МЕЖГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«БЕЛОРУССКО-РОССИЙСКИЙ УНИВЕРСИТЕТ»

Кафедра «Автоматизированные системы управления»

СИСТЕМЫ АНАЛИТИЧЕСКОГО ПРОГРАММИРОВАНИЯ

Методические рекомендации к лабораторным работам для студентов специальности 1-40 80 02 «Системный анализ, управление и обработка информации» очной и заочной форм обучения

Часть 2

Могилев 2020

УДК 004.43
ББК 32.973-018
С40

Рекомендовано к изданию
учебно-методическим отделом
Белорусско-Российского университета

Одобрено кафедрой «Автоматизированные системы управления» «10» ноября 2020 г., протокол № 3

Составители: д-р техн. наук, доц. А. И. Якимов;
канд. техн. наук Е. А. Якимов

Рецензент канд. техн. наук, доц. И. В. Лесковец

Даны методические указания к выполнению лабораторных работ по дисциплине «Системы аналитического программирования», а также приведены контрольные вопросы и список литературы для подготовки.

Учебно-методическое издание

СИСТЕМЫ АНАЛИТИЧЕСКОГО ПРОГРАММИРОВАНИЯ

Часть 2

Ответственный за выпуск	А. И. Якимов
Корректор	И. В. Голубцова
Компьютерная верстка	Н. П. Полевничая

Подписано в печать . Формат 60 × 84/16. Бумага офсетная. Гарнитура Таймс.
Печать трафаретная. Усл. печ. л. . Уч.-изд. л. . Тираж 16 экз. Заказ №

Издатель и полиграфическое исполнение:
Межгосударственное образовательное учреждение высшего образования
«Белорусско-Российский университет».
Свидетельство о государственной регистрации издателя,
изготовителя, распространителя печатных изданий
№ 1/156 от 07.03.2019.
Пр-т Мира, 43, 212022, Могилев.

Содержание

Введение 4

1 Лабораторная работа № 7. Моделирование взаимосвязей с помощью систем одновременных регрессионных уравнений 5

2 Лабораторная работа № 8. Моделирование и сценарное прогнозирование динамики показателей на основе многомерных временных рядов 19

3 Лабораторная работа № 9. Нелинейные модели для классификации и регрессии 23

4 Лабораторная работа № 10. Вероятностный вывод для дискретных моделей 34

5 Лабораторная работа № 11. Обучение с неполными данными 44

Список литературы 48

Введение

Целью преподавания дисциплины «Системы аналитического программирования» является приобретение студентами теоретических знаний и практических навыков в области основ теории обучения машин, современных методов восстановления зависимостей по эмпирическим данным, включая дискриминантный, кластерный и регрессионный анализ, овладение навыками практического решения задач интеллектуального анализа данных.

1 Лабораторная работа № 7. Моделирование взаимосвязей с помощью систем одновременных регрессионных уравнений

Цель работы: овладеть на практике основами моделирования взаимосвязей с помощью систем одновременных регрессионных уравнений.

Порядок выполнения работы

1 Изучить основные теоретические положения, сделав необходимые выписки в конспект.
2 Получить задание у преподавателя, выполнить вариант задания.
3 Сделать выводы по результатам работы.
4 Оформить отчет.

Требования к отчету

1 Цель работы.
2 Постановка задачи.
3 Результаты разработки программного обеспечения.
4 Выводы.

Методические указания

Системы одновременных уравнений могут быть представлены в *структурной* и *приведенной* формах.

Основными составляющими обеих форм записи являются *эндогенные* и *экзогенные* переменные. Эндогенные переменные (y) определяются внутри модели и являются зависимыми переменными. Экзогенные переменные (x) определяются вне системы и являются независимыми переменными. Предполагается, что экзогенные переменные не коррелируют с ошибкой регрессии в соответствующем уравнении.

Простейшая структурная форма модели имеет вид:

$$\begin{cases} y_1 = b_{12} y_2 + a_{11} x_1 + \varepsilon_1; \\ y_2 = b_{21} y_1 + a_{22} x_2 + \varepsilon_2. \end{cases}$$

Классификация переменных на эндогенные и экзогенные зависит от теоретической концепции принятой модели. Экономические переменные могут выступать в одних моделях как эндогенные, в других – как экзогенные переменные. Внеэкономические переменные, например климатические условия, входят в систему как экзогенные переменные. В качестве экзогенных могут рассматриваться значения эндогенных переменных за предшествующий период времени (лаговые переменные). Так, потребление текущего года может зависеть не только от ряда экономических факторов, но и от уровня потребления в преды-

дущем году. Целесообразно в качестве экзогенных переменных выбирать те, которые могут быть объектом регулирования.

Структурная форма модели в правой части содержит:

– коэффициенты при эндогенной переменной – b_i;

– коэффициенты при экзогенной переменной – a_j;

– переменные модели выражены в отклонениях от среднего уровня, т. е. под x подразумевается $x - \bar{x}$, а под y – соответственно $y - \bar{y}$. Поэтому свободный член в каждом уравнении отсутствует.

Использование МНК для оценивания коэффициентов структурной модели дает, как принято считать в теории, смещенные и несостоятельные оценки. Поэтому обычно для определения структурных коэффициентов структурная форма модели преобразуется в приведенную.

Приведенная форма модели представляет собой систему линейных функций эндогенных переменных от экзогенных. Для простейшей структурной модели соответствующая приведенная модель имеет вид:

$$\begin{cases} y_1 = \delta_{11}x_1 + \delta_{12}x_2 + u_1; \\ y_2 = \delta_{21}x_1 + \delta_{22}x_2 + u_2. \end{cases}$$

Ее можно получить, выразив y_2 из первого уравнения структурной модели:

$$y_2 = \frac{y_1 - a_{11}x_1}{b_{12}}.$$

Выполнив подстановку во второе уравнение, после необходимых преобразований получим

$$y_1 = \frac{a_{11}}{1 - b_{12}b_{21}}x_1 + \frac{a_{22}b_{12}}{1 - b_{12}b_{21}}x_2.$$

Аналогично, выразив y_1 из второго уравнения и произведя подстановку в первое, имеем

$$y_2 = \frac{a_{11}b_{21}}{1 - b_{12}b_{21}}x_1 + \frac{a_{22}}{1 - b_{12}b_{21}}x_2.$$

Применяя МНК, можно оценить δ, а затем найти значения эндогенных переменных через экзогенные.

Эконометрические модели обычно включают в систему не только уравнения, но и тождества. Они устанавливают соотношения между эндогенными переменными, но не содержат случайных составляющих. Например, Т. Хаавелмо в 1947 г., исследуя линейную зависимость потребления (c) от дохода (y), предложил одновременно учитывать и тождество дохода. В этом случае модель имеет вид:

$$\begin{cases} c = a + by; \\ y = c + x, \end{cases}$$

где a и b – параметры линейной зависимости c от y;

x – инвестиции в основной капитал и запасы экспорта и импорта.

Оценки параметров должны учитывать тождество дохода в отличие от параметров обычной линейной регрессии.

Проблема идентификации

При переходе от приведенной формы модели к структурной исследователь сталкивается с проблемой идентификации. Идентификация – это единственность соответствия между приведенной и структурной формой модели. В зависимости от условий определения структурных коэффициентов модели по приведенным коэффициентам любая структурная модель может быть отнесена к одному из трех классов: идентифицируемая, неидентифицируемая и сверхидентифицируемая.

Модель идентифицируема, если все структурные коэффициенты однозначно определяются через приведенные коэффициенты.

Модель неидентифицируема, если структурные коэффициенты невозможно найти по приведенным коэффициентам.

Модель сверхидентифицируема, если структурные коэффициенты, выраженные через приведенные коэффициенты, имеют два и более числовых значения.

В идентифицируемой модели количество структурных и приведенных коэффициентов одинаково. Если структурных коэффициентов больше (меньше), чем приведенных, то модель, соответственно, неидентифицируема (сверхидентифицируема).

Проверка структурной модели на идентифицируемость позволяет установить степень возможности оценивания коэффициентов структурных уравнений по коэффициентам приведенных уравнений.

Структурная модель всегда представляет собой систему совместных уравнений, каждое из которых необходимо проверять на идентификацию. Модель считается идентифицируемой, если каждое ее уравнение идентифицируемо. Если хотя бы одно из уравнений системы неидентифицируемо, то и вся модель считается неидентифицируемой. Сверхидентифицируемая модель содержит хотя бы одно сверхидентифицируемое уравнение.

Рассмотрим *необходимое условие идентификации*.

Если обозначить число эндогенных переменных уравнения через H, а число экзогенных переменных, которые содержатся в системе, но *не входят* в данное уравнение, – через D, то необходимое условие идентифицируемости модели может быть записано в виде следующего счетного правила:

$D + 1 = H$ – уравнение идентифицируемо;

$D + 1 < H$ – уравнение неидентифицируемо;

$D + 1 > H$ – уравнение сверхидентифицируемо.

Предположим, рассматривается следующая система одновременных уравнений:

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + a_{11}x_1 + a_{12}x_2; \\ y_2 = b_{21}y_1 + a_{21}x_1 + a_{22}x_2 + a_{23}x_3; \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{33}x_3 + a_{34}x_4. \end{cases}$$

Для первого уравнения $H = 3$, $D = 2$. Уравнение идентифицируемо.

Для второго уравнения $H = 2$, $D = 1$. Уравнение идентифицируемо.

Для третьего уравнения $H = 3$, $D = 2$. Уравнение идентифицируемо.

Для оценки параметров структурной модели система должна быть идентифицируема или сверхидентифицируема.

Более точным (*достаточным*) *условием идентификации* является следующее.

Уравнение идентифицируемо, если по отсутствующим в нем переменным (эндогенным и экзогенным) можно из коэффициентов при них в других уравнениях системы получить матрицу, определитель которой не равен нулю, а ранг матрицы не меньше, чем число эндогенных переменных в системе без единицы.

Для решения идентифицируемых уравнений применяется косвенный метод наименьших квадратов, для решения сверхидентифицированных – двухшаговый метод наименьших квадратов.

Косвенный МНК состоит в следующем:

1) составляют приведенную форму модели и определяют численные значения параметров для каждого ее уравнения в отдельности с помощью обычного МНК;

2) путем алгебраических преобразований переходят от приведенной формы к уравнениям структурной формы модели, получая тем самым численные оценки структурных параметров.

Двухшаговый МНК заключается в следующем:

1) составляют приведенную форму модели и определяют численные значения параметров каждого ее уравнения в отдельности с помощью обычного МНК;

2) выявляют эндогенные переменные, находящиеся в правой части структурного уравнения (параметры которого определяют двухшаговым МНК) и находят расчетные значения по полученным на первом этапе соответствующим уравнениям приведенной формы модели;

3) с помощью обычного МНК определяют параметры каждого структурного уравнения в отдельности, используя в качестве исходных данных фактические значения предопределенных переменных и расчетные значения эндогенных переменных, стоящих в правой части данного структурного уравнения, полученные на втором этапе.

Пример 1 – Рассмотрим следующую структурную модель:

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + a_{11}x_1 + a_{12}x_2; \\ y_2 = b_{21}y_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4; \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{31}x_1 + a_{32}x_2. \end{cases}$$

Проверим каждое уравнение системы на необходимое и достаточное условие идентификации.

Для первого уравнения $H = 3$ (y_1, y_2, y_3) и $D = 2$ (x_3, x_4) отсутствуют, т. е. $D + 1 = H$, необходимое условие идентификации выдержано. Для проверки на достаточное условие идентификации заполним таблицу коэффициентов при отсутствующих в первом уравнении переменных (таблица 1.1).

Таблица 1.1 – Коэффициенты при отсутствующих в первом уравнении переменных

Уравнение	Переменные	
	x_3	x_4
2	a_{23}	a_{24}
3	0	0

Определитель матрицы

$$\begin{pmatrix} a_{23} & a_{24} \\ 0 & 0 \end{pmatrix}$$

равен нулю. Достаточное условие идентифицируемости не выполняется, и первое уравнение нельзя считать идентифицируемым.

Для второго уравнения $H = 2$ (y_1, y_2) и $D = 1$ (x_1), т. е. $D + 1 = H$, необходимое условие идентификации выдержано. Для проверки на достаточное условие идентификации заполним таблицу коэффициентов при отсутствующих во втором уравнении переменных (таблица 1.2).

Таблица 1.2 – Коэффициенты при отсутствующих во втором уравнении переменных

Уравнение	Переменные	
	y_3	x_1
1	b_{31}	a_{11}
3	-1	a_{31}

Определитель матрицы не равен нулю, следовательно, ранг матрицы равен 2, что не меньше числа эндогенных переменных в системе минус 1. Достаточное условие идентифицируемости выполняется, и второе уравнение точно идентифицируемо.

В третьем уравнении системы $H = 3$ и $D = 2$. Необходимое условие выполняется. Для проверки на достаточное условие идентификации построена таблица 1.3.

Таблица 1.3 – Коэффициенты при отсутствующих в третьем уравнении переменных

Уравнение	Переменные	
	x_3	x_4
1	0	0
2	a_{23}	a_{24}

Делаем вывод о невыполнении достаточного условия идентифицируемости. Уравнение неидентифицируемо.

Вывод: структурная модель неидентифицируема.

Пример 2 – Предположим, что следующая структурная модель признана идентифицируемой:

$$\begin{cases} y_1 = b_{13}y_3 + a_{11}x_1 + a_{13}x_3; \\ y_2 = b_{21}y_1 + a_{22}x_2 + b_{23}y_3; \\ y_3 = b_{32}y_1 + a_{31}x_1 + a_{33}x_3. \end{cases}$$

Исходя из приведенной формы (уравнения которой найдены через *Анализ данных/Регрессия*)

$$y_1 = 2x_1 + 4x_2 + 10x_3;$$

$$y_2 = 3x_1 - 6x_2 + 2x_3;$$

$$y_3 = -5x_1 + 8x_2 + 5x_3,$$

найти структурные коэффициенты модели.

1 Из третьего уравнения приведенной формы выразим x_2 (так как его нет в первом уравнении структурной формы):

$$x_2 = \frac{y_3 + 5x_1 - 5x_3}{8}.$$

Данное выражение содержит переменные y_3, x_1 и x_3, которые входят в правую часть первого уравнения структурной формы модели (СФМ). Подставим полученное выражение x_2 в первое уравнение приведенной формы модели (ПФМ):

$$y_1 = 2x_1 + 4 \cdot \frac{y_3 + 5x_1 - 5x_3}{8} + 10x_3.$$

Откуда получим первое уравнение СФМ в виде

$$y_1 = 0{,}5y_3 + 4{,}5x_1 + 7{,}5x_3.$$

2 Во втором уравнении СФМ нет переменных x_1 и x_3. Структурные параметры второго уравнения СФМ можно определить в два этапа.

Первый этап: выразим x_1 в данном случае из первого или третьего уравнения ПФМ. Например, из первого уравнения

$$x_1 = \frac{y_1 - 4x_2 - 10x_3}{2} = 0{,}5y_1 - 2x_2 - 5x_3.$$

Подстановка данного выражения во второе уравнение ПФМ не решило бы задачу до конца, т. к. в выражении присутствует x_3, которого нет в СФМ.

Выразим x_3 из третьего уравнения ПФМ:

$$x_3 = \frac{y_3 + 5x_1 - 8x_2}{5}.$$

Подставим его в выражение для x_1:

$$x_1 = 0{,}5y_1 - 2x_2 - 5 \cdot \frac{y_3 + 5x_1 - 8x_2}{5} = 0{,}5y_1 - y_3 + 6x_2 - 5x_1;$$

$$x_1 = \frac{0{,}5y_1 - y_3 + 6x_2}{6}.$$

Второй этап: аналогично, чтобы выразить x_3 через искомые y_3, x_1 и x_2, заменим в выражении x_3 значение x_1 на полученное из первого уравнения ПФМ:

$$x_3 = \frac{y_3 + 5 \cdot (0{,}5y_1 - 2x_2 - 5x_3) - 8x_2}{5} = 0{,}2y_3 + 0{,}5y_1 - 3{,}6x_2 - 5x_3.$$

Следовательно, $x_3 = 0{,}333y_3 + 0{,}083y_1 - 0{,}6x_2$.

Подставим полученные x_1 и x_3 во второе уравнение ПФМ:

$$y_2 = 3 \cdot \frac{0{,}5y_1 - y_3 + 6x_2}{6} - 6x_2 + 2 \cdot (0{,}033y_3 + 0{,}083y_1 - 0{,}6x_2).$$

В результате имеем второе уравнение СФМ

$$y_2 = 0{,}416y_1 - 0{,}434y_3 - 4{,}2x_2.$$

3 Из второго уравнения ПФМ выразим x_2, т. к. его нет в третьем уравнении СФМ:

$$x_2 = \frac{-y_2 + 3x_1 + 2x_3}{6} = -0{,}167y_2 + 0{,}5x_1 + 0{,}333x_3.$$

Подставим полученное выражение в третье уравнение ПФМ:

$$y_3 = -5x_1 + 8 \cdot (-0{,}167y_2 + 0{,}5x_1 + 0{,}333x_3) + 5x_3.$$

В результате имеем третье уравнение СФМ

$$y_3 = -1{,}336y_2 - x_1 + 7{,}644x_3.$$

Таким образом, СФМ примет вид:

$$y_1 = 0{,}5y_3 + 4{,}5x_1 + 7{,}5x_3;$$

$$y_2 = 0{,}416y_1 - 0{,}434y_3 - 4{,}2x_2;$$

$$y_3 = -1{,}336y_2 - x_1 + 7{,}644x_3.$$

Пример 3 – Рассматривается следующая модель:

$$C_t = a_1 + b_{11}Y_t + b_{12}C_{t-1} + u_1;$$

$$I_t = a_2 + b_{21}Y_t + b_{22}I_{t-1} + u_2;$$

$$r_t = a_3 + b_{31}Y_t + b_{32}M_{t-1} + b_{35}r_{t-1} + u_3;$$

$$Y_t = C_t + I_t + G_t,$$

где C_t – расходы на потребление в период t;

Y_t – совокупный доход в период t;

I_t – инвестиции в период t;

r_t – процентная ставка в период t;

M_t – денежная масса в период t;

G_t – государственные расходы в период t;

C_{t-1} – расходы на потребление в период $t-1$;

I_{t-1} – инвестиции в период $t-1$;

u_1, u_2, u_3 – случайные ошибки.

Решение

В этой системе все уравнения сверхидентифицированы. Для оценки параметров каждого уравнения применяем двухшаговый МНК.

Шаг 1. Запишем приведенную форму модели в общем виде:

$$C_t = A_1 + A_2 C_{t-1} + A_3 I_{t-1} + A_4 M_t + A_5 G_t + v_1;$$

$$I_t = B_1 + B_2 C_{t-1} + B_3 I_{t-1} + B_4 M_t + B_5 G_t + v_2;$$

$$Y_t = D_1 + D_2 C_{t-1} + D_3 I_{t-1} + D_4 M_t + D_5 G_t + v_3;$$

$$r_t = E_1 + E_2 C_{t-1} + E_3 I_{t-1} + E_4 M_t + E_5 G_t + v_4,$$

где v_1, v_2, v_3, v_4 – случайные ошибки.

Определим параметры каждого уравнения отдельно обычным МНК (*Сервис/Анализ данных/Регрессия*).

Затем по созданным уравнениям регрессии найдем расчетные значения эндогенных переменных Y_t и r_t. Обозначим их соответственно $\hat{Y}_t$ и $\hat{r}_t$.

Шаг 2. В исходных структурных уравнениях заменим эндогенные переменные, выступающие в роли факторных признаков, их расчетными значениями:

$$C_t = a_1 + b_{11}\hat{Y}_t + b_{12} C_{t-1} + u_1;$$

$$I_t = a_2 + b_{21}\hat{r}_t + b_{22} I_{t-1} + u_2;$$

$$r_t = a_3 + b_{31}\hat{Y}_t + b_{32} M_{t-1} + u_3.$$

Применяем к каждому из полученных уравнений обычный МНК, заканчиваем процедуру параметризации.

Практическое задание

Необходимо:

1) выделить эндогенные и экзогенные переменные;

2) применив необходимое и достаточное условия идентификации, определить, идентифицировано ли каждое из уравнений системы;

3) если система идентифицируется, записать приведенную форму модели;

4) определить коэффициенты приведенной формы модели;

5) определить коэффициенты структурной формы модели.

Примечание – Исходные статистические данные для факторов, используемых в моделях, брать из таблиц 1.4 и 1.5.

Таблица 1.4 – Исходные данные к практическому заданию (часть 1)

Текущий период	Процентная ставка *R*, %	ВВП *Y*, млн р.	Денежная масса *M*, млн р.	Внутренние инвестиции *I*, млн р.	Национальный доход *Y*, млн р.	Расходы на личное потребление *C*, млн р.	Валовая прибыль экономики *Q*, млн р.
1	10,01	1 398,5	0,12	211	310	450	725,6
2	6,25	19 005,5	0,95	2670	5328	7500	11 390,5
3	6,00	171 509,5	9,20	27 125	49730	40600	76961,7
4	7,14	610745,2	33,20	108 810	172380	124 000	251 944,4
5	8,83	152404,9	98,70	266 974	437 007	310000	662 374,4
6	8,27	2 145655,5	220,80	375 998	558 500	260 000	790819,2
7	8,44	2478594,1	288,30	408 797	711600	390 000	881 001,1
8	8,35	2741051,2	374,10	407 086	686 000	490 000	1 032 768,6
9	7,99	4757233,7	448,30	970 439	1 213 600	990 000	2 050 276,8
10	7,83	7 063 392,8	704,70	1 165 181	2 097 700	1 650 000	3 033 247,2

Таблица 1.5 – Исходные данные к практическому заданию (часть 2)

Текущий период	Индекс стоимости жизни *P*, %	Объем продукции промышленности *R*, млн р.	Государственные расходы *G*, млн р.	Доля импорта в ВВП *M*	Реальный объем чистого экспорта *X*, млн р.	Налоги *T*, млн р.	Запас капитала *K*, млн р.	Зарплата *S*, тыс. р.
1	200	600	348	0,1295	186	152	325	0,6
2	210	1 300	5970	0,4826	1 1847	3893	4550	6,0
3	220	18500	57674	0,3050	65524	28672	34965	58,7
4	238	129000	230385	0,2321	169 534	85044	133209	220,4
5	195	384 000	486 112	0,2429	426 735	253 326	327 941	472,4
6	208	1 108 000	652 700	0,2060	532 239	380 685	454 369	790,2
7	229	1 469 000	839 000	0,2094	592 332	471 657	482451	950,2
8	204	1 626 000	842 100	0,2350	840 596	520 534	485 452	1051,5
9	180	1 707 000	1258 000	0,2689	2 090 687	875 751	766 672	1522,6
10	181	3 150000	1960 100	0,2493	3232388	1 348 178	1 293 750	2223,4

Вариант 1

Модель денежного рынка

$$R_t = a_1 + b_{11}M_t + b_{12}Y_t + \varepsilon_1;$$

$$Y_t = a_2 + b_{21}R_t + b_{22}I_t + \varepsilon_2,$$

где R – процентная ставка;
Y – ВВП;
M – денежная масса;
I – внутренние инвестиции;
t – текущий период.

Вариант 2

Модель Менгеса

$$Y_t = a_1 + b_{11}Y_{t-1} + b_{12}I_t + \varepsilon_1;$$

$$I_t = a_2 + b_{21}Y_t + b_{22}Q_t + \varepsilon_2;$$

$$C_t = a_3 + b_{31}Y_t + b_{32}C_{t-1} + b_{33}P_t + \varepsilon_3;$$

$$Q_t = a_4 + b_{41}Q_{t-1} + b_{42}R_t + \varepsilon_4,$$

где Y – национальный доход;
C – расходы на личное потребление;
I – чистые инвестиции;
Q – валовая прибыль экономики;
P – индекс стоимости жизни;
R – объем продукции промышленности;
t – текущий период;
$(t-1)$ – предыдущий период.

Вариант 3

Модифицированная модель Кейнса

$$C_t = a_1 + b_{11}Y_t + b_{12}Y_{t-1} + \varepsilon_1;$$

$$I_t = a_2 + b_{21}Y_t + \varepsilon_2;$$

$$Y_t = C_t + I_t + G_t,$$

где C – потребление;
Y – ВВП;
I – валовые инвестиции;
G – государственные расходы;
t – текущий период;
$(t-1)$ – предыдущий период.

Вариант 4

Модель мультипликатора-акселератора

$$C_t = a_1 + b_{11}R_t + b_{12}C_{t-1} + \varepsilon_1;$$

$$I_t = a_2 + b_{21}(R_t - R_{t-1}) + \varepsilon_2;$$

$$R_t = C_t + I_t,$$

где C – расходы на личное потребление;
R – доход;
t – текущий период;
$(t-1)$ – предыдущий период.

Вариант 5

Конъюнктурная модель имеет вид:

$$C_t = a_1 + b_{11}Y_t + b_{12}C_{t-1} + \varepsilon_1;$$

$$I_t = a_2 + b_{21}r_t + b_{22}I_{t-1} + \varepsilon_2;$$

$$r_t = a_3 + b_{31}Y_t + b_{32}M_t + \varepsilon_3;$$

$$Y_t = C_t + I_t + G_t,$$

где C – расходы на потребление;
Y – ВВП;
I – инвестиции;
r – процентная ставка;
M – денежная масса;
G – государственные расходы;
t – текущий период;
$t-1$ – предыдущий период.

Вариант 6

Макроэкономическая модель (упрощенная версия модели Клейна)

$$C_t = a_1 + b_{11} \cdot Y_t + b_{13} \cdot T_t + \varepsilon_1;$$

$$I_t = a_2 + b_{21} \cdot Y_t + b_{24} \cdot K_{t-1} + \varepsilon_2;$$

$$Y_t = C_t + I_t,$$

где C – потребление;
Y – доход;
T – налоги;
K – запас капитала;
t – текущий период;
$(t-1)$ – предыдущий период.

Вариант 7

Макроэкономическая модель экономики России (одна из версий)

$$C_t = a_1 + b_{11} Y_t + b_{12} C_{t-1} + \varepsilon_1;$$

$$I_t = a_2 + b_{21} Y_t + b_{23} r_t + \varepsilon_2;$$

$$r_t = a_3 + b_{31} Y_t + b_{34} M_t + b_{35} r_{t-1} + \varepsilon_3;$$

$$Y_t = C_t + I_t + G_t,$$

где C – расходы на потребление;
Y – ВВП;
I – инвестиции;
r – процентная ставка;
M – денежная масса;
G – государственные расходы;
t – текущий период;
$(t-1)$ – предыдущий период.

Вариант 8

Одна из версий модифицированной модели Кейнса

$$C_t = a_1 + b_{11}Y_t + b_{12}Y_{t-1} + \varepsilon_1;$$

$$I_t = a_2 + b_{21}Y_t + b_{22}Y_{t-1} + \varepsilon_2;$$

$$Y_t = C_t + I_t + G_t,$$

где C – расходы на личное потребление;
Y – национальный доход;
I – чистые инвестиции;
G – государственные расходы;
t – текущий период,
$(t-1)$ – предыдущий период.

Вариант 9

Модель денежного и товарного рынков

$$R_t = a_1 + b_{11}Y_t + b_{13}r_t + \varepsilon_1;$$

$$Y_t = a_2 + b_{21}R_t + b_{23}I_t + b_{25}G_t + \varepsilon_2;$$

$$I_t = a_3 + b_{31}R_t + \varepsilon_2,$$

где Y – ВВП;
I – внутренние инвестиции;
R – процентная ставка;
M – денежная масса;
G – реальные государственные расходы.

Вариант 10

Модифицированная модель Кейнса

$$C_t = a_1 + b_{11}Y_t + \varepsilon_1;$$

$$I_t = a_2 + b_{21}Y_t + b_{22}Y_{t-1} + \varepsilon_2;$$

$$Y_t = C_t + I_t + G_t,$$

где C – расходы на личное потребление;
Y – доход;
I – инвестиции;
G – государственные расходы;

t – текущий период;
$(t-1)$ – предыдущий период.

Контрольные вопросы

1 Что такое эндогенные переменные?
2 Что такое экзогенные переменные?
3 Приведите условие идентифицируемой модели.
4 Приведите условие неидентифицируемой модели.
5 Приведите условие сверхидентифицируемой модели.

2 Лабораторная работа № 8. Моделирование и сценарное прогнозирование динамики показателей на основе многомерных временных рядов

Цель работы: ознакомиться с моделированием и сценарным прогнозированием динамики показателей на основе многомерных временных рядов.

Порядок выполнения работы

1 Изучить основные теоретические положения, сделав необходимые выписки в конспект.
2 Получить задание у преподавателя, выполнить типовые задания.
3 Исследовать построение простой линейной регрессии.
4 Оформить отчет.

Требования к отчету

1 Цель работы.
2 Постановка задачи.
3 Результаты построения простой линейной регрессии.
4 Выводы.

Основные теоретические положения

Компоненты динамическою ряда и упрощенные приемы прогнозирования. Прогнозирование в экономике, как правило, связано с анализом временного ряда, который позволяет характеризовать закономерность изменения явления и экстраполировать ее на будущее (период прогноза).

При построении моделей по временным рядам необходимо учитывать компоненты динамического ряда. Уровни динамического ряда формируются под действием разных факторов. Одни из них являются основными на данном этапе исторического развития, а другие – случайными, несущественными с точ-

ки зрения содержания динамики. Фактическую величину уровня динамического ряда (Y_t) можно представить как функцию трех компонент:

$$Y_t = f(T, C, E), \tag{2.1}$$

где Y_t – фактический уровень динамического ряда в период времени *t*;

T – тренд ряда (тенденция);

C – периодические колебания (циклические, сезонные);

E – случайная компонента.

При анализе временного ряда прежде всего изучается *тенденция* ряда, определяющая основное направление развития явления за длительный период времени – *тренд* ряда (*T*); вместе с тем могут иметь место регулярные *периодические колебания* (циклические – длительностью в несколько лет, а также сезонные – внутригодичные), вызванные особенностями существования явления в одни периоды по сравнению с другими (*C*), что должно быть учтено при прогнозировании. Анализ будет не полным, если не исследовать *случайные колебания*, связанные с действием разного рода второстепенных факторов – случайная компонента (*E*).

Названные компоненты временного ряда необязательно присущи каждому временному ряду. Могут быть ряды динамики, в которых отсутствует как тенденция, так и периодические колебания. В этом случае уровни ряда являются функцией случайной компоненты: они колеблются вокруг среднего уровня, что характерно для так называемого *стационарного ряда.* Прогноз по стационарному ряду основан на предположении о неизменности в будущем среднего уровня динамического ряда и может быть представлен в виде

$$Y_P = \overline{Y} \pm S_P, \tag{2.2}$$

где Y_P – прогнозное значение;

$\overline{Y}$ – среднее значение уровня динамического ряда;

S_P – средняя ошибка прогноза, определяемая как

$$S_P = \sigma \cdot \sqrt{1 + \frac{1}{n}},$$

где σ – среднее квадратическое отклонение по временному ряду;

n – длина ряда.

Для прогноза принято считать предельную ошибку (Δp), вероятность которой обычно не должна превышать 5 %:

$$\Delta p = t_{\alpha=0{,}05,\, n-1} \cdot S_P,$$

где $t_{\alpha=0{,}05,\, n-1}$ – табличное значение *t*-критерия Стьюдента при уровне значимости $\alpha = 0{,}05$ и числе степеней свободы $n-1$.

Рассматривается динамика потребления сахара на душу населения за год по региону, кг (таблица 2.1).

Таблица 2.1 – Динамика потребления сахара на душу населения

2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
32	32	32	30	31	32	32	33	31	33	32	33

Требуется дать прогноз потребления сахара на душу населения в регионе на 2012 г.

В данном ряду динамики нет четко выраженной тенденции и периодических колебаний. Поэтому прогноз можно сделать исходя из среднего уровня временного ряда. По простой средней арифметической получим $\overline{Y} = 31{,}9167$; $\sigma = 0{,}862$. Табличное значение *t*-критерия Стьюдента при уровне значимости $\alpha = 0{,}05$ и числе степеней свободы $n - 1 = 11$ равно 2,201. Тогда предельная ошибка прогноза составит:

$$2{,}201 \cdot 0{,}862 \cdot \sqrt{1 + \frac{1}{12}} = 1{,}9747 .$$

В результате прогноз на 2012 г. окажется равным 31,9167 ± 1,9747, т. е. в интервале от 29,9 до 33,9 кг.

Однако большинство динамических рядов в экономике характеризуются тенденцией и случайными колебаниями. В этом случае прогноз можно дать с помощью обобщающих показателей динамики. Предполагая стабильным средний абсолютный прирост, прогноз можно представить в виде следующей экстраполяции:

$$Y_P = Y_n + \overline{\Delta} L ,$$

где Y_n – конечный уровень динамического ряда;

$\overline{\Delta}$ – средний абсолютный прирост уровня ряда в единицу времени;

L – период упреждения, т. е. на сколько временных интервалов дается экстраполяция.

Средний абсолютный прирост можно найти по формуле

$$\overline{\Delta} = (Y_n - Y_1) / (n - 1), \qquad (2.3)$$

где Y_1 – начальный уровень динамического ряда.

Если конечный уровень динамического ряда не характерен для исследуемого временного промежутка (резких колебаний в уровнях), то для прогноза используется более стабильный уровень (при этом дается его обоснование).

Практическое задание

1 Временной ряд приведен в таблице 2.1.

Используя средства MS Excel:

1) построить график временного ряда;

2) добавить линию тренда и ее уравнение;

3) найти уравнение тренда методом наименьших квадратов, сравнить уравнения (выше на графике и полученное);

4) построить график временного ряда и полученной функции тренда в одной системе координат;

5) определить параметры уравнения тренда с помощью функции Excel.

2 Имеются данные об изменении пропускной способности сети (байт/с) в зависимости от числа абонентов, полученные при замерах через каждые полчаса. Эти данные приведены в таблице 2.2.

Таблица 2.2 – Динамика изменения пропускной способности сети

t	1	2	3	4	5	6	7
x_{ti}	68	68	60	63	67	66	66
y_{ti}	11250000	10750000	10000000	10625000	11000000	10000000	9500000

3 Имеются данные об изменении пропускной способности сети (байт/с) и интенсивности абонентов (байт/с), полученные при замерах через каждые полчаса. Эти данные приведены в таблице 2.3

Таблица 2.3 – Динамика изменения пропускной способности сети и интенсивности абонентов

t	1	2	3	4	5	6	7
x_{ti}	11250000	10750000	10000000	10625000	11000000	10000000	9500000
y_{ti}	0,00002	0,000023	0,000025	0,000021	0,00002	0,000025	0,000023

Контрольные вопросы

1 Какие компоненты определяют фактическую величину уровня динамического ряда?

2 Как определяется средняя ошибка прогноза?

3 Что такое период упреждения прогноза?

3 Лабораторная работа № 9. Нелинейные модели для классификации и регрессии

Цель работы: ознакомиться с нелинейными регрессионными моделями и методами оценки параметров нелинейных моделей.

Порядок выполнения работы

1 Изучить основные теоретические положения, сделав необходимые выписки в конспект.
2 Получить задание у преподавателя, выполнить типовые задания.
3 Сделать выводы по результатам исследований.
4 Оформить отчет.

Требования к отчету

1 Цель работы.
2 Постановка задачи.
3 Результаты исследования.
4 Выводы.

Основные теоретические положения

Линейные регрессионные модели, рассмотренные в предыдущих темах, обладают тем свойством, что они *линейны по переменным* (переменные входят в модель в первой степени) *и линейны по параметрам* (параметры выступают в качестве коэффициентов при переменных). Однако не все зависимости можно выразить или достаточно адекватно приблизить линейными функциями. Многие зависимости не являются линейными по своей сути, поэтому использование для их изучения линейных моделей может привести к неадекватным результатам. Нелинейные зависимости часто встречаются в экономике. Так, при исследовании зависимости спроса от цены часто используют логарифмические модели, а при рассмотрении производственных функций – степенные.

Рассмотрим некоторые (наиболее часто используемые на практике) нелинейные модели, для которых возможно сведение к линейным. Для того чтобы свести нелинейную модель к линейной (*линеаризовать модель*), обычно с помощью некоторых преобразований переменных нелинейную модель представляют в виде линейного соотношения между преобразованными переменными, оценивают коэффициенты этого соотношения и затем, с помощью обратного преобразования, находят оценки параметров исходной нелинейной модели. Сразу заметим, что не всякая нелинейная модель может быть оценена подобным образом, в ряде случаев невозможно подобрать подходящее преобразование, линеаризующее модель. В этом случае приходится использовать методы нелинейной оптимизации.

Говоря о нелинейных моделях, часто выделяют модели, *нелинейные по переменным* (но *линейные относительно параметров*), и модели, *нелинейные по*

оцениваемым параметрам.

Оценка моделей, *нелинейных по объясняющим переменным*, *но линейных по оцениваемым параметрам*, не представляет особой сложности: в этом случае обычно используют *замену переменных* для сведения модели к линейной и оценки параметров с помощью обычного МНК (примененного к модели с замененными переменными).

Так, в случае ***полиномиальной зависимости*** степени k

$$y = a_0 + a_1 x + a_2 x^2 + ... + a_k x^k + \varepsilon$$

с помощью замены переменных

$$z_1 = x,\ z_2 = x^2,\ ...,\ z_k = x^k$$

получаем линейную модель множественной регрессии с k объясняющими переменными

$$y = a_0 + a_1 z_1 + a_2 z_2 + ... + a_k z_k + \varepsilon.$$

Оценки параметров этой линейной модели находят с помощью обычного МНК.

На практике среди подобных полиномиальных регрессий наиболее часто встречаются полиномы второй степени (квадратичная или параболическая регрессия)

$$y = a_0 + a_1 x + a_2 x^2 + \varepsilon$$

и третьей степени (кубическая регрессия)

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \varepsilon.$$

Обычно эта функция используется тогда, когда внутри рассматриваемого интервала изменения фактора прямой или обратный характер зависимости изменяется на противоположный. Полиномиальные функции хорошо подходят для моделирования эффекта масштаба, анализа максимумов и минимумов.

Модель вида

$$y = \alpha + \frac{\beta}{x} + \varepsilon$$

называется ***обратной (гиперболической) моделью***.

Эта модель сводится к линейной с помощью замены

$$z = \frac{1}{x}.$$

Данная модель обычно применяется в тех случаях, когда неограниченное увеличение объясняющей переменной x асимптотически приближает зависимую переменную y к некоторому пределу. Обратные функции хорошо подходят для моделирования эффектов полного насыщения и ограниченности. В зависимости от знаков коэффициентов можно выделить следующие характерные случаи (рисунок 3.1).

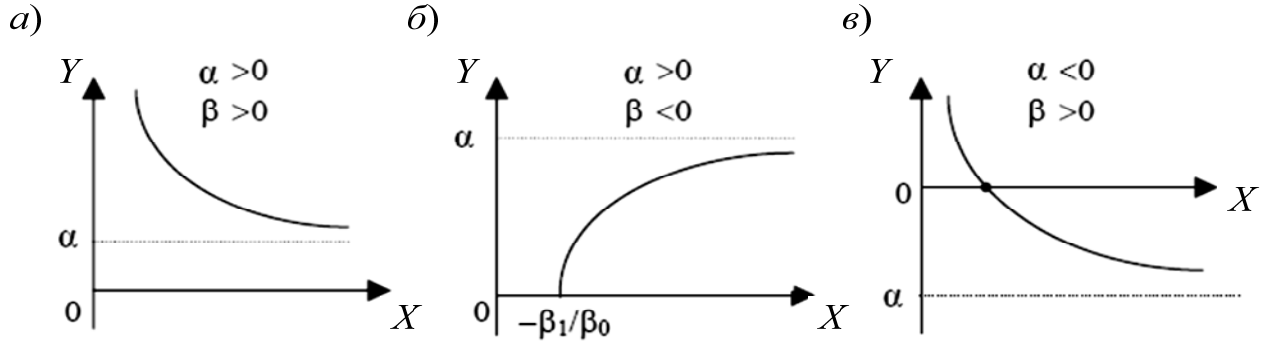


Рисунок 3.1 – Графики обратной модели при различных знаках коэффициентов

Зависимость, изображенная на рисунке 3.1, *а*, может отражать зависимость между объемом выпуска и средними фиксированными издержками. График, изображенный на рисунке 3.1, *б*, – зависимость между доходом и спросом на блага (например, на товары первой необходимости либо товары относительной роскоши) – так называемые *функции Торнквиста*. Примером такой зависимости такого вида могут служить также *кривые Энгеля*, отражающие взаимосвязь доли расходов на товары длительного пользования и общих сумм расходов (или доходов). Важным случаем графика, изображенного на рисунке 3.1, *в*, является *кривая Филлипса,* отражающая зависимость между процентным изменением заработной платы от уровня безработицы, выраженного в процентах.

Модель вида

$$y = \frac{1}{\alpha + \beta x + \varepsilon}$$

также является обратной моделью и может быть приведена к линейной модели: обращая обе части равенства, получаем линейную форму относительно переменной $1/y$

$$\frac{1}{y} = \alpha + \beta x + \varepsilon,$$

которая окончательно линеаризуется с помощью замены $y^{\%} = 1/y$.

Возможны и другие модели, нелинейные по объясняющим переменным, которые линеаризуются заменой переменных.

Например, ***линейно логарифмические*** **(*полулогарифмические*)** зависимости

$$y = \alpha + \beta \ln x + \varepsilon,$$

которые приводятся к линейной форме заменой $z = \ln x$.

Подобные зависимости также используются при моделировании *кривых Энгеля* и характеризуются тем, что логарифм при объясняющей переменной снижает влияние роста этой переменной (степень влияния x снижается с ростом x). Таким образом можно моделировать эффекты насыщения на уровне скорости роста: «возрастание с убывающей скоростью».

С помощью замены переменных возможна также оценка зависимостей ***с квадратными корнями***, например

$$y = \alpha + \beta\sqrt{x} + \varepsilon,$$

которые с помощью замены $z = \sqrt{x}$ приводятся к линейной модели.

Видим, что произвольная комбинация вышеприведенных зависимостей может быть линеаризована с помощью соответствующих замен переменных.

Несколько более сложным является случай *нелинейности модели по параметрам*, т. к. непосредственное применение МНК для их оценивания невозможно. Подходящим преобразованием (обычно связанным с логарифмированием по основанию e) иногда удается привести модель к линейному виду.

Так, в случае ***степенной*** зависимости

$$y = \alpha x^{\beta} \varepsilon,$$

прологарифмировав обе части, получим

$$\ln y = \ln \alpha + \beta \ln x + \ln \varepsilon.$$

Данная линейная модель, в которой и зависимая, и объясняющая переменные заданы в логарифмическом виде, иногда называется ***двойной логарифмической*** моделью. После замены переменных

$$y^{\%} = \ln y,\ x^{\%} = \ln x,\ \alpha^{\%} = \ln \alpha,\ \varepsilon^{\%} = \ln \varepsilon$$

получаем линейную модель

$$y^{\%} = \alpha^{\%} + \beta x^{\%} + \varepsilon^{\%}.$$

Степенная функция может использоваться при изучении зависимости спроса y на некоторое благо от его цены x (в данном случае $\beta < 0$) или от дохода x (в данном случае $\beta > 0$); при такой интерпретации переменных такая функция называется *функцией Энгеля*. Эта функция может также отражать зависимость объема выпуска от использования ресурса (производственная функция), в которой $0 < \beta < 1$, а также ряд других зависимостей. Примерные графи-

ки рассматриваемых степенных (логарифмических) зависимостей приведены на рисунке 3.2.

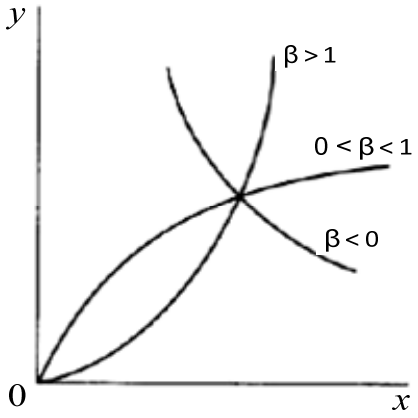


Рисунок 3.2 – Графики степенных зависимостей

Следует также заметить, что изокванты часто встречаются в экономических моделях (рисунок 3.3).

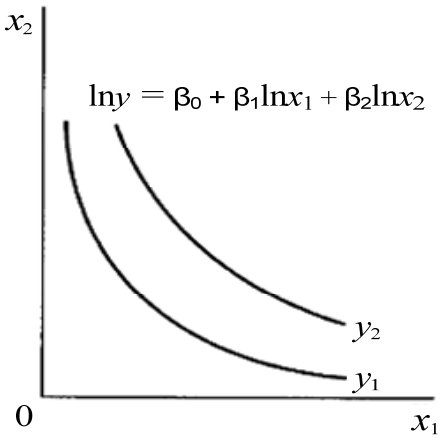


Рисунок 3.3 – Изокванты

В эконометрическом анализе широко применяется ***показательная*** модель вида

$$y = \alpha\beta^{x}\varepsilon .$$

Эта зависимость также приводится к линейной форме с помощью логарифмирования

$$\ln y = \ln\alpha + \ln\beta \cdot x + \ln\varepsilon$$

и замены переменных

$$y^{\%} = \ln y,\ \alpha^{\%} = \ln\alpha,\ \beta^{\%} = \ln\beta,\ \varepsilon^{\%} = \ln\varepsilon .$$

В результате получаем линейную модель

$$y^{\%} = \alpha^{\%} + \beta^{\%} x + \varepsilon^{\%}.$$

Аналогично проводится линеаризация ***показательной* (*экспоненциальной*)** зависимости вида

$$y = \alpha e^{\beta x} \varepsilon.$$

На практике эта зависимость часто встречается при анализе изменения зависимой переменной *y*, которая имеет постоянный темп прироста во времени. В этом случае часто пишут $y = \alpha e^{\beta t} \varepsilon$. Данная функция, путем логарифмирования и замены переменных, также сводится к линейной модели

$$y^{\%} = \alpha^{\%} + \beta x + \varepsilon^{\%},$$

где $y^{\%} = \ln y, \alpha^{\%} = \ln \alpha, \varepsilon^{\%} = \ln \varepsilon$.

Можно заметить, что последние две модели сводятся к ***логарифмически линейной* (*логлинейной*)** зависимости

$$\ln y = \alpha + \beta x + \varepsilon,$$

которая легко сводится к линейной заменой $y^{\%} = \ln y$.

Подобные зависимости используются при моделировании эффектов насыщения на уровне скорости роста – так называемое «возрастание с возрастающей скоростью». Примерами использования подобных зависимостей являются *кривые Энгеля* для товаров роскоши и товаров, спрос на которые проявляется при большом доходе, моделирование оплаты труда (процентная надбавка за стаж или опыт). Эти функции также хорошо подходят для моделирования эффектов, которые проявляются в процентном выражении в ответ на абсолютный рост факторов (вознаграждение).

Частным случаем логлинейной модели является зависимость, хорошо известная в банковском и финансовом деле,

$$y_t = y_0 (1 + r)^t,$$

где y_0 – начальная величина переменной *y* (например, первоначальный вклад в банке);

r – сложный темп прироста величины *y* (процентная ставка);

y_t – значение величины *y* в момент времени *t* (вклад в банке в момент времени *t*).

Логлинейная модель легко сводится к полулогарифмической модели: прологарифмировав, получаем

$$\ln y_t = \ln y_0 + t \ln(1 + r).$$

Обозначив $\ln y_0 = \alpha,\ \ln(1 + r) = \beta$, имеем логлинейную зависимость

$$\ln y_t = \alpha + \beta t.$$

График логлинейной зависимости представлен на рисунке 3.4.

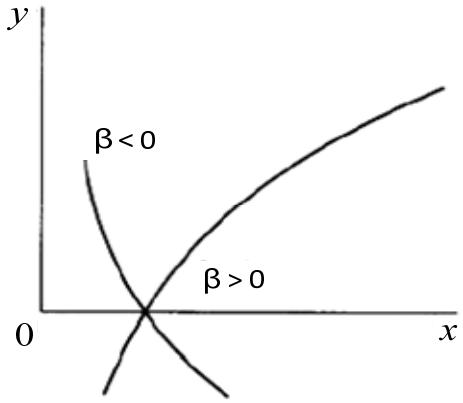


Рисунок 3.4 – График логлинейной зависимости

К линейному виду приводима и ***логистическая функция*** вида

$$y = \frac{\alpha}{1 + \beta \cdot e^{-\gamma x + \varepsilon}}.$$

Обращая обе части равенства, имеем

$$\beta \cdot e^{-\gamma x + \varepsilon} = \frac{\alpha}{y} - 1.$$

Прологарифмировав теперь обе части, получим линейную форму

$$\ln \beta - \gamma x + \varepsilon = \ln(\tfrac{\alpha}{y} - 1),$$

в которой осталось только сделать замены

$$y^{\%} = \ln(\tfrac{\alpha}{y} - 1),\ \beta^{\%} = \ln \beta.$$

Вышеприведенные модели легко обобщаются на случай нескольких переменных и композиции вышеперечисленных функций. Например, для оценки хорошо известной производственной ***функции Кобба–Дугласа*** (с учетом влияния случайных возмущений, присущих всякому экономическому явлению)

$$y = AK^{\alpha}L^{\beta}\varepsilon$$

после логарифмирования получаем линейную двойную логарифмическую форму

$$\ln y = \ln A + \alpha \ln K + \beta \ln L + \ln \varepsilon.$$

Для этой модели множественной линейной регрессии несложно найти оценки параметров с помощью классического МНК и получить оценки параметров исходной зависимости. Коэффициенты α и β представляют собой эластичности, их сумма указывает на эффект масштаба.

С учетом научно-технического прогресса производственная функция Кобба–Дугласа принимает вид

$$y = AK^{\alpha}L^{\beta}e^{rt}\varepsilon.$$

После логарифмирования также получаем линейную модель

$$\ln y = \ln A + \alpha \ln K + \beta \ln L + rt + \ln \varepsilon,$$

с помощью которой можно найти оценки параметров исходной зависимости.

Однако не всякая нелинейная зависимость может быть приведена к линейной с помощью замены переменных и/или их преобразования. Так, например, зависимости

$$y = \alpha + \beta x^{\gamma} + \varepsilon;$$

$$y = \alpha x_1^{\beta_1} x_2^{\beta_2} + \varepsilon$$

не могут быть преобразованы в линейную форму. В таких случаях используют специальные нелинейные итеративные процедуры оценивания параметров.

Практическое задание

Для решения задачи необходимо воспользоваться средствами ***Microsoft Excel.***

Необходимо построить производственную ***функцию Кобба–Дугласа***

$$Q = \alpha L^{\beta_1} K^{\beta_2},$$

где Q – объем выпуска продукции, ден. ед.;

L и K – трудозатраты и капиталовложения соответственно, ден. ед.;

α, β_1, β_2 – неизвестные коэффициенты, подлежащие определению по выборочным данным q_i, l_i, k_i (i = 1, 2, ..., n).

Логарифмируя функцию Кобба–Дугласа, приходим к задаче построения регрессионной зависимости

$$\ln Q = \beta_0 + \beta_1 \ln L + \beta_2 \ln K \mid \beta_0 = \ln \alpha.$$

Эта задача легко решается в пакете Microsoft Excel. После определения коэффициентов регрессии функцию Кобба–Дугласа можно считать построенной ($\alpha = e^{\beta_0}$).

Так как построенная модель не является линейной, то качество (адекватность) модели можно проверить с помощью средней относительной ошибки аппроксимации

$$E = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{q_i - \hat{Q}_i}{q_i} \right| \cdot 100\ \% \mid \hat{Q}_i = \alpha (l_i)^{\beta_1} (k_i)^{\beta_2}.$$

Если E < 10 %, то построенную функцию можно считать адекватной исходным данным и применять ее для прогнозирования.

Задание состоит в следующем.

1 На основе входных данных, соответствующих варианту, найти коэффициенты α, β_1, β_2 и составить функцию Кобба–Дугласа.

2 Оценить адекватность модели с помощью средней относительной ошибки аппроксимации.

Варианты заданий представлены в таблице 3.1.

Таблица 3.1 – Варианты заданий

Вариант 1				**Вариант 2**			
Номер точки	Q	L	K	Номер точки	Q	L	K
1	2350,2	2034,2	1870,4	1	2866,6	2440,8	2244
2	2470,1	2125,2	2150,1	2	3013	2550	2580
3	2110,2	1931	1450,4	3	2573,8	2316	1740
4	2560,5	2165	2242,5	4	3122,8	2595,6	2688
5	2650,4	2266,2	2751,6	5	3232,6	2718	3300
6	2240,3	1979,8	1640,9	6	2732,4	2373,6	1968
7	2430,6	2084,8	2003,6	7	2964,2	2496	2400
8	2530,8	2142,6	2166,4	8	3086,2	2564,4	2592
9	2550,7	2148,8	2184,9	9	3110,6	2575,2	2616
10	2450,4	2103,8	2091,6	10	2988,6	2523,6	2508
11	2290,2	2001,2	1780,4	11	2793,4	2401,2	2136
12	2160,1	1953,3	1540,1	12	2634,8	2343,6	1848
13	2400,3	2067,3	1960,9	13	2927,6	2480,4	2352
14	2490	2130,2	2150	14	3037,4	2556	2580
15	2590,4	2170,5	2301,6	15	3159,4	2604	2760

Продолжение таблицы 3.1

Вариант 3				**Вариант 4**			
Номер точки	*Q*	*L*	*K*	Номер точки	*Q*	*L*	*K*
1	2997	2542,5	2337,5	1	2478,3	2135,7	1963,5
2	3150,1	2656,2	2687,5	2	2604,9	2231,2	2257,5
3	2691	2412,5	1812,5	3	2225,2	2026,5	1522,5
4	3264,9	2703,7	2800	4	2699,8	2271,1	2352
5	3379,6	2831,2	3437,5	5	2794,7	2378,2	2887,5
6	2856,8	2472,5	2050	6	2362,3	2076,9	1722
7	3099,1	2600	2500	7	2562,7	2184	2100
8	3226,6	2671,2	2700	8	2668,1	2243,8	2268
9	3252,1	2682,5	2725	9	2689,2	2253,3	2289
10	3124,6	2628,7	2612,5	10	2583,8	2208,1	2194,5
11	2920,5	2501,2	2225	11	2415	2101	1869
12	2754,7	2441,2	1925	12	2277,9	2050,6	1617
13	3060,8	2583,7	2450	13	2531	2170,3	2058
14	3175,6	2662,5	2687,5	14	2626	2236,5	2257,5
15	3303,1	2712,5	2875	15	2731,4	2278,5	2415
Вариант 5				**Вариант 6**			
Номер точки	*Q*	*L*	*K*	Номер точки	*Q*	*L*	*K*
1	2607,2	2237,4	2057	1	2684,8	2298,4	2113,1
2	2740,4	2337,5	2365	2	2821,9	2401,2	2429,5
3	2340,9	2123	1595	3	2410,6	2180,9	1638,5
4	2840,2	2379,3	2464	4	2924,7	2444,1	2531,2
5	2940,1	2491,5	3025	5	3027,6	2559,4	3107,5
6	2485,2	2175,8	1804	6	2559,1	2235,1	1853,2
7	2696	2288	2200	7	2776,2	2350,4	2260
8	2806,9	2350,7	2376	8	2890,5	2414,8	2440,8
9	2829,1	2360,6	2398	9	2913,3	2424,9	2463,4
10	2718,2	2313,3	2299	10	2799,1	2376,3	2361,7
11	2540,7	2201,1	1958	11	2616,3	2261,1	2011,4
12	2396,4	2148,3	1694	12	2467,7	2206,8	1740,2
13	2662,7	2273,7	2156	13	2741,9	2335,7	2214,8
14	2762,5	2343	2365	14	2844,8	2406,9	2429,5
15	2873,5	2387	2530	15	2959	2452,1	2599

Окончание таблицы 3.1

Вариант 7				**Вариант 8**			
Номер точки	*Q*	*L*	*K*	Номер точки	*Q*	*L*	*K*
1	2788,6	2379,7	2187,9	1	2840,6	2420,4	2225,3
2	2931	2486,2	2515,5	2	2985,6	2528,7	2558,5
3	2503,8	2258,1	1696,5	3	2550,5	2296,7	1725,5
4	3037,8	2530,7	2620,8	4	3094,4	2573,9	2665,6
5	3144,6	2650	3217,5	5	3203,2	2695,3	3272,5
6	2658	2314,2	1918,8	6	2707,6	2353,8	1951,6
7	2883,5	2433,6	2340	7	2937,3	2475,2	2380
8	3002,2	2500,2	2527,2	8	3058,2	2543	2570,4
9	3025,9	2510,8	2550,6	9	3082,3	2553,7	2594,2
10	2907,2	2460,5	2445,3	10	2961,5	2502,5	2487,1
11	2717,4	2341,1	2082,6	11	2768	2381,1	2118,2
12	2563,1	2285	1801,8	12	2610,9	2324	1832,6
13	2847,9	2418,3	2293,2	13	2901	2459,7	2332,4
14	2954,7	2492,1	2515,5	14	3009,8	2534,7	2558,5
15	3073,4	2538,9	2691	15	3130,7	2582,3	2737

Вариант 9				**Вариант 10**			
Номер точки	*Q*	*L*	*K*	Номер точки	*Q*	*L*	*K*
1	2918,7	2481,4	2281,4	1	3049,3	2583,1	2374,9
2	3067,8	2592,5	2623	2	3205,1	2698,7	2730,5
3	2620,6	2354,6	1769	3	2737,9	2451,1	1841,5
4	3179,5	2638,8	2732,8	4	3321,8	2747	2844,8
5	3291,3	2763,3	3355	5	3438,6	2876,5	3492,5
6	2782,1	2413,1	2000,8	6	2906,6	2512	2082,8
7	3018,1	2537,6	2440	7	3153,2	2641,6	2540
8	3142,3	2607,1	2635,2	8	3282,9	2713,9	2743,2
9	3167,1	2618,1	2659,6	9	3308,9	2725,4	2768,6
10	3042,9	2565,6	2549,8	10	3179,1	2670,8	2654,3
11	2844,2	2441,2	2171,6	11	2971,5	2541,2	2260,6
12	2682,7	2382,6	1878,8	12	2802,8	2480,3	1955,8
13	2980,8	2521,7	2391,2	13	3114,2	2625	2489,2
14	3092,6	2598,6	2623	14	3231	2705,1	2730,5
15	3216,8	2647,4	2806	15	3360,8	2755,9	2921

Контрольные вопросы

1 В чём отличие нелинейных моделей от линейных?

2 Назовите модели, нелинейные по переменным (но линейные относительно параметров).

3 Каким образом производится замена переменных в случае полиномиальной зависимости?

4 В каких случаях применяется обратная (гиперболическая) модель?

5 Какие зависимости используются при моделировании кривых Энгеля?

6 Назовите модели, нелинейные по оцениваемым параметрам.

7 Какую функцию описывает двойная логарифмическая модель?

8 В каких случаях применяется показательная (экспоненциальная) модель?

9 Приведите примеры нелинейных зависимостей, которые не могут быть приведены к линейным с помощью замены переменных и/или их преобразования.

4 Лабораторная работа № 10. Вероятностный вывод для дискретных моделей

Цель работы: ознакомиться с концепцией вероятностного программирования при помощи библиотеки pymc3 языка Python.

Порядок выполнения работы

1 Изучить основные теоретические положения, сделав необходимые выписки в конспект.

2 Получить задание у преподавателя, выполнить типовые задания.

3 Сделать выводы по результатам исследований.

4 Оформить отчет.

Требования к отчету

1 Цель работы.

2 Постановка задачи.

3 Результаты исследования.

4 Выводы.

Основные теоретические положения

Для выполнения лабораторной работы потребуется библиотека ***pymc3*** языка программирования Python.

```
!pip3 install pymc3 # установить библиотеку
```

Введение в байесовский статистический вывод. Байесовская статистика начинает построение своей модели при помощи понятия априорной вероятности, с помощью которой описывается текущее состояние знаний, относительно параметров распределения. ***Априорная вероятность***, таким образом, – это степень уверенности в том, что исследуемый параметр примет то или иное значение ещё до начала сбора исходных статистических данных. На этом основании байесовское понимание вероятности относят к группе субъективистских трактовок вероятности. Чаще всего предполагается, что для оценки степени уверенности необходимо привлечь экспертов, чьё субъективное свидетельство позволит избежать действительной многократной реализации интересующего эксперимента.

Следующий элемент – это исходные статистические данные. По мере их поступления статистик пересчитывает распределение вероятностей анализируемого параметра, переходя от априорного распределения к апостериорному, используя для этого формулу Байеса

$$P(\theta \mid y) = \frac{P(y \mid \theta)P(\theta)}{P(y)},$$

где $P(\theta)$ – априорная вероятность гипотезы θ, выражает знания о значениях параметров, предшествующие построению модели;

$P(y \mid \theta)$ – апостериорная вероятность, представляет собой распределение вероятностей для параметра в модели, рассчитанное с учётом априорных знаний и правдоподобия новых;

$P(\theta \mid y)$ – правдоподобие, вероятность наблюдать данные y при истинности гипотезы θ;

$P(y)$ – свидетельство (evidence, marginal likelihood), полная вероятность гипотезы θ. Другими словами – вероятность наблюдения данных, усреднённая по всем возможным значениям, которые могут принимать параметры.

Суть формулы в том, что она позволяет переставить причину и следствие: по известному факту события вычислить вероятность того, что оно было вызвано данной причиной. Эту формулу также называют формулой обратной вероятности. Процесс пересмотра вероятностей, связанных с высказываниями, по мере поступления новой информации составляет существо машинного обучения и является одним из возможных способов формализации и операционализации следующего тезиса: «степень нашей разумной уверенности в некотором утверждении (касающемся, например, неизвестного численного значения интересующего нас параметра) возрастает и корректируется по мере пополнения имеющейся у нас информации относительно исследуемого явления» (Айвазян, 2008). В частотном подходе данный тезис интерпретируется в свойстве состоятельности оценки неизвестного параметра: чем больше объём выборки, на основании которой строим оценку, тем большей информацией об этом параметре располагаем и тем ближе к истине заключение. Специфика байесовского подхода к интерпретации этого тезиса основана на том, что вероятность, понимаемая как количественное значение степени разумной уверенности в справедливости некоторого

утверждения, пересматривается по мере изменения информации, касающейся этого утверждения. Поэтому в данном подходе вероятность всегда есть условная вероятность при условии нынешнего состояния информации.

У байесовского метода имеется несколько недостатков. Одним из них является необходимость привлекать для расчёта априорные данные, которые могут быть недоступны. А если они и доступны, то, как отмечалось ранее, часто носят субъективный характер. Другой недостаток – сложность вычислений. В вышеописанном примере для вычисления байесовской вероятности необходимо было вычислить частотную вероятность, полную вероятность и, наконец, собственно байесовскую вероятность. Сложность байесовских вычислений частично объясняет тот факт, что байесовские методы вновь обрели популярность с развитием вычислительной техники. Следующий недостаток байесовского метода – неинтуитивность, непонятность его результатов для обыденного сознания. Именно на этой неинтуитивности построен знаменитый парадокс Монти Холла, который легко решает с помощью формулы Байеса.

Развивая статистическую модель данных, байесовский подход предоставляет дополнительные вероятностные модели для всех неизвестных параметров в модели данных. Этот подход заключается в моделировании неуверенности в параметрах с помощью научной информации эксперта. Эта информация называется априорной. Экспертная информация должна быть получена независимо от анализируемых данных.

Байесовская статистика концептуально очень проста: есть некоторые данные, которые являются фиксированными в том смысле, что не можем изменить то, что измерили, и есть параметры, значения которых представляют интерес, и, следовательно, исследуем их вероятные значения.

Все неопределенности, которые имеем, моделируются с использованием вероятностей. В других статистических парадигмах существуют разные типы неизвестных величин; в байесовских рамках все, что неизвестно, рассматривается одинаково. Если не знаем количество, назначаем ему распределение вероятностей. Затем теорема Байеса используется для преобразования предшествующего распределения вероятности $p(\theta)$ (что знаем о данной проблеме до наблюдения данных) в апостериорное распределение $p(\theta \mid D)$ (то, что знаем после наблюдения данных). Другими словами, байесовская статистика является формой обучения.

Вероятностное программирование. Возможность автоматизации части логического вывода привела к разработке вероятностных языков программирования (PPL), которые позволяют четко разделить создание модели и логический вывод. Используя PPL, пользователи могут задать вероятностную модель, написав несколько строк кода, после чего вывод делается автоматически.

Бросание монетки – частотный подход. В большинстве учебников по теории вероятностей изучение начинается с бросания монетки. То же самое и в байесовской статистике. Пусть кто-то бросает монету, и наблюдаем исходы: 1 – выпал орёл, 0 – выпала решка. Задача – оценить вероятность выпадения орла (т. е. θ в данном случае – это количество выпадений орла).

Эта случайная величина – результаты бросания монетки – подчиняется распределению Бернулли. Также знаем, что последовательность независимых случайных величин $Y = X_1 + X_2 + \cdots + X_n$, имеющих одинаковое распределение Бернулли, имеет биномиальное распределение с параметрами n и p, где p – вероятность успеха в n испытаниях. Это записывается в виде $Y \sim \mathrm{Bin}(n, p)$.

Функция вероятности данной случайной величины задаётся формулой Бернулли

$$P_n^k = C_n^k p^k (1-p)^{n-k},$$

где k – количество успешных исходов.

Программирование формулы на Python.

```
%matplotlib inline
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import pymc3 as pm
from scipy import stats
from math import factorial
import arviz as az
import matplotlib.pyplot as plt
#%config InlineBackend.figure_format = 'svg'
plt.rcParams['figure.figsize'] = (10, 6)
SEED = 12


def num_of_successes(n, k):
    return factorial(n)/(factorial(k) * factorial(n - k))


def probability_of_success(p, n, k):
    C_kn = num_of_successes(n, k)
    return C_kn * (p**k) * (1 - p)**(n - k)
```

Формула Бернулли позволяет ответить на вопрос, какова вероятность в десяти бросках монеты получить девять «орлов», если монетка честная (вероятность «орла» составляет 50 %).

```
probability_of_success(p=0.5, n=10, k=9)
```

```
Code output
0.009765625
```

Как видно, вероятность достаточно небольшая.

Вероятность успешных испытаний для бросков монетки моделируется при помощи ***биноминального распределения***. Это дискретное распределение показывает вероятность успеха в серии из N испытаний при фиксированном значении θ, если все испытания независимы друг от друга. В терминах байесовской статистики это будет правдоподобие $p(y \mid \theta)$. Следующий код генерирует 9 биномиальных распределений; у каждого есть своя легенда, где показаны параметры: n – количество испытаний и p – вероятность успеха в каждом испытании.

```
    n_params = [1, 4, 12]
    p_params = [0.25, 0.5, 0.75]

    f, ax = plt.subplots(len(n_params), len(p_params), sharex=True,
sharey=True)
    for i in range(3):
        for j in range(3):
            n = n_params[i]
            p = p_params[j]
            y = [probability_of_success(p=p, n=n, k=i) for i in
range(n+1)]
            ax[i,j].vlines(range(0, n + 1), 0, y, colors='b', lw=5)
            ax[i,j].set_ylim(0, 1)
            ax[i,j].plot(0, 0, label="n = {:3.2f}\np ={:3.2f}".format(n,
p), alpha=0)
            ax[i,j].legend(fontsize=10)
    ax[2,1].set_xlabel('$\\theta$', fontsize=14)
    ax[1,0].set_ylabel('$p(y|\\theta)$', fontsize=14);
```

Результаты работы программы представлены на рисунке 4.1.

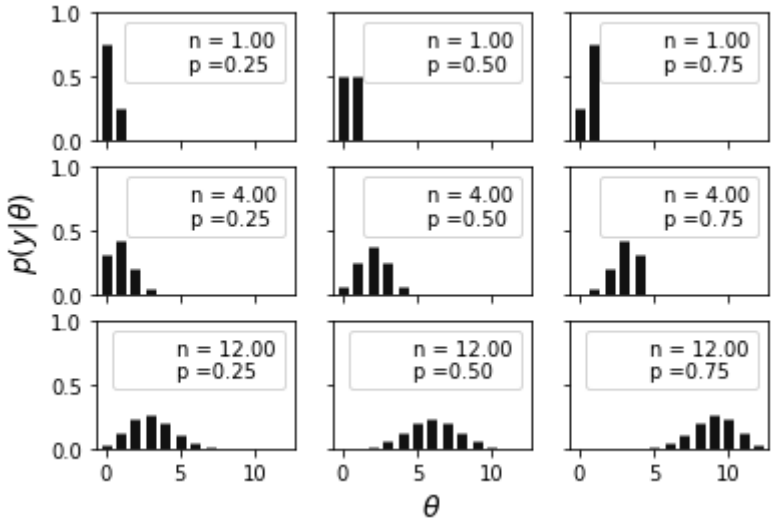


Рисунок 4.1 – Оценка правдоподобия в терминах байесовской статистики

На практике не известно, чему равен θ. Однако это не проблема для байесовского подхода, где каждый раз, когда отсутствует значение параметра, необ-

ходимо придумать для него априорное распределение, а затем подобрать его параметры.

В данном примере в роли априорного распределения будет выступать ***бета-распределение***. Оно выражает уверенность в том, что θ принимает именно такое значение, какое декларируется. Выбрано бета-распределение, поскольку оно является сопряжённым к биномиальному распределению, т. е. может принимать тот же вид, но при других параметрах. Также если функцию правдоподобия для распределения Бернулли умножить на плотность бета-распределения, то снова получают бета-распределение. Таким образом, переходя от априорного распределения к апостериорному, вид распределения не меняется, что удобно.

Пример программы:

```
n_trials = [0, 1, 2, 3, 4, 8, 16, 32, 50, 150]
data = [0, 1, 1, 1, 1, 4, 6, 9, 13, 48]
theta_real = 0.35

beta_params = [(1, 1), (20, 20), (1, 4)]
x = np.linspace(0, 1, 200)

for idx, N in enumerate(n_trials):
    if idx == 0:
        plt.subplot(4, 3, 2)
    else:
        plt.subplot(4, 3, idx+3)
    y = data[idx]
    for (a_prior, b_prior) in beta_params:
        # это получилось после перемножения биноминального на бета
       p_theta_given_y = stats.beta.pdf(x, a_prior + y, b_prior + N - y)
        plt.fill_between(x, 0, p_theta_given_y, alpha=0.7)
    plt.xlabel('θ')
    plt.axvline(theta_real, ymax=0.3, color='k')
    plt.plot(0, 0, label=f'{N:4d} trials\n{y:4d} heads', alpha=0)
    plt.legend()
    plt.tight_layout()
```

Результат представлен на рисунке 4.2.

Графически изображено (см. рисунок 4.2) распределение вероятностей для значений θ после учёта имеющихся данных, принимая во внимание априорную информацию. Эти распределения удалось легко вычислить, потому что биноминальное и бета-распределения сопряжённые. Предположим, однако, что априорное распределение не является сопряжённым и трудно решить задачу, так сказать, вручную. На практике обычно бывает именно так.

Линейная регрессия в байесовском стиле. Обычно линейная регрессия выглядит так:

$$Y = X\beta + \varepsilon.$$

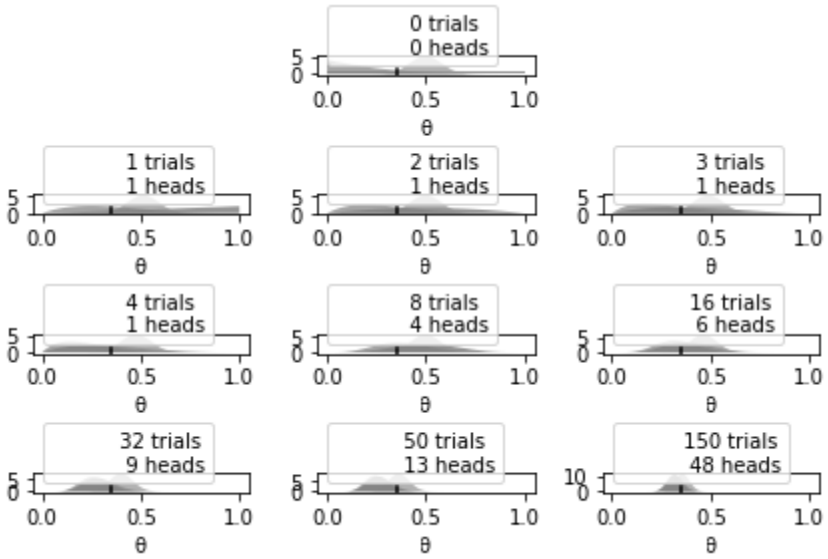


Рисунок 4.2 – Распределение вероятностей для значений θ

В байесовской статистике выражают эту модель в терминах распределения вероятностей. Вышеупомянутая линейная регрессия может быть переписана как

$$Y \sim N(X\beta, \sigma^2).$$

Таким образом, рассматриваем Y как случайную переменную (или случайный вектор), элементы которой распределены нормально. Среднее значение этого нормального распределения обеспечивается линейным предиктором с дисперсией σ^2.

Хотя это, по сути, одна и та же модель. Есть два критических преимущества байесовского подхода:

1) наличие априорного распределения: можем количественно оценить любые предварительные знания, которые могут быть, выбрав соответствующие априорные распределения. Например, если думаем, что дисперсия, вероятно, будет небольшой, выбрали бы априоное распределение, где вероятность небольших значений выше;

2) количественная оценка неопределенности: не получаем единственную оценку β, как указано ранее, но вместо этого получаем полное апостериорное распределение о том, насколько вероятны различные значения. Например, с небольшим количеством данных неопределенность для β будет очень высокой, и получим очень широкие постериорные распределения для этого параметра.

Сгенерируем 200 точек, которые описываются уравнением $y = 2x + 1 + \varepsilon$, и нанесём их и прямую на график (рисунок 4.3).

```
size = 200
true_intercept = 1
```

```
    true_slope = 2

    x = np.linspace(0, 1, size)
    # y = a + b*x
    true_regression_line = true_intercept + true_slope * x
    # add noise
    y = true_regression_line + np.random.normal(scale=.5, size=size)

    data = dict(x=x, y=y)
    fig = plt.figure(figsize=(7, 7))
    ax = fig.add_subplot(111, xlabel='x', ylabel='y', title='Generated
data and underlying model')
    ax.plot(x, y, 'x', label='sampled data')
    ax.plot(x, true_regression_line, label='true regression line',
lw=2.)
    plt.legend(loc=0);
```

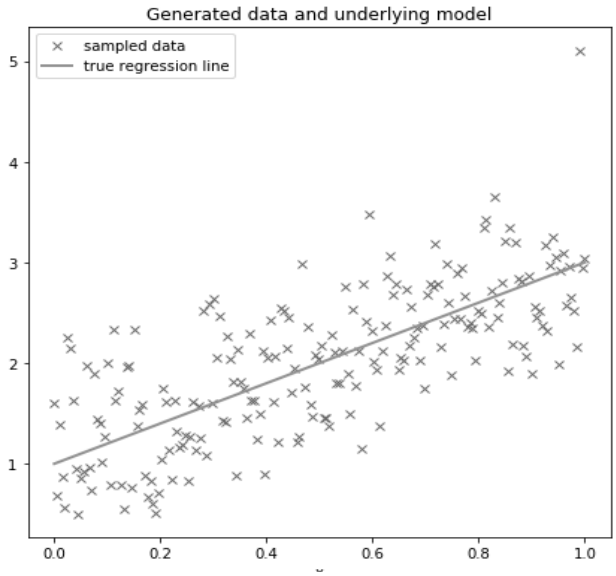
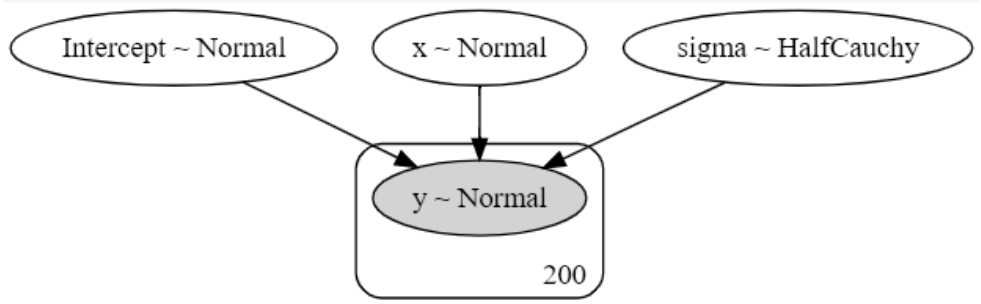


Рисунок 4.3 – Сгенерированные данные по заданному закону

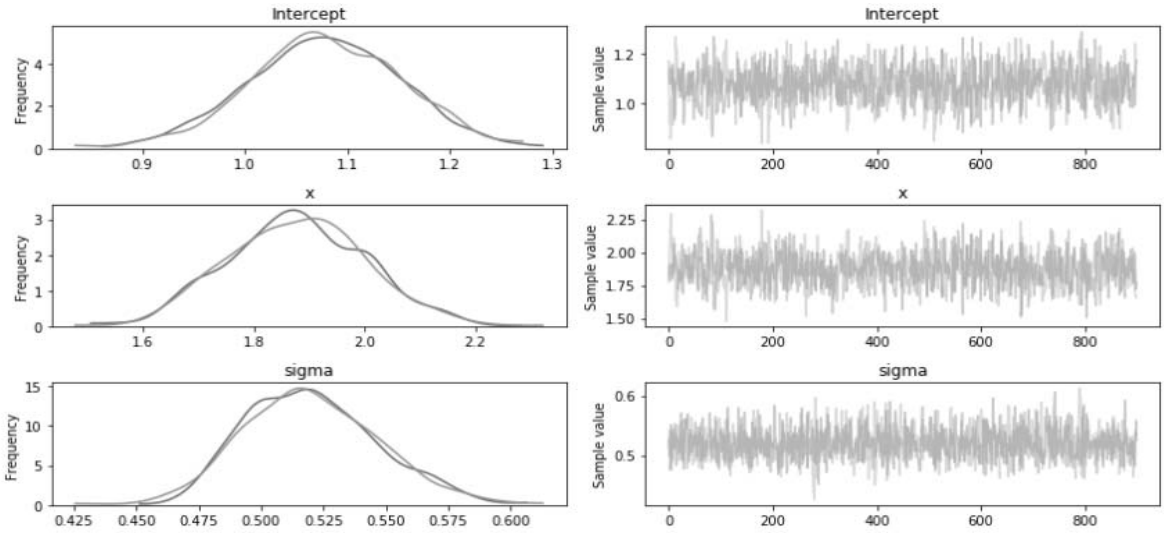
```
    with pm.Model() as simple_linear_model:
        # Define priors
        sigma = pm.HalfCauchy('sigma', beta=10, testval=1)
        intercept = pm.Normal('Intercept', 0, sd=20)
        x_coeff = pm.Normal('x', 0, sd=20)

        # Define likelihood
        likelihood = pm.Normal('y',
                               mu=intercept + x_coeff * x,
                               sd=sigma,
                               observed=y)

        # Inference!
        trace = pm.sample(1000) # draw 3000 posterior samples using NUTS
sampling
```

```
Code output
Auto-assigning NUTS sampler...
Initializing NUTS using jitter+adapt_diag...
Multiprocess sampling (2 chains in 2 jobs)
NUTS: [x, Intercept, sigma]
Sampling 2 chains: 100%|██████████| 3000/3000 [00:07<00:00, 400.94draws/s]
```

```
pm.model_to_graphviz(simple_linear_model)
```



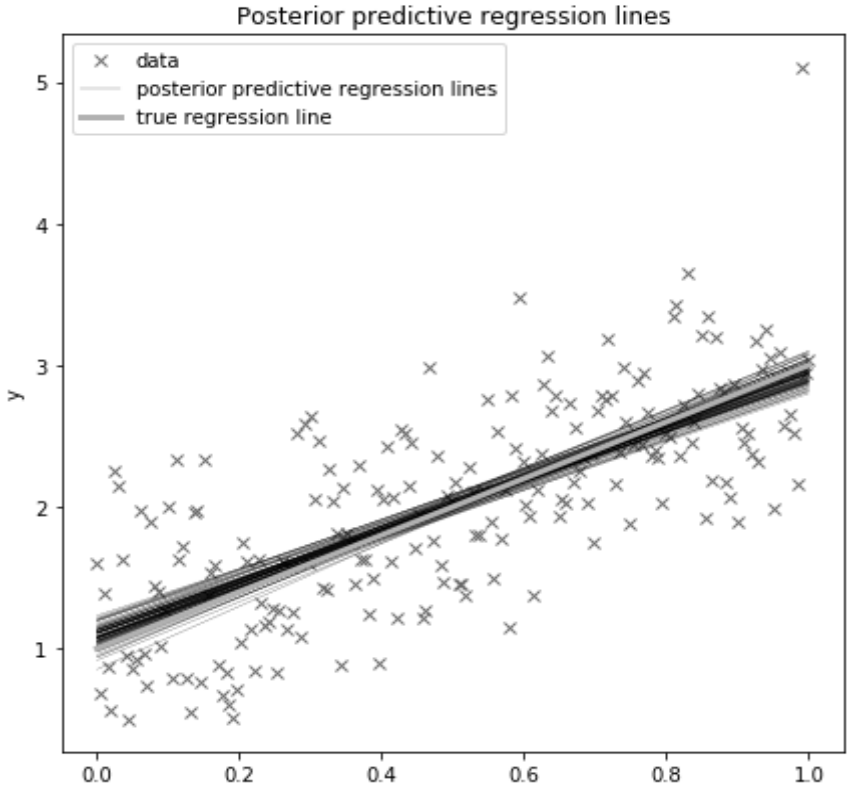

```
pm.traceplot(trace, skip_first=100);
```



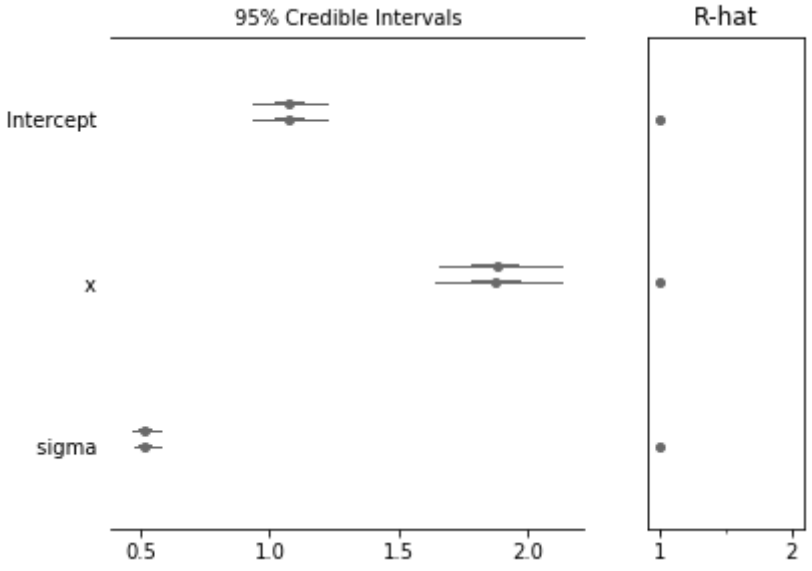
Таким образом, есть не только одна наиболее подходящая линия регрессии, но и множество. Апостериорный прогностический график берет несколько выборок из апостериорного (пересечения и наклоны) и строит линию регрессии для каждого из них. Здесь используем для этого вспомогательную функцию *plot_posterior_predictive_glm ().*

```
plt.figure(figsize=(7, 7))
plt.plot(x, y, 'x', label='data')
pm.plot_posterior_predictive_glm(trace,
                                 samples=100,
                                 label='posterior predictive regres-
sion lines')
plt.plot(x, true_regression_line, label='true regression line',
lw=3., c='y')

plt.title('Posterior predictive regression lines')
```

```
plt.legend(loc=0)
plt.xlabel('x')
plt.ylabel('y');
```




```
pm.forestplot(trace);
```



Практическое задание

Для получения практических навыков требуется выполнить задания по программированию, представленные в теоретическом разделе.

Контрольные вопросы

1 Что такое априорная вероятность ?
2 Что такое апостериорная вероятность ?
3 Что такое правдоподобие в рамках байесовской статистики ?
4 Что такое свидетельство в рамках байесовской статистики ?
5 Опишите формулу Бейеса и её составляющие.
6 Перечислите недостатки байесовского метода.
7 Что такое биноминальное распределение ?
8 Опишите преимущества байесовского подхода в рассмотрении линейной регрессии.

5 Лабораторная работа № 11. Обучение с неполными данными

Цель работы: изучить методы обработки неполных экспериментальных данных.

Порядок выполнения работы

1 Изучить основные теоретические положения, сделав необходимые выписки в конспект.
2 Получить задание у преподавателя, выполнить типовые задания.
3 Сделать выводы по результатам исследований.
4 Оформить отчет.

Требования к отчету

1 Цель работы.
2 Постановка задачи.
3 Результаты исследования.
4 Выводы.

Основные теоретические положения

В настоящее время гибридные системы искусственного интеллекта стали активно применяться при решении широкого круга задач классификации и прогнозирования. В основе одной из реализаций систем данного класса лежит нечеткая экспертная система, база знаний которой генерируется в процессе структурной и параметрической идентификации на основе имеющихся в наличии статистических данных. Идентификация системы осуществляется в процессе проведения кластеризации доступных в виде баз данных статистических данных с последующим отображением структуры кластеров в структуру нечеткой

базы знаний. Однако в большинстве случаев имеющиеся базы данных имеют значительное количество пропусков в таблицах. Объективными причинами этого являются поломки оборудования при измерении тех или иных характеристик, потеря ретроспективной информации, ограничение доступа к информации и др. Субъективные причины обусловлены человеческим фактором при накоплении и обработке информации. Таким образом, необходимым условием построения гибридных систем данного класса является привлечение того или иного способа обработки или предварительного восстановления пропусков в статистических данных при проведении их кластеризации.

В настоящее время разработано множество методов восстановления пропусков. Наиболее распространенные методы данного класса с указанием их основных особенностей представлены далее.

1 *Исключение строк с наличием пропусков.* Данный метод легко реализуем, но необходимым условием его применения является следование данных требованию MCAR (missing completely at random), т. е. пропуски в данных по переменным должны быть полностью случайными. Кроме того, он обычно применяется лишь при незначительном количестве пропусков в таблице, иначе полученная на выходе таблица данных становится непредставительной. Главный недостаток такого подхода обусловлен потерей информации при исключении неполных данных.

2 *Заполнение пропусков средними по столбцу значениями.* Данный метод также легко реализуем, но его применение имеет смысл только в случае, когда пропуски в данных по переменным являются случайными и сам механизм пропусков несущественен. К недостаткам метода относят вносимые искажения в распределения данных, уменьшение дисперсии.

3 *Метод ближайших соседей.* В основе метода лежит механизм поиска строк таблицы, которые по определенному критерию являются ближайшими к строке с *пропусками*. Для заполнения пропуска значения данной переменной (в фиксированном столбце) у соседних строк усредняются с определенными весовыми коэффициентами, обратно пропорциональными расстоянию к строке с пропуском. При большом количестве пропусков данный метод также практически неприменим, поскольку базируется на существовании связей между строками в таблице.

4 *Регрессионный анализ.* Из условий применения этого метода можно выделить требование о следовании данных условию MAR (хотя для частных случаев возможно применение более слабых требований) и требования, относящиеся к выполнению предпосылок регрессионного анализа. Недостатки метода очевидны: качество восстановления пропусков напрямую зависит от успешного выбора взятой за основу регрессионной модели.

5 *Метод сплайн-интерполяции.* Для успешного применения необходимо, чтобы данные следовали условию MAR. Недостатки метода следуют из самой его идеи. Например, в случае восстановления группы пропусков, следующих подряд друг за другом, результат аппроксимации сплайном данной группы не всегда может дать оценки, приближающиеся с достаточной точностью к значениям, которые могли бы быть на месте пропусков.

6 *Метод максимальной правдоподобности и EM-алгоритм.* Метод требует проверки гипотез о распределении значений переменных. Применение осложняется при большом количестве пропущенных значений переменной. Особенность данного метода состоит в построении модели порождения пропусков с последующим получением выводов на основании функции правдоподобия, построенной при условии справедливости данной модели, с оцениванием параметров методами типа максимального правдоподобия. Отметим, что для данных методов возможно построение моделей, учитывающих конкретную специфику области, и, как следствие, постановка более слабых условий к данным (слабее MAR).

7 *Алгоритмы ZET и ZetBraid.* По сути, алгоритм ZET является детально проработанной и апробированной технологией верификации экспериментальных данных, основанной на гипотезе их избыточности. Главная идея алгоритма ZET заключается в подборе «компетентной матрицы». Используя данные из нее, находят параметры зависимости, которая применяется для прогнозирования пропущенного значения. Субъективизм определения размерности «компетентной матрицы» приводит к учету неинформативных и шумовых факторов и смещению оценки неизвестного значения. Основное отличие алгоритма ZetBraid состоит в определении оптимального размера «компетентной матрицы». Данные алгоритмы хорошо показали себя, но статистическая оценка неизвестного значения исключительно на основе корреляционно-регрессионного анализа и необходимость задания ряда важных параметров приводит к необходимости убедиться в правдоподобности восстановленных значений.

8 *Resampling метод.* Метод является итеративным и имеет две модификации, которые основаны на построении регрессионных моделей с последующим усреднением полученных оценок для пропущенных значений. Преимуществом метода является повторное использование исходных данных, ведь увеличение числа подвыборок позволяет наиболее полно использовать исходную информацию. С другой стороны, объем новой информации уменьшается для каждой новой подвыборки, т. к. увеличивается вероятность того, что данные элементы выборки были уже выбраны раньше, – это основной недостаток метода вкупе с отсутствием процедур его оптимизации.

9 *Метод кластерного анализа.* Особенность метода – его применение не опирается на какую-либо вероятностную модель, но при этом оценить его свойства в статистических терминах не представляется возможным. Однако этот метод обладает существенным достоинством в виде алгоритмической простоты его реализации, а также позволяет указать предпочтительный порядок восстановления данных и выявить случаи, когда пропуски не могут быть восстановлены по имеющимся данным.

Обобщая результаты обзора наиболее распространенных методов восстановления пропусков в статистических данных, следует отметить, что оценки для пропущенных значений, как правило, вычисляются по присутствующим данным, что вносит искусственную зависимость между наблюдениями. Кроме того, распределение данных после заполнения будет отличаться от истинного, даже если пренебречь зависимостью, указанной ранее. Этот факт особенно

нагляден для простых методов заполнения (средневыборочных, по регрессии) и при существенном количестве пропусков.

Пропущенные данные в наборе обучающих данных могут уменьшить мощность / пригодность модели или привести к необъективной модели, поскольку неправильно были проанализированы поведение и взаимосвязь с другими переменными. Это может привести к неправильному прогнозированию или классификации (рисунок 5.1).

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55		Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57		Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	2	1	50%
M	4	2	50%
Missing	2	2	100%

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55	F	Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57	F	Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	4	3	75%
M	4	2	50%

Рисунок 5.1 – Данные по игре крикет по гендеру

Следует обратить внимание на пропущенные значения на рисунке 5.1: в левом сценарии не рассмотрены пропущенные значения. Вывод из этого набора данных состоит в том, что шансы играть в крикет у мужчин выше, чем у женщин. С другой стороны, если посмотреть на вторую таблицу, которая показывает данные после обработки пропущенных значений (в зависимости от пола), видно, что женщины имеют более высокие шансы играть в крикет по сравнению с мужчинами.

Практическое задание

На основе данных в предложенном файле SourceData.csv разработать программный модуль, который позволяет спрогнозировать расход топлива. Далее случайным образом изменить данные путём удаления данных в ячейках или же строки данных, после чего восстановить данные, используя методы 1–3, описанные в теоретическом разделе, и сравнить прогноз.

Для реализации программного модуля рекомендуется использовать язык программирования Python.

Для защиты необходимо представить отчёт с результатами прогнозирования и программный код.

Контрольные вопросы

1 Выделите несколько категорий задач, решаемых с помощью машинного обучения.

2 Какие способы восстановления данных вы знаете?

3 Назовите минусы восстановленных данных.

Список литературы

1 **Дадян, Э. Г.** Методы, модели, средства хранения и обработки данных: учебник / Э. Г Дадян, Ю. А. Зеленков. – Москва: Вузовский учебник; ИНФРА-М, 2017. – 168 с.

2 **Дюк, В. А.** Применение технологий интеллектуального анализа данных в естественно-научных, технических и гуманитарных областях / В. А. Дюк, А. В. Флегонтов, И. К. Фомина // Изв. Рос. гос. педагогического ун-та им. А. И. Герцена. – 2011. – № 138. – С. 77–84.

3 **Замятин, А. В.** Интеллектуальный анализ данных: учебное пособие / А. В. Замятин. – Томск: Томский гос. ун-т, 2016. – 120 с.

4 **Каштанов, В. А.** Случайные процессы: учебник и практикум для прикладного бакалавриата / В. А. Каштанов, Н. Ю. Энатская. – Москва: Юрайт, 2019. – 156 с.

5 **Боев, В. Д.** Имитационное моделирование систем: учебное пособие для прикладного бакалавриата / В. Д. Боев. – Москва: Юрайт, 2019. – 253 с.

6 **Дронов, В. А.** Django 2.1. Практика создания веб-сайтов на Python / В. А. Дронов. – Санкт-Петербург: BHV, 2019. – 672 с.