} + \underset{(3,23)}{35{,}9}\, x_i^{(2)} \quad \text{– 3-итерация, значимы } \theta_0, \theta_1, \theta_2$$

$$\hat y_i = \underset{(40,1)}{-241} + \underset{(0,81)}{20{,}6}\, x_i^{(1)} + \underset{(3,56)}{36{,}0}\, x_i^{(2)} \quad \text{– 4-итерация, значимы } \theta_0, \theta_1, \theta_2$$

$$\hat y_i = \underset{(40,1)}{-241} + \underset{(0,81)}{20{,}6}\, x_i^{(1)} + \underset{(3,51)}{36{,}0}\, x_i^{(2)} \quad \text{– 5-итерация, значимы } \theta_0, \theta_1, \theta_2$$

y – реал. обменный курс
$x^{(1)}$ – время

год	y	$x^{(0)}$	$x^{(1)}$
1995	2,5	1	0
1996	2,3	1	1
1997	2	1	2
1998	1,7	1	3
1999	3,5	1	4
2000	3,3	1	5
2001	2,8	1	6
2002	2,4	1	7
2003	2,2	1	8
2004	2,1	1	9
2005	2	1	10

$$\hat y_i = \underset{(0,331)}{2{,}527} - \underset{(0,0559)}{0{,}0182}\, x_i^{(1)} \quad \text{– 1-итерация}$$

$$d = \sum_{i=2}^{n} (\hat\varepsilon_i - \hat\varepsilon_{i-1})^2 \Big/ \sum_{i=1}^{n} \hat\varepsilon_i^2 = \frac{3{,}995}{3{,}089} = 1{,}28,$$

$d_h(0{,}05; 11) = 0{,}98$, $d_u(0{,}05; 11) = 1{,}32$ – подтвердить или опровергнуть наличие положительной автокорреляции невозможно

$$\varepsilon_i = 0{,}354\varepsilon_{i-1} + \delta_i,\quad \tilde y_1 = \sqrt{1 - 0{,}354^2}\, y_1,\quad \tilde y_i = y_i - \rho y_{i-1},$$
$$\tilde x_1^{(j)} = \sqrt{1 - 0{,}354^2}\, x_1^{(j)},\quad \tilde x_i^{(j)} = x_i^{(j)} - \rho x_{i-1}^{(j)}$$

$$\hat y_i = \underset{(0,428)}{2{,}550} - \underset{(0,0704)}{0{,}0260}\, x_i^{(1)} \quad \text{– 2-итерация}$$

$$\hat y_i = \underset{(0,430)}{2{,}550} - \underset{(0,0706)}{0{,}0261}\, x_i^{(1)} \quad \text{– 3-итерация}$$

Точечный прогноз в моделях линейной регрессии

Наиболее распространенная задача: предсказывать y по известному X.

$$\begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix} \begin{pmatrix} x_1^{(0)}=1 & x_1^{(1)} & \dots & x_1^{(p)} \\ \dots & \dots & \dots & \dots \\ x_n^{(0)}=1 & x_n^{(1)} & \dots & x_n^{(p)} \end{pmatrix} \quad - \text{ известные данные.}$$

$$\boxed{(y_{n+1})} \left(x_{n+1}^{(0)}=1 \quad x_{n+1}^{(1)} \quad \dots \quad x_{n+1}^{(p)}\right).$$

неизвестное значение

Также известен характер ковариационных связей регрессионного остатка ε_{n+1}:

$\sigma_\varepsilon^{(n+1)} = \left(M(\varepsilon_1\varepsilon_{n+1}),\dots,M(\varepsilon_n\varepsilon_{n+1})\right)^T$, $M\varepsilon_{n+1}=0$, $M\varepsilon_{n+1}^2=\Delta^2$.

Наилучший (в смысле среднего квадрата ошибки) несмещенный прогноз для y_{n+1}:

$\hat{y}_{n+1} = X_{n+1}^T\hat{\Theta}_{OMHK} + \left(\sigma_\varepsilon^{(n+1)}\right)^T\Sigma_\varepsilon^{-1}\hat{\varepsilon}$.

Только когда $M(\varepsilon_i\varepsilon_{n+1})=0,\ i=1,\dots,n$ прогноз $\hat{y}_{n+1}$ совпадает со значением функции регрессии.

Классический случай и случай гетероскедастичных остатков (диагональная матрица Σ_0):

$\hat{y}_{n+1} = X_{n+1}^T\hat{\Theta}_{OMHK}$.

Случай автокоррелированных остатков:

$$\left(\sigma_\varepsilon^{(n+1)}\right)^T = \sigma^2\left(\rho^n \quad \rho^{n-1} \quad \dots \quad \rho\right).$$

$$\left(\sigma_\varepsilon^{(n+1)}\right)^T\Sigma_\varepsilon^{-1} = \sigma^2\left(\rho^n \quad \rho^{n-1} \quad \dots \quad \rho\right)\times\frac{1}{\sigma^2\left(1-\rho^2\right)}\begin{pmatrix} 1 & -\rho & \dots & 0 \\ -\rho & 1+\rho^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix} = \frac{\left(0 \quad 0 \quad \dots \quad 0 \quad \rho\left(1-\rho^2\right)\right)}{\left(1-\rho^2\right)}.$$

$$\left(\sigma_\varepsilon^{(n+1)}\right)^T\Sigma_\varepsilon^{-1} = \left(0 \quad 0 \quad \dots \quad 0 \quad \rho\right).$$

$$\hat{y}_{n+1} = X_{n+1}^T\hat{\Theta}_{OMHK} + \rho\hat{\varepsilon}_n.$$

Интервальный прогноз в моделях линейной регрессии

Для построения доверительного интервала необходима оценка точности точечного прогноза.

1. Классическая модель:

$$\sigma_{n+1}^2 = M(\hat{y}_{n+1}-y_{n+1})^2 = \hat{\sigma}^2\left(X_{n+1}^T\left(X^TX\right)^{-1}X_{n+1}+1\right).$$

Парная регрессия:

$$X^TX = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{pmatrix}, \quad \left(X^TX\right)^{-1} = \frac{1}{\sum(x_i-\bar{x})^2}\begin{pmatrix} \frac{1}{n}\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix},$$

$$\sum(x_i-\bar{x})^2 = \sum x_i^2 - 2\sum x_i\bar{x} + n\bar{x}^2 = \sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum x_i^2 - n\bar{x}^2,$$

$$X_{n+1}^T\left(X^TX\right)^{-1}X_{n+1} = \frac{1}{\sum(x_i-\bar{x})^2}\begin{pmatrix}1 & x_{n+1}\end{pmatrix}\begin{pmatrix} \frac{1}{n}\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}\begin{pmatrix} 1 \\ x_{n+1} \end{pmatrix} = \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum(x_i-\bar{x})^2},$$

$$y_{n+1} \in \left[\hat{y}_{n+1} - u_{\frac{1+\gamma}{2}}\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{\sum(x_i-\bar{x})^2}};\quad \hat{y}_{n+1} + u_{\frac{1+\gamma}{2}}\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{\sum(x_i-\bar{x})^2}}\right].$$

2. Обобщенная модель (отличия от классической):

1) $\sigma_{n+1}^2 = \hat{\sigma}^2\left(X_{n+1}^T\left(X^T\Sigma_0^{-1}X\right)^{-1}X_{n+1}+1\right)$;

2) $\hat{\sigma}^2$ – получены на последней итерации практически реализуемого ОМНК;

3) $y_{n+1} \in \left[\hat{y}_{n+1} - t_{1-\gamma}(n-p-1)\sigma_{n+1};\quad \hat{y}_{n+1} + t_{1-\gamma}(n-p-1)\sigma_{n+1}\right]$.

Точность регрессионной модели в реалистической ситуации

Возможно неверное предположение о линейности функции регрессии:

$M(y:X)=\theta_0+\theta_1x^{(1)}+...+\theta_px^{(p)}\ \Rightarrow$ аппроксимация неизвестной функции – линейной.

1. Доверительные интервалы оказываются неточными – как для неизвестных значений параметров, так и для прогнозов.
2. Особая осторожность необходима при экстраполяции – при восстановлении значения функции регрессии по значениям объясняющих переменных, лежащим вне статистически обследованной области.

Стратегия:

1. Разбиваем исходную выборку (X,Y) объема n на обучающую выборку $(X1,Y1)$ объема n_1 и экзаменующую выборку $(X2,Y2)$ объема n_2 k различными способами.
2. По каждой обучающей выборке $j=1,...,k$ получаем ОМНК-оценки $\hat{\Theta}_j^*=\left(\hat{\theta}_{0j}^*,\hat{\theta}_{1j}^*,...,\hat{\theta}_{pj}^*\right)$.
3. По экзаменующим выборкам оцениваем точность

$$(\hat{\sigma}*)^2=\frac{1}{\left(\sum_{j=1}^{k}n_{2j}\right)-p-1}\sum_{j=1}^{k}\left(Y2_j-X2_j\hat{\Theta}_j^*\right)^T\Sigma_0^{-1}\left(Y2_j-X2_j\hat{\Theta}_j^*\right).$$

Метод скользящего экзамена

Способ разбиения: j-обучающая выборка – исходная выборка X без j-наблюдения:
$k=n,\ n_{1j}=n-1,\ n_{2j}=1$.

Полученное значение $(\hat{\sigma}*)^2$ используется вместо оценки $\hat{\sigma}^2=\frac{1}{n-p-1}\left(Y-X\hat{\Theta}\right)^T\Sigma_0^{-1}\left(Y-X\hat{\Theta}\right)$, получаемой по формулам ОМНК. Оно, как правило, оказывается существенно больше.

Помесячные данные о печати фотографий в некоторой фирме за 24 месяца

	месяц	***y*, количество, шт.**	***x*(1), цена индекс., руб.**	***x*(2), реклама индекс., руб.**
1	январь, 2003	12 500	2,50	0
2	февраль, 2003	7 600	3,03	0
3	март, 2003	6 900	2,97	0
4	апрель, 2003	13 500	2,97	4 950
……	……	……	……	……
24	декабрь, 2004	11 800	3,15	1 575

$\hat{y}_i=\underset{(3052)}{19206}-\underset{(929)}{3344}\,x_i^{(1)}+\underset{(0,1972)}{0,9889}\,x_i^{(2)}$, $\boxed{\hat{\sigma}=1484}$.

Без 1-й строки $\hat{y}_i=\underset{(3499)}{16732}-\underset{(1046)}{2643}\,x_i^{(1)}+\underset{(0,1956)}{1,0291}\,x_i^{(2)}$, $\hat{\varepsilon}_1^*=2376$;

Без 2-й строки $\hat{y}_i=\underset{(3098)}{19803}-\underset{(937)}{3486}\,x_i^{(1)}+\underset{(0,2002)}{0,9501}\,x_i^{(2)}$, $\hat{\varepsilon}_2^*=-1639$;

……

Без 24-й строки $\hat{y}_i=\underset{(3059)}{18816}-\underset{(930)}{3245}\,x_i^{(1)}+\underset{(0,1964)}{0,9845}\,x_i^{(2)}$, $\hat{\varepsilon}_{24}^*=-1653$.

$$\boxed{\hat{\sigma}*=\sqrt{\frac{1}{21}\left(2376^2+(-1639)^2+...+(-1653)^2\right)}=1675}.$$

Необязательно проводить полный расчет, при вычеркивании *i*-строки:

$$\hat{\varepsilon}_i^*=\frac{\hat{\varepsilon}_i}{1-q_i},\quad \hat{\Theta}_i^*=\hat{\Theta}+\frac{y_i-X_i^T\hat{\Theta}}{1-q_i}\left(X^TX\right)^{-1}X_i,\quad q_i=X_i^T\left(X^TX\right)^{-1}X_i.$$

Линейные регрессионные модели с переменной структурой

Проблема неоднородности данных: y зависит не только от X, но и от уровня сопутствующих переменных Z (как правило, не являющихся количественными).

Сезонность (сезон, квартал, месяц), пол, возраст, социальная страта, регион,…

Способы решения:

1. **Разбиение имеющихся статистических данных на однородные порции** (внутри каждой подвыборки значения переменных Z постоянны).
   Для каждой подвыборки своя функция регрессия $\hat{f}\left(X, Z^*\right) = \hat{\theta}_0\left(Z^*\right) + \hat{\theta}_1\left(Z^*\right)x^{(1)} + \ldots + \hat{\theta}_p\left(Z^*\right)x^{(p)}$.
   При этом $\hat{f}\left(X, Z^*\right)$ и $\hat{f}\left(X, Z^{**}\right)$ значимо отличаются.
   Проблемы:
   1) сопутствующие переменные Z ненаблюдаемы, либо эти значения не были зарегистрированы при сборе исходных данных $\Rightarrow$ прямое разбиение выборки невозможно, необходимо использование методов классификации объектов (расщепление смеси вероятностных распределений, кластер-анализ);
   2) прямое разбиение возможно, но приводит к слишком малым подвыборкам.
2. **Введение дамми-переменных** (фиктивных переменных, переменных-манекенов)
   Преимущества:
   1) сильно повышается статистическая надежность оценок;
   2) одновременно появляется возможность проверки гипотез о значимом влиянии сопутствующих переменных.

Если категоризованная переменная $z^{(j)}$ имеет k_j градаций, то требуется ввести $\left(k_j - 1\right)$ бинарных дамми-переменных (принимающих значения 0 или 1)!

Социальная страта (низкодоходная / среднедоходная / высокодоходная), $k_1 - 1 = 2$

$$z_i^{(1.1)} = \begin{cases} 1, & \text{если } i\text{-наблюдение за среднедоходным домашним хозяйством,} \\ 0, & \text{иначе;} \end{cases}$$

$$z_i^{(1.2)} = \begin{cases} 1, & \text{если } i\text{-наблюдение за высокодоходным домашним хозяйством,} \\ 0, & \text{иначе.} \end{cases}$$

Сезонность (зима / весна / лето / осень), $k_2 - 1 = 3$

$$z_i^{(2.1)} = \begin{cases} 1, & \text{если } i\text{-наблюдение осуществлено весной,} \\ 0, & \text{иначе;} \end{cases}$$

$$z_i^{(2.2)} = \begin{cases} 1, & \text{если } i\text{-наблюдение осуществлено летом,} \\ 0, & \text{иначе;} \end{cases}$$

$$z_i^{(2.3)} = \begin{cases} 1, & \text{если } i\text{-наблюдение осуществлено осенью,} \\ 0, & \text{иначе.} \end{cases}$$

Возможны различные варианты зависимостей, например, следующие:

Вариант 1. При переходе из страты в страту и из сезона в сезон меняется только свободный член регрессии θ_0 (абсолютное потребление); θ_1 (склонность к потреблению) постоянна:

$$\hat{y} = \theta_0 + \theta_1 x + \theta_{1.1} z^{(1.1)} + \theta_{1.2} z^{(1.2)} + \theta_{2.1} z^{(2.1)} + \theta_{2.2} z^{(2.2)} + \theta_{2.3} z^{(2.3)}.$$

Вариант 2. При переходе из страты в страту меняется склонность к потреблению; фактор сезонности по-прежнему влияет только на потребляемое количество:

для низкодоходной страты склонность к потреблению $\hat{\theta}_1$;

для среднедоходной страты склонность к потреблению $\hat{\theta}_1 + \hat{\theta}_{1.1}$;

для высокодоходной страты склонность к потреблению $\hat{\theta}_1 + \hat{\theta}_{1.2}$;

$$\hat{y} = \theta_0 + \theta_1 x + \theta_{1.1}\left(z^{(1.1)} x\right) + \theta_{1.2}\left(z^{(1.2)} x\right) + \theta_{2.1} z^{(2.1)} + \theta_{2.2} z^{(2.2)} + \theta_{2.3} z^{(2.3)}.$$

Статистическая надежность:

Точность модели зависит от соотношения $n/(p+1)$ – чем оно больше, тем точнее оценки.

Обследование проведено на 12 семьях (по 4 от каждой страты) за 12 месяцев.

1) Изолированная оценка по стратам-сезонам $n/(p+1)=12/2=6$.
2) Оценка по всем наблюдениям $n/(p+1)=144/7\approx 20{,}6$ – **точнее в 3,5 раза!**

Проверка наличия значимого влияния сопутствующих переменных на структуру модели:

$\hat{y} = \underset{(2,1)}{6{,}8} + \underset{(0,006)}{0{,}025}\,x + \underset{(1,2)}{3{,}1}\,z^{(1.1)} + \underset{(2,1)}{5{,}2}\,z^{(1.2)} + \underset{(1,19)}{0{,}8}\,z^{(2.1)} + \underset{(1,1)}{4{,}7}\,z^{(2.2)} - \underset{(1,18)}{1{,}2}\,z^{(2.3)}$.

Из переменных сезонности $z^{(2.1)}$ и $z^{(2.3)}$ являются незначимыми, значимое влияние оказывает только лето $z^{(2.2)}$, остальные переменные можно исключить.

Если ни одна переменная $z^{(j)}$ не влияет значимо на *y*, то неоднородности данных нет!

Ловушка, связанная с введением дамми-переменных:

Если у переменной $z^{(j)}$ есть *k* градаций, то есть риск ввести *k* дамми-переменных.

$z_i^{(2.4)} = \begin{cases} 1, & \text{если } i\text{-наблюдение осуществлено зимой,} \\ 0, & \text{иначе.} \end{cases}$

месяц	$z^{(2.1)}$	$z^{(2.2)}$	$z^{(2.3)}$	$z^{(2.4)}$
январь	0	0	0	1
февраль	0	0	0	1
март	1	0	0	0
апрель	1	0	0	0
май	1	0	0	0
июнь	0	1	0	0
июль	0	1	0	0
август	0	1	0	0
сентябрь	0	0	1	0
октябрь	0	0	1	0
ноябрь	0	0	1	0
декабрь	0	0	0	1

В данной модели присутствует линейная зависимость переменных: $z^{(2.1)} + z^{(2.2)} + z^{(2.3)} + z^{(2.4)} = x^{(0)} \equiv 1$ (полная мультиколлинеарность).

Матрица X^TX вырожденная, обратной матрицы $\left(X^TX\right)^{-1}$ не существует, формулы МНК не работают!

Число дамми-переменных должно быть на единицу меньше числа градаций соответствующей категоризованной переменной!

Объем продаж мороженого (млн.шт.) за 5 лет в зависимости от цены (руб.) и сезона

год	**сезон**	***y*, кол-во**	**цена**	**индекс цен**	***x*, цена инд.**	$z^{(1)}$**, весна**	$z^{(2)}$**, лето**	$z^{(3)}$**, осень**
1999	весна	1,5	3	1	3,00	1	0	0
	лето	2,6	4	1,11	3,60	0	1	0
	осень	1,7	3,5	1,15	3,04	0	0	1
	зима	0,9	3,5	1,26	2,78	0	0	0
2000	весна	1,4	4	1,34	2,99	1	0	0
	лето	3	4	1,40	2,86	0	1	0
	осень	2,8	4	1,45	2,76	0	0	1
	зима	1,6	4	1,52	2,63	0	0	0
2001	весна	1,9	4,5	1,59	2,83	1	0	0
	лето	3,2	5	1,63	3,07	0	1	0
	осень	2,7	4,5	1,68	2,68	0	0	1
	зима	2	4,5	1,78	2,53	0	0	0
2002	весна	2,2	5	1,87	2,67	1	0	0
	лето	3,4	5	1,95	2,56	0	1	0
	осень	2,6	5	2,01	2,49	0	0	1
	зима	2,1	5	2,09	2,39	0	0	0
2003	весна	2,9	5	2,16	2,31	1	0	0
	лето	3,3	6	2.19	2,74	0	1	0
	осень	2,5	6	2,24	2,68	0	0	1
	зима	2,2	6	2,32	2,59	0	0	0

Исходная модель: $\hat{y} = \underset{(1,56)}{3{,}02} - \underset{(0,56)}{0{,}25}\,x$, $\hat{R}^2 = 0{,}011 = 1{,}1\%$, $\hat{\sigma} = 0{,}71$.

Модель с дамми-переменными: $\hat{y} = \underset{(0,74)}{5{,}35} - \underset{(0,28)}{1{,}39}x + \underset{(0,2)}{0{,}47}z^{(1)} + \underset{(0,22)}{1{,}87}z^{(2)} + \underset{(0,2)}{0{,}9}\,z^{(3)}$, $\hat{R}^2 = 0{,}84 = 84\%$, $\hat{\sigma} = 0{,}31$.

Учет эффекта взаимодействия сопутствующих переменных

До сих пор сопутствующие переменные влияли на результирующий показатель независимо, рассмотрим случай их взаимодействия.

Категоризованной переменной $z^{(j)}$ соответствуют дамми-переменные $z^{(j.1)}, z^{(j.2)}, \ldots, z^{(j.k_j-1)}$;
Категоризованной переменной $z^{(l)}$ соответствуют дамми-переменные $z^{(l.1)}, z^{(l.2)}, \ldots, z^{(l.k_l-1)}$.

↓

вводится $N = (k_j - 1)(k_l - 1)$ дамми-переменных, образуемых всевозможными попарными произведениями $z_{qs} = z^{(j.q)} z^{(l.s)}$.

y – заработная плата, $x^{(1)}, \ldots, x^{(p)}$ – количественные объясняющие переменные (стаж, экспертные оценки качества работы…), $z^{(1)}$ – образование (начальное / среднее / высшее), $z^{(2)}$ – пол (мужской / женский).

образование	**пол**	$z^{(1.1)}$	$z^{(1.2)}$	$z^{(2.1)}$	$z^{(3.1)}=z^{(1.1)}z^{(2.1)}$	$z^{(3.1)}=z^{(1.2)}z^{(2.1)}$
начальное	мужской	0	0	0	0	0
начальное	женский	0	0	1	0	0
среднее	мужской	1	0	0	0	0
среднее	женский	1	0	1	1	0
высшее	мужской	0	1	0	0	0
высшее	женский	0	1	1	0	1

$\hat{y} = \theta_0 + \theta_1 x^{(1)} + \ldots + \theta_p x^{(p)} + \theta_{1.1} z^{(1.1)} + \theta_{1.2} z^{(1.2)} + \theta_{2.1} z^{(2.1)} + \theta_{3.1} z^{(3.1)} + \theta_{3.1} z^{(3.1)}$ –
модель, полученная после логарифмирования переменных (осуществлен переход от мультипликативной модели к аддитивной).

Статистически незначимые оценки означают, что данный фактор (или взаимодействие нескольких факторов) в модели отсутствует!

Проверка регрессионной однородности двух групп наблюдений

Случай 1. Большая выборка В1 (n_1 наблюдений) + большая выборка В2 (n_2 наблюдений):

Статистическая проверка $\theta_0^{(1)} = \theta_0^{(2)}, \theta_1^{(1)} = \theta_1^{(2)}, \ldots, \theta_p^{(1)} = \theta_p^{(2)}$.

Например, построить доверительные интервалы для $\theta_0^{(1)}, \theta_1^{(1)}, \ldots, \theta_p^{(1)}$, проверить, входят ли в эти интервалы значения $\hat{\theta}_0^{(2)}, \hat{\theta}_1^{(2)}, \ldots, \hat{\theta}_p^{(2)}$.

Случай 2. Большая выборка В1 + малая выборка В2:

Критерий Чоу:

1) Выбираем уровень значимости α.
2) По выборке В1 строим МНК-оценки и вычисляем невязки $\hat{\varepsilon}^{(1)} = Y^{(1)} - X^{(1)}\hat{\Theta}^{(1)}$.
3) По выборке В2 строим МНК-оценки и вычисляем невязки $\hat{\varepsilon}^{(2)} = Y^{(2)} - X^{(2)}\hat{\Theta}^{(2)}$.
4) По полной выборке В1+В2 строим МНК-оценки и вычисляем невязки $\hat{\varepsilon} = Y - X\hat{\Theta}$.
5) $$F_{\text{эмп}} = \frac{\left(\sum_{i=1}^{n_1+n_2} (\hat{\varepsilon}_i)^2 - \sum_{i=1}^{n_1} \left(\hat{\varepsilon}_i^{(1)}\right)^2 - \sum_{i=1}^{n_2} \left(\hat{\varepsilon}_i^{(2)}\right)^2\right) \Big/ (p+1)}{\left(\sum_{i=1}^{n_1} \left(\hat{\varepsilon}_i^{(1)}\right)^2 + \sum_{i=1}^{n_2} \left(\hat{\varepsilon}_i^{(2)}\right)^2\right) \Big/ (n_1 + n_2 - 2p - 2)}.$$
6) $F_{\text{эмп}} > FРАСПОБР(\alpha;\ p+1;\ n_1 + n_2 - 2p - 2) \Rightarrow$ гипотеза об однородности выборок В1 и В2 отвергается.

Случай 3. Большая выборка B1 + сверхмалая выборка B2:

По B2 нельзя построить значимые оценки коэффициентов регрессии (например, в случае $n_2 < p+1$) – в частности, когда к основной выборке B1 добавляется небольшая порция данных B2, и вопрос в том, можно ли их объединять).

Модифицированный критерий Чоу:

$$F_{\text{эмп}} = \frac{\left(\sum_{i=1}^{n_1+n_2} (\hat{\varepsilon}_i)^2 - \sum_{i=1}^{n_1} \left(\hat{\varepsilon}_i^{(1)}\right)^2 \right) \Big/ n_2}{\sum_{i=1}^{n_1} \left(\hat{\varepsilon}_i^{(1)}\right)^2 \Big/ (n_1 - p - 1)},$$

$F_{\text{эмп}} > FРАСПОБР(\alpha; n_2; n_1 - p - 1) \Rightarrow$ гипотеза об однородности выборок отвергается.

Зависимость зарплаты y, руб./мес. от стажа x, лет на некотором предприятии.

зарплата	стаж
4 949	2
9 094	15
9 167	7
11 836	11
9 683	3
9 927	1
11 970	24
10 607	10
5 747	2
15 327	14
9 844	9
4 953	8
6 152	1
9 109	4
1 6235	7
2 621	1
13 702	12
5 771	6
15 416	9
12 035	5

Имеется две дополнительных порции данных. Вопрос: какая из них относится к обследуемому предприятию?

Основная выборка: $\boxed{\hat{y}^{(1)} = 7143 + 339{,}6x}$, $\sum \left(\hat{\varepsilon}_i^{(1)}\right)^2 = 204858003$.

Дополнительная выборка 1:

зарплата	стаж
15 585	4
19 816	3

$\boxed{\hat{y} = 8756 + 233{,}7x}$, $\sum \hat{\varepsilon}_i^2 = 367336733$,

$$F_{\text{эмп}} = \frac{(367\,336\,733 - 204\,858\,003)/2}{204\,858\,003/(20-1-1)} = 7{,}14,\ F_{\text{крит}} = 3{,}55,$$

$F_{\text{эмп}} > F_{\text{крит}} \Rightarrow$ гипотеза об однородности отвергается.

Дополнительная выборка 2:

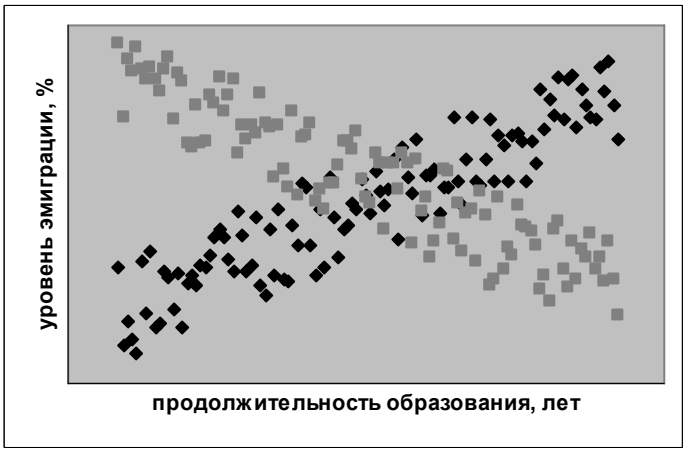
зарплата	стаж
10 271	8
21 044	17

$\boxed{\hat{y} = 6729 + 439{,}7x}$, $\sum \hat{\varepsilon}_i^2 = 260\,015\,984$,

$$F_{\text{эмп}} = \frac{(260\,015\,984 - 204\,858\,003)/2}{204\,858\,003/(20-1-1)} = 2{,}42,\ F_{\text{крит}} = 3{,}55,$$

$F_{\text{эмп}} < F_{\text{крит}} \Rightarrow$ гипотеза об однородности принимается.

Пример неоднородности данных при неизвестных сопутствующих переменных



Регрессионный анализ свидетельствует об отсутствии связи. Геометрически – 2 пересекающиеся крестом подвыборки.

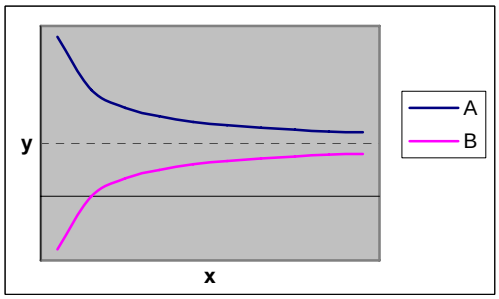
Вывод: имеется скрытый сопутствующий признак – тип образования (гуманитарное / естественно-техническое).

Проблема: при $p \geq 3$ визуальный анализ принципиально невозможен.

Нелинейные модели, поддающиеся непосредственной линеаризации

Зависимости гиперболического типа

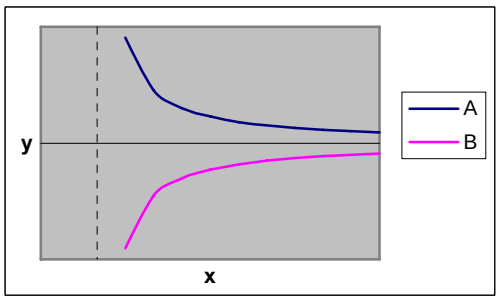
1) $y = \theta_0 + \dfrac{\theta_1}{x} + \varepsilon, \quad x \in (0;\ +\infty)$.



Случай A: $\theta_0 > 0,\ \theta_1 > 0$.
Случай B: $\theta_0 > 0,\ \theta_1 < 0$.
Вертикальная асимптота $x = 0$.
Горизонтальная асимптота $y = \theta_0$.
Замена: $\tilde{x}_i = 1/x_i$.

$$\boxed{y = \theta_0 + \theta_1 \tilde{x} + \varepsilon}.$$

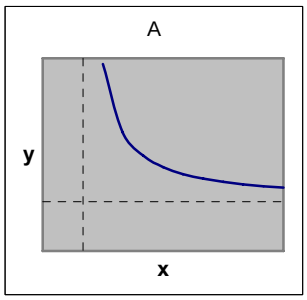
2) $y = \dfrac{1}{\theta_0 + \theta_1 x + \varepsilon}, \quad x \in \left(-\dfrac{\theta_0}{\theta_1};\ +\infty\right)$.

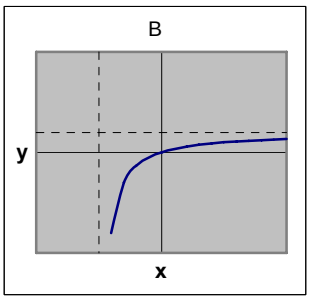


Случай A: $\theta_0 < 0,\ \theta_1 > 0$.
Случай B: $\theta_0 > 0,\ \theta_1 < 0$.
Вертикальная асимптота $x = -\theta_0/\theta_1$.
Горизонтальная асимптота $y = 0$.
Замена: $\tilde{y}_i = 1/y_i$.

$$\boxed{\tilde{y} = \theta_0 + \theta_1 x + \varepsilon}.$$

3) $y = \dfrac{x}{\theta_0 x + \theta_1 + x\varepsilon}, \quad x \in \left(-\dfrac{\theta_1}{\theta_0};\ +\infty\right)$.



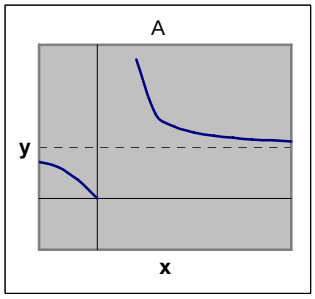


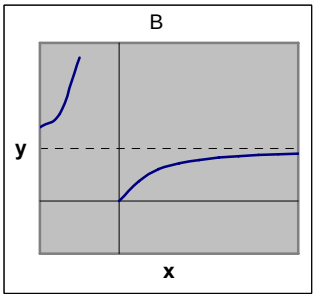
Случай A: $\theta_0 > 0,\ \theta_1 < 0$.
Случай B: $\theta_0 > 0,\ \theta_1 > 0$.
Вертикальная асимптота $x = -\theta_1/\theta_0$.
Горизонтальная асимптота $y = 1/\theta_0$.
Замена: $\tilde{x}_i = 1/x_i$, $\tilde{y}_i = 1/y_i$.

$$\boxed{\tilde{y} = \theta_0 + \theta_1 \tilde{x} + \varepsilon}.$$

Зависимости экспоненциального типа

4) $y = \theta_0\, e^{\theta_1/x + \varepsilon}, \quad x \in (0;\ +\infty)$.





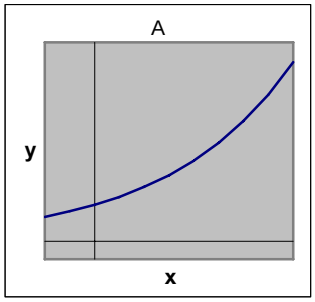
Случай A: $\theta_0 > 0,\ \theta_1 > 0$.
Случай B: $\theta_0 > 0,\ \theta_1 < 0$.
Вертикальная асимптота $x = 0$.
Горизонтальная асимптота $y = \theta_0$.
Замена: $\tilde{x}_i = 1/x_i$, $\tilde{y}_i = \ln y_i$, $\tilde{\theta}_0 = \ln \theta_0$.

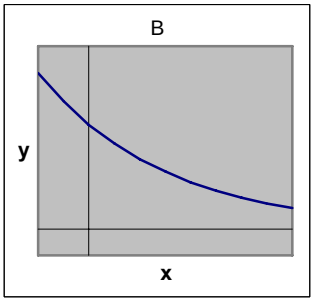
$$\boxed{\tilde{y} = \tilde{\theta}_0 + \theta_1 \tilde{x} + \varepsilon}.$$

1А, 2А, 3А, 4А – спрос от цены,
1Б, 3Б, 4Б – спрос от дохода (кривые Энгеля).

5) Постоянный темп относительного прироста во времени
5% / год = 132 раза /век, 10% / год = 13781 раз / век. $\boxed{132>6,\ \ 13781>11}$

$y = \theta_0\, e^{\theta_1 x+\varepsilon}$, $y \in (0;\ +\infty)$.



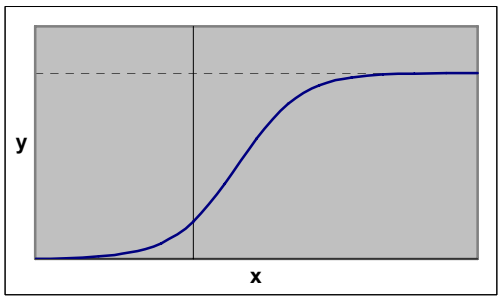


Случай A: $\theta_0 > 0,\ \theta_1 > 0$.
Случай B: $\theta_0 > 0,\ \theta_1 < 0$.
Горизонтальная асимптота $y = 0$.
Замена: $\tilde{y}_i = \ln y_i$, $\tilde{\theta}_0 = \ln \theta_0$.

$$\boxed{\tilde{y} = \tilde{\theta}_0 + \theta_1 x + \varepsilon}.$$

6) Логистическая кривая

$$y = \frac{1}{\theta_0 + \theta_1 e^{-x} + \varepsilon},\quad x \in (-\infty;\ +\infty).$$



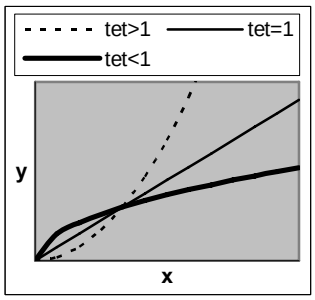
В данном случае $\theta_1 > \theta_0 > 0$.
Горизонтальные асимптоты $y = 0$ и $y = 1/\theta_0$.
Пересечение оси ординат в точке $y = 1/(\theta_0 + \theta_1)$.
Замена: $\tilde{x}_i = e^{-x_i}$, $\tilde{y}_i = 1/y_i$.

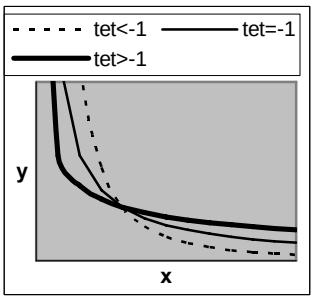
$$\boxed{\tilde{y} = \theta_0 + \theta_1 \tilde{x} + \varepsilon}.$$

Моделирование насыщаемых показателей.

Зависимости степенного типа (возможна множественная регрессия)

7) $y = \theta_0 \left(x^{(1)}\right)^{\theta_1} \ldots \left(x^{(p)}\right)^{\theta_p} e^{\varepsilon}$, $x^{(j)} \in (0;\ +\infty)$, $j = 1,\ldots, p$.





Замена: $\tilde{x}_i^{(j)} = \ln x_i^{(j)}$, $\tilde{y}_i = \ln y_i$, $\tilde{\theta}_0 = \ln \theta_0$.

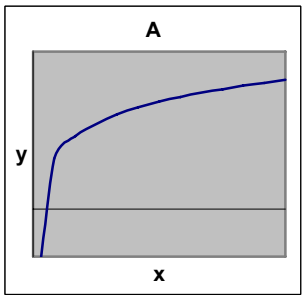
$$\boxed{\tilde{y} = \tilde{\theta}_0 + \theta_1 \tilde{x}^{(1)} + \ldots + \theta_p \tilde{x}^{(p)} + \varepsilon}.$$

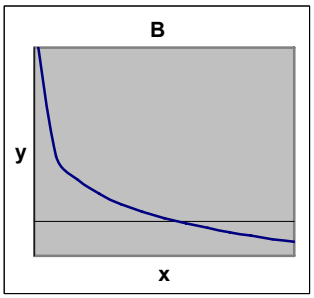
θ_j, $j = 1,\ldots, p$ – коэффициенты эластичности по переменной $x^{(j)}$.
Если эластичность постоянна, то зависимость всегда степенного типа.

Производственные функции, функции спроса от цены, дохода, других факторов.
Степенные функции обычно используются в динамических моделях, если результирующий показатель зависит от мультипликативно воздействующих объясняющих.

Зависимость логарифмического типа

8) $y = \theta_0 + \theta_1 \ln x + \varepsilon$, $x \in (0;\ +\infty)$.





Случай A: $\theta_1 > 0$.
Случай B: $\theta_1 < 0$.
Вертикальная асимптота $x = 0$.
Замена: $\tilde{x}_i = \ln x_i$.

$$\boxed{y = \theta_0 + \theta_1 \tilde{x} + \varepsilon}.$$

Выбор вида зависимости

Задача: выбрать из всевозможных видов моделей наилучшую.

1. Метод проб и ошибок

5. Построить различные варианты моделей (полиномы, гиперболического, экспоненциального, степенного, логарифмического типа и др.).
6. Оценить модели (найти значения всех коэффициентов модели).
7. Выбрать наилучшую из этих моделей:
   1) Используя методы проверки гипотезы о виде функции регрессии
   2) По максимальному значению множественного коэффициента корреляции с учетом количества параметров модели (если y входит в модель линейно!):

$\hat{R}^2_{y.X} = 1 - \dfrac{D\varepsilon}{Dy} = 1 - \dfrac{\sum \varepsilon_i^2}{\sum (y_i - \bar{y})^2}$ $\qquad$ $\hat{R}^2_{y.X} = 1 - \dfrac{|R|}{|R|_{00}}$, R – корреляционная матрица.

$\hat{R}^{2*}_{y.X} = 1 - \left(1 - \hat{R}^2_{y.X}\right)\dfrac{n-1}{n-p-1}$ – несмещенная оценка.

2. Метод Бокса-Кокса

Метод Бокса-Кокса – формализованная процедура подбора линеаризующего преобразования:

$$\tilde{y}_i(\lambda) = \frac{y_i^{\lambda} - 1}{\lambda}, \quad \tilde{x}_i^{(j)}(\lambda) = \frac{\left(x_i^{(j)}\right)^{\lambda} - 1}{\lambda}, \quad i = 1,..,n, \quad j = 1,...,p.$$

Гипотеза: существует значение λ^*, такое что

$\tilde{y}_i(\lambda^*) = \theta_0 + \theta_1 \tilde{x}_i^{(1)}(\lambda^*) + ... + \theta_p \tilde{x}_i^{(p)}(\lambda^*) + \varepsilon_i$ или $\tilde{y}_i(\lambda^*) = \theta_0 + \theta_1 x_i^{(1)} + ... + \theta_p x_i^{(p)} + \varepsilon_i$.

Замечание 1

Преобразования применяются исключительно к положительным переменным. Если по некоторой переменной имеются отрицательные значения, осуществляется сдвиг:

$$\tilde{y}_i(\lambda) = \frac{\left(y_i + c^{(0)}\right)^{\lambda} - 1}{\lambda}, \quad \tilde{x}_i^{(j)}(\lambda) = \frac{\left(x_i^{(j)} + c^{(j)}\right)^{\lambda} - 1}{\lambda}, \quad c^{(j)} > \left|\min_{i=1,...,n} x_i^{(j)}\right|, \quad i = 1,..,n, \quad j = 1,...,p.$$

Замечание 2

$\lambda^* = 1$ – линейная зависимость y и $x^{(1)},...,x^{(p)}$.

$\lambda^* = 0$ – степенная или экспоненциальная зависимость y и $x^{(1)},...,x^{(p)}$:

$$\tilde{y}_i(0) = \lim_{\lambda \to 0} \frac{y_i^{\lambda} - 1}{\lambda} = \ln y_i, \quad \tilde{x}_i^{(j)}(0) = \lim_{\lambda \to 0} \frac{\left(x_i^{(j)}\right)^{\lambda} - 1}{\lambda} = \ln x_i^{(j)}, \quad i = 1,..,n, \quad j = 1,...,p;$$

$\ln y = \theta_0 + \theta_1 \ln x^{(1)} + ... + \theta_p \ln x^{(p)}$, $\boxed{y = e^{\theta_0} \left(x^{(1)}\right)^{\theta_1} \left(x^{(p)}\right)^{\theta_p}}$ или

$\ln y = \theta_0 + \theta_1 x^{(1)} + ... + \theta_p x^{(p)}$, $\boxed{y = e^{\theta_0 + \theta_1 x^{(1)} + ... + \theta_p x^{(p)}}}$.

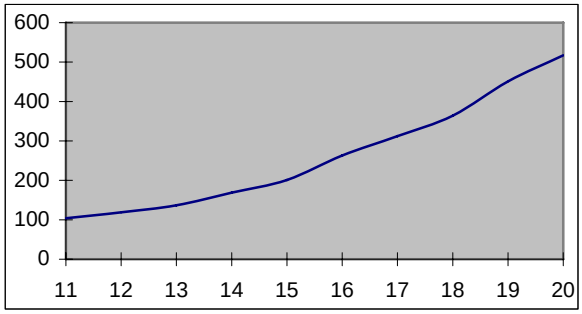
При других λ^* получаем связь каких-то степеней исходных переменных.

Оценка λ^* (решетчатая процедура)

1. Задается интервал $\lambda \in [\lambda_{\min}; \lambda_{\max}]$, часто $\lambda \in [-1; 2]$.
2. С некоторым шагом $\Delta\lambda$
   1) Вычисляются значения $\tilde{y}_i(\lambda)$ и, при необходимости, $\tilde{x}_i^{(j)}(\lambda)$;
   2) Находятся оценки $\hat{\theta}_j(\lambda)$ и множественный коэффициент корреляции $\hat{R}^2(\lambda)$.
3. Строится зависимость $\hat{R}^2(\lambda)$ и находится $\lambda^* = \arg\max_{\lambda} \hat{R}^2(\lambda)$.

Объем предложения акций на фондовом рынке в зависимости от цены

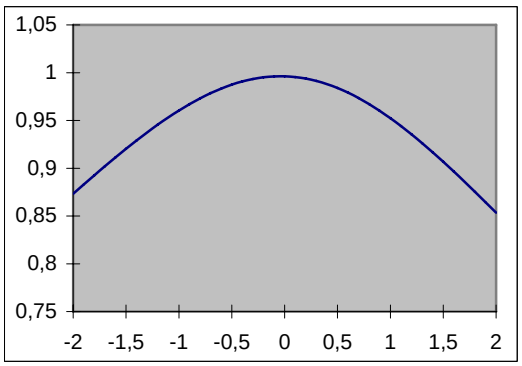
***x*, цена, $**	***y*, объем, тыс.шт.**
11	104
12	119
13	137
14	169
15	201
16	263
17	312
18	364
19	451
20	517



$$\tilde{y}(\lambda) = \theta_0 + \theta_1 x$$

x	$\tilde{y}(-1)$	$\tilde{y}(-0{,}5)$	$\tilde{y}(-0{,}1)$	$\tilde{y}(0)$	$\tilde{y}(0{,}1)$	$\tilde{y}(0{,}5)$	$\tilde{y}(1)$	$\tilde{y}(2)$
11	0,9904	1,8039	3,7151	4,6444	5,9112	18,396	103	5408
12	0,9916	1,8167	3,7992	4,7791	6,127	19,817	118	7080
13	0,9927	1,8291	3,886	4,92	6,3558	21,409	136	9384
14	0,9941	1,8462	4,013	5,1299	6,7028	24	168	14280
15	0,9950	1,8589	4,1159	5,3033	6,9949	26,355	200	20200
16	0,9962	1,8767	4,272	5,5722	7,458	30,435	262	34584
17	0,9968	1,8868	4,369	5,743	7,7589	33,327	311	48672
18	0,9973	1,8952	4,4551	5,8972	8,0348	36,158	363	66248
19	0,9978	1,9058	4,5727	6,1115	8,4254	40,474	450	101700
20	0,9981	1,912	4,6463	6,248	8,6788	43,475	516	133644

$\lambda = -1$, $\tilde{y} = 0{,}9814 + 0{,}000876x$, $\hat{R}^2 = 0{,}9604$.
$\lambda = -0{,}5$, $\tilde{y} = 1{,}6689 + 0{,}0125x$, $\hat{R}^2 = 0{,}9875$.
$\lambda = -0{,}1$, $\tilde{y} = 2{,}5062 + 0{,}1083x$, $\hat{R}^2 = 0{,}9961$.
$\lambda = 0$, $\tilde{y} = 2{,}5459 + 0{,}1864x$, $\hat{R}^2 = 0{,}9962$.
$\lambda = 0{,}1$, $\tilde{y} = 2{,}2638 + 0{,}3214x$, $\hat{R}^2 = 0{,}9955$.
$\lambda = 0{,}5$, $\tilde{y} = -15{,}341 + 2{,}8855x$, $\hat{R}^2 = 0{,}9840$.
$\lambda = 1$, $\tilde{y} = -457{,}53 + 46{,}467x$, $\hat{R}^2 = 0{,}9524$.
$\lambda = 2$, $\tilde{y} = -164270 + 13445x$, $\hat{R}^2 = 0{,}8535$.



$\max\limits_{\lambda} \hat{R}^2(\lambda) = \hat{R}^2(0) = 0{,}9962$; с точностью до сотых $\lambda^* = -0{,}04$.

$\tilde{y} = \ln y = 2{,}5459 + 0{,}1864x$, $\boxed{y = e^{2{,}5459 + 0{,}1864x} = 12{,}755e^{0{,}1864x}}$.

Замечание 1

При практической реализации решетчатой процедуры сначала можно оценить значение λ^* достаточно грубо, используя то, что при $\lambda \in (-\infty; \lambda^*)$ $\hat{R}^2(\lambda)$ монотонно возрастает, а при $\lambda \in (\lambda^*; +\infty)$ – монотонно убывает.

Замечание 2

На некоторых практических задачах λ^* находится вне интервала $\lambda \in [-1; 2]$.

$\lambda^* = 0{,}5$ – квадратичная зависимость между исходными переменными:

$$\tilde{y} = \frac{y^{0,5} - 1}{0{,}5} = \theta_0 + \theta_1 x,\ y^{0,5} = 1 + 0{,}5\theta_0 + 0{,}5\theta_1 x,\ \boxed{y = (1 + 0{,}5\theta_0)^2 + (1 + 0{,}5\theta_0)\theta_1 x + 0{,}25\theta_1^2 x^2}.$$

Бинарные результирующие показатели и связанные с ними логит- и пробит-модели

$x^{(1)}, x^{(2)}, ..., x^{(p)} \Rightarrow y = \begin{cases} 0 \\ 1 \end{cases}$

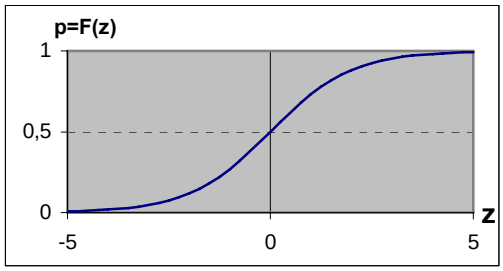
Возраст, образование, стаж, желаемый уровень зарплаты $\Rightarrow$ безработный $= \begin{cases} 0, \text{нет} \\ 1, \text{да} \end{cases}$

Если построить линейную регрессионную зависимость $y = \theta_0 + \theta_1 x^{(1)} + ... + \theta_p x^{(p)}$, непонятна интерпретация значений $\hat{y}_i = X_i\hat{\Theta}$, измеренных в непрерывной количественной шкале.

Выход: построить регрессионную зависимость вероятности $P(y=1)$ от $X\Theta$!

Непосредственная зависимость $P(y=1) = \theta_0 + \theta_1 x^{(1)} + ... + \theta_p x^{(p)}$ плоха, так как вероятность выходит за пределы отрезка [0; 1], лучше подобрать функцию $F(X\Theta)$:

$$P(y=1) = F(X\Theta) = F(z).$$



Свойства:

1. $F(z)$ – монотонно возрастает.
2. $F(z) \in [0; 1]$.
3. $F(z) \to 0$ при $z \to -\infty$.
4. $F(z) \to 0$ при $z \to +\infty$.

Из моделей бинарных показателей наиболее распространены логит- и пробит-модели!

Логит-модель: $P(y_i = 1 : X_i) = \dfrac{e^{X_i\Theta}}{1 + e^{X_i\Theta}} = \Lambda(X_i\Theta)$ – логистическая функция.

Пробит-модель: $P(y_i = 1 : X_i) = \dfrac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{X_i\Theta} e^{-t^2/2} dt = \Phi(X_i\Theta)$ – стандартная нормальная функция.

Обе функции симметричны относительно $X\Theta = 0$.

Оценивание параметров в логит- и пробит-моделях

Необходимы повторяющиеся исходные данные:

Вариант 1. Несколько наблюдений при каждом значении объясняющей переменной.

Вариант 2. Несколько наблюдений для каждого интервала группировки.

$$\left.\begin{array}{l} y_{11};\ y_{12}; ...;\ y_{1n_1};\ X_{11} = X_{12} = ... = X_{1n_1} = X_1 \\ y_{21};\ y_{22}; ...;\ y_{2n_2};\ X_{21} = X_{22} = ... = X_{2n_2} = X_2 \\ ..... \\ y_{N1}; y_{N2}; ...; y_{Nn_N};\ X_{N1} = X_{N2} = ... = X_{Nn_N} = X_N \end{array}\right\} \Rightarrow p_j = \frac{\sum\limits_{i=1}^{n_j} y_{ji}}{n_j},\ j = 1, ..., N.$$

p_j – относительная частота появления единиц для j-значения объясняющей переменной.

Логит-модель:	**Пробит-модель:**
$p = \dfrac{e^z}{1+e^z} \Rightarrow z = \ln\dfrac{p}{1-p}$	$p = НОРМСТРАСП(z) \Rightarrow z = НОРМСТОБР(p)$

$z = \theta_0 + \theta_1 x^{(1)} + ... + \theta_p x^{(p)}$, находим МНК-оценки $\hat{\theta}_0, \hat{\theta}_1, ..., \hat{\theta}_p$ – первое приближение, нужно учесть гетероскедастичность, разделив переменные на $\sqrt{D\varepsilon_j}$, далее итеративная процедура.

$D\varepsilon_j = \dfrac{F_j(1-F_j)}{n_j f_j^2} = \dfrac{1}{n_j \Lambda_j (1-\Lambda_j)}$	$D\varepsilon_j = \dfrac{НОРМСТРАСП(X_j\Theta)(1 - НОРМСТРАСП(X_j\Theta))}{n_j (НОРМРАСП(X_j\Theta; 0; 1; 0))^2}$

Доля иркутян-владельцев персональных компьютеров (p) в зависимости от среднедушевого дохода $(x$, тыс.руб./мес.$)$; объем выборки $n = 20*10 = 200$.

Логит-модель

x	p	$z = \ln\frac{p}{1-p}$	$\hat{z}$	$\Lambda(\hat{z})$	$\hat{\varepsilon}$	$D\varepsilon$	$\hat{z}$	$\Lambda(\hat{z})$	$\hat{\varepsilon}$
1	0,2	-1,386	-1,478	0,186	0,014	0,661	-1,475	0,186	0,014
2	0,1	-2,197	-1,276	0,218	-0,118	0,586	-1,267	0,220	-0,120
3	0,2	-1,386	-1,074	0,255	-0,055	0,527	-1,059	0,258	-0,058
4	0,3	-0,847	-0,872	0,295	0,005	0,481	-0,851	0,299	0,001
5	0,2	-1,386	-0,670	0,339	-0,139	0,447	-0,643	0,345	-0,145
6	0,6	0,405	-0,468	0,385	0,215	0,422	-0,435	0,393	0,207
7	0,4	-0,405	-0,266	0,434	-0,034	0,407	-0,227	0,444	-0,044
8	0,8	1,386	-0,064	0,484	0,316	0,400	-0,018	0,495	0,305
9	0,5	0	0,138	0,535	-0,035	0,402	0,190	0,547	-0,047
10	0,6	0,405	0,340	0,584	0,016	0,412	0,398	0,598	0,002
11	0,6	0,405	0,542	0,632	-0,032	0,430	0,606	0,647	-0,047
12	0,8	1,386	0,744	0,678	0,122	0,458	0,814	0,693	0,107
13	0,7	0,847	0,946	0,720	-0,020	0,496	1,022	0,735	-0,035
14	0,8	1,386	1,148	0,759	0,041	0,547	1,230	0,774	0,026
15	0,8	1,386	1,350	0,794	0,006	0,612	1,438	0,808	-0,008
16	0,9	2,197	1,552	0,825	0,075	0,693	1,646	0,838	0,062
17	0,7	0,847	1,754	0,852	-0,152	0,795	1,854	0,865	-0,165
18	0,8	1,386	1,956	0,876	-0,076	0,921	2,062	0,887	-0,087
19	0,9	2,197	2,158	0,896	0,004	1,077	2,270	0,906	-0,006
20	0,9	2,197	2,360	0,914	-0,014	1,269	2,478	0,923	-0,023

1-итерация: $\hat{z} = -1{,}680 + 0{,}202x$, $\hat{K}_d = 1 - \frac{D\varepsilon}{Dy} = 1 - \frac{0{,}0125}{0{,}0715} = 0{,}826$.

2-итерация: $\hat{z} = -1{,}681 + 0{,}208x$, $\hat{K}_d = 1 - \frac{D\varepsilon}{Dy} = 1 - \frac{0{,}0123}{0{,}0715} = 0{,}828$.

Пробит-модель

x	p	$z = НОРМСТОБР(p)$	$\hat{z}$	$\Phi(\hat{z})$	$\hat{\varepsilon}$	$D\varepsilon$	$\hat{z}$	$\Phi(\hat{z})$	$\hat{\varepsilon}$
1	0,2	-0,842	-0,879	0,190	0,010	0,209	-0,882	0,189	0,011
2	0,1	-1,282	-0,759	0,224	-0,124	0,194	-0,759	0,224	-0,124
3	0,2	-0,842	-0,639	0,262	-0,062	0,182	-0,636	0,262	-0,062
4	0,3	-0,524	-0,518	0,302	-0,002	0,173	-0,513	0,304	-0,004
5	0,2	-0,842	-0,398	0,345	-0,145	0,166	-0,390	0,348	-0,148
6	0,6	0,253	-0,278	0,391	0,209	0,162	-0,267	0,395	0,205
7	0,4	-0,253	-0,157	0,438	-0,038	0,158	-0,144	0,443	-0,043
8	0,8	0,842	-0,037	0,485	0,315	0,157	-0,021	0,492	0,308
9	0,5	0	0,083	0,533	-0,033	0,157	0,102	0,541	-0,041
10	0,6	0,253	0,204	0,581	0,019	0,159	0,225	0,589	0,011
11	0,6	0,253	0,324	0,627	-0,027	0,163	0,348	0,636	-0,036
12	0,8	0,842	0,444	0,672	0,128	0,169	0,471	0,681	0,119
13	0,7	0,524	0,565	0,714	-0,014	0,177	0,595	0,724	-0,024
14	0,8	0,842	0,685	0,753	0,047	0,187	0,718	0,764	0,036
15	0,8	0,842	0,805	0,790	0,010	0,200	0,841	0,800	0,000
16	0,9	1,282	0,926	0,823	0,077	0,216	0,964	0,832	0,068
17	0,7	0,524	1,046	0,852	-0,152	0,236	1,087	0,861	-0,161
18	0,8	0,842	1,166	0,878	-0,078	0,262	1,210	0,887	-0,087
19	0,9	1,282	1,287	0,901	-0,001	0,294	1,333	0,909	-0,009
20	0,9	1,282	1,407	0,920	-0,020	0,334	1,456	0,927	-0,027

1-итерация: $\hat{z} = -1{,}000 + 0{,}120x$, $\hat{K}_d = 1 - \frac{D\varepsilon}{Dy} = 1 - \frac{0{,}0127}{0{,}0715} = 0{,}822$.

2-итерация: $\hat{z} = -1{,}005 + 0{,}123x$, $\hat{K}_d = 1 - \frac{D\varepsilon}{Dy} = 1 - \frac{0{,}0125}{0{,}0715} = 0{,}824$.

Типовые задания для контрольных работ

Задание 1

Имеются в наличии данные по 10 фирмам, продающим компакт-диски, – объемы продаж, тыс. шт. / мес. (y), цены, руб. ($x^{(1)}$), вложения в рекламу, тыс. руб. / мес. ($x^{(2)}$).

1) Построить регрессионную зависимость $y = \theta^{(0)} + \theta^{(1)}x^{(1)} + \theta^{(2)}x^{(2)}$.
2) Проверить гипотезу о значимости регрессоров при уровнях значимости $\alpha = 0{,}05$ и $\alpha = 0{,}01$.
3) Построить доверительные интервалы для коэффициентов регрессии $\theta^{(0)}$, $\theta^{(1)}$, $\theta^{(2)}$ с вероятностью $\gamma = 0{,}95$.
4) Вычислить множественный коэффициент корреляции, проверить гипотезу о значимости модели при уровнях значимости $\alpha = 0{,}05$ и $\alpha = 0{,}01$.

y	15	18	10	17	14	26	11	25	6	12
$x^{(1)}$	80	100	90	75	120	85	100	70	120	75
$x^{(2)}$	25	40	0	10	60	80	10	0	15	5

Задание 2

Имеются данные по ценам на квартиры, тыс.руб. (y) в зависимости от общей площади, м^2 ($x^{(1)}$) и площади кухни, м^2 ($x^{(2)}$).

1) Построить регрессионную зависимость $y = \theta^{(0)} + \theta^{(1)}x^{(1)} + \theta^{(2)}x^{(2)}$.
2) Обосновать наличие гетероскедастичности.
3) С помощью обобщенного метода наименьших квадратов построить зависимость с учетом гетероскедастичности.

y	995	1200	780	1150	750	1650	1880	930	2400	835
$x^{(1)}$	46	48	30	48	31	73	88	44	73	31
$x^{(2)}$	6	8	6	9	5	9	12	5	12	6

Задание 3

Известны посезонные данные по объемам продаж сноубордов, шт. (y) в зависимости от цены, тыс.руб. (x). Построить линейную регрессионную модель с учетом сезонности.

	2003				2004				2005			
	весна	лето	осень	зима	весна	лето	осень	зима	весна	лето	осень	зима
y	49	67	101	163	86	43	190	204	118	50	201	216
x	4,5	5	6	6,5	5	5,5	5,5	7	3,5	5	5	6

Задание 4

Известны данные по числу преступлений на 100 тысяч человек, тыс. (y) в зависимости от среднедушевого дохода, тыс.руб. (x) по 10 регионам России. Построить модель с помощью линеаризующего преобразования Бокса-Кокса. Величину λ найти с точностью до десятых.

y	4,62	2,87	3,55	2,34	2,30	1,92	1,85	1,30	2,39	1,38
x	4,9	6,5	6,9	7,2	7,6	8,8	9,5	11,2	15,6	17,4

Вопросы к экзамену

1. Эконометрика: цели, методы, проблемы, типы переменных.
2. Структурная и приведенная форма модели.
3. Этапы эконометрического исследования.
4. Классическая линейная модель множественной регрессии. Метод наименьших квадратов.
5. Свойства оценок. Состоятельность, несмещенность, эффективность.
6. Проверка гипотезы о значимости регрессоров и о значимости модели, доверительные интервалы для коэффициентов регрессии.
7. Возможные ошибки спецификации модели. Мультиколлинеарность.
8. Методы устранения мультиколлинеарности.
9. Обобщенная линейная модель множественной регрессии. Обобщенный метод наименьших квадратов.
10. Обобщенная линейная модель множественной регрессии с гетероскедастичными остатками.
11. Обобщенная линейная модель множественной регрессии с автокоррелированными остатками.
12. Практические рекомендации по построению регрессионной модели. Практически реализуемый обобщенный метод наименьших квадратов.
13. Точечный прогноз.
14. Интервальный прогноз.
15. Реалистическая ситуация. Метод скользящего экзамена.
16. Проблема неоднородности данных. Способы решения.
17. Дамми-переменные.
18. Учет эффекта взаимодействия сопутствующих переменных.
19. Проверка регрессионной однородности двух групп наблюдений. Критерий Чоу.
20. Нелинейные модели, поддающиеся непосредственной линеаризации.
21. Подход Бокса-Кокса подбора линеаризующего преобразования.
22. Бинарные результирующие показатели. Логит- и пробит-модели. Их оценивание.

Содержание

От автора…… 3
Лекция 1. Введение в эконометрику, системы одновременных уравнений, структурная и приведенная формы…….. 4
Лекция 2. Пример «потребление, инвестиции, ВВП», этапы эконометрического исследования…… 6
Лекция 3. Регрессионный анализ, классическая линейная модель множественной регрессии, метод наименьших квадратов…….. 8
Лекция 4. Свойства оценок, проверка гипотез о значимости регрессора и модели в целом, доверительные интервалы…… 10
Лекция 5. Ошибки спецификации модели, мультиколлинеарность, методы ее устранения…… 12
Лекция 6. Обобщенная линейная модель множественой регрессии. Обобщенный метод наименьших квадратов. Случай гетероскедастичных остатков…….…. 14
Лекция 7. Случай автокоррелированных остатков. Проверка гипотез о гетероскедастичности и автокорреляции…… 16
Лекция 8. Практические рекомендации по построению регрессионной модели. Процедура Кохрейна-Оркатта…….. 18
Лекция 9. Точечный и интервальный прогноз в моделях линейной регрессии. Метод скользящего экзамена…… 20
Лекция 10. Линейные регрессионные модели с переменной структурой. Использование дамми-переменных…….. 22
Лекция 11. Взаимодействие сопутствующих переменных. Проверка регрессионной однородности двух групп наблюдений…… 24
Лекция 12. Нелинейные модели, поддающиеся непосредственной линеаризации…… 26
Лекция 13. Выбор вида зависимости. Метод проб и ошибок. Процедура Бокса-Кокса……..…… 28
Лекция 14. Бинарные результирующие показатели. Логит- и пробит-модели. Их оценивание…… 30
Типовые задания для контрольных работ…… 32
Вопросы для экзамена…….. 33

Александр Юрьевич Филатов,
e-mail: fial@irlan.ru, ICQ 10793366

Другие авторские разработки в области математической экономики выложены на сайтах
http://polnolunie.baikal.ru/me/metrix.htm
http://polnolunie.baikal.ru/me/mat_ec.htm
http://matec.isu.ru
http://fial_.livejournal.com

КОНСПЕКТ ЛЕКЦИЙ ПО ЭКОНОМЕТРИКЕ
учебное пособие

Редактор Э.А. Невзорова

Темплан 2006. Поз.16.

Подписано в печать 14.02.2006. Формат 60×84 1/16.
Бумага писчая белая. Печать трафаретная. Уч.-изд.л. 2,3.
Тираж 150 экз.

Редакционно-издательский отдел
Иркутского государственного университета
664003, Иркутск, бул. Гагарина, 36