МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САМАРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА»
(САМАРСКИЙ УНИВЕРСИТЕТ)

ЭКОНОМЕТРИКА (ПРОДВИНУТЫЙ УРОВЕНЬ)

Методические указания

Рекомендовано редакционно-издательским советом федерального государственного автономного образовательного учреждения высшего образования «Самарский национальный исследовательский университет имени академика С.П. Королева» в качестве методических указаний для студентов Самарского университета, обучающихся по основной образовательной программе высшего образования по направлению подготовки 38.04.01 Экономика

Составители: *Е.А. Блинова,*
О.А. Кузнецова

Самара
Издательство Самарского университета
2020

УДК 330.4(075)
ББК 65в6я7
Э400

Составители: ***Е.А. Блинова, О.А. Кузнецова***

Рецензент: д-р экон. наук, проф. Д.Ю. Иванов

Эконометрика (продвинутый уровень): методические указания по выполнению лабораторных работ / *Е.А. Блинова, О.А. Кузнецова*. – Электрон. текст. дан. (1,1 Мб). – Самара: Издательство Самарского университета, 2020. – 1 опт. компакт-диск (CD-RM). – Систем. требования: PC, процессор Pentium, 160 МГц; оперативная память 32 Мб, на винчестере 16 Мб; Microsoft Windows XP/7/10; разрешение экрана 1024x768 с глубиной цвета 16 бит; DVD-ROM 2-х и выше, мышь; Adobe Acrobat Reader. – Загл. с титул. экрана.

Приведены примеры выполнения и задания для самостоятельного выполнения лабораторных работ по четырем темам.

Предназначено для студентов Самарского университета, обучающихся по направлению подготовки 38.04.01 Экономика.

Подготовлено на кафедре Математических методов в экономике.

Авторы данных методических указаний – победители грантового конкурса Стипендиальной программы Владимира Потанина 2018/19.

УДК 330.4(075)
ББК 65в6я7

Редактор Л.Р. Дмитриенко
Компьютерная верстка Л.Р. Дмитриенко

Подписано для тиражирования 22.07.2020.
Объем издания 1,1 Мб.
Количество носителей 1 диск.
Тираж 10 дисков.

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САМАРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА»
(САМАРСКИЙ УНИВЕРСИТЕТ)
443086 Самара, Московское шоссе, 34.

Изд-во Самарского университета.
443086 Самара, Московское шоссе, 34.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Целями дисциплины являются формирование у обучающихся:

1) способности составления прогнозных моделей основных социально-экономических показателей;

2) способности оценивать эффективность проектов с учётом фактора неопределённости.

Задачи:

- изучить методы прогнозирования основных социально-экономических показателей;
- научить обучающихся применять инструменты для прогнозирования основных социально-экономических показателей;
- развить навыки прогнозирования основных социально-экономических показателей;
- освоить методы оценки эффективности проектов с учётом фактора неопределённости;
- научить обучающихся применять инструменты оценки неопределённости;
- развить навыки оценки эффективности проектов с учётом фактора неопределённости.

Лабораторная работа 1
ПАРНАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

Цель работы: научиться строить модель парной линейной регрессии с помощью MS Excel и оценивать ее качество.

Пример выполнения работы:

Таблица 1. **Исходные данные**

№	*y*	*x*
1	33,66	120,00
2	28,56	112,30
3	20,40	107,25
4	27,54	107,25
5	48,96	127,25
6	21,42	112,20
7	22,44	113,85
8	43,86	122,10
9	47,94	122,10
10	72,42	132,10
11	79,56	127,05
12	55,08	127,05
13	72,42	128,70
14	77,52	132,00
15	81,60	133,65
16	56,10	128,60
17	77,52	140,25
18	51,00	130,25
19	66,30	130,25

В табл. 1 приведена статистика распределения расходов на потребление продуктов питания *y* и средней заработной платы *x*.

Общий вид парной линейной регрессии

$$y = a + b \bullet x + \varepsilon.$$

Необходимо: определить коэффициенты парной линейной, оценить качество полученной модели.

Порядок выполнения работы:

1. Проверяем достаточность статистической информации – данных должно быть, как минимум, в 7 раз больше, чем факторных переменных.

В данном случае факторная переменная 1, число наблюдений 19 – условие выполняется.

2. Построить область распределения данных *y* от *x* на основе статистической информации (рис. 1).

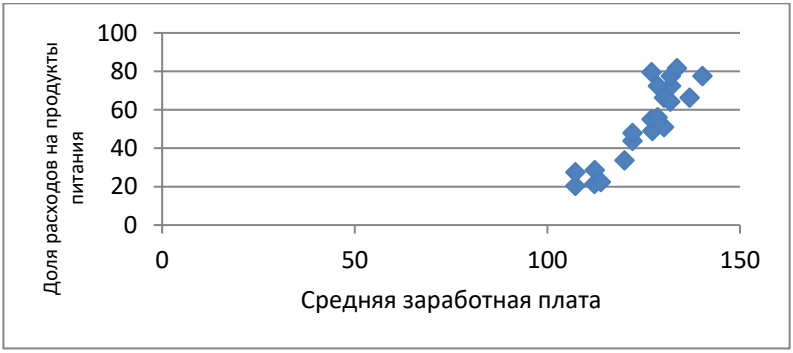

Рис. 1. Распределения расходов на потребление продуктов питания от средней заработной платы

На рис. 1 наблюдается линейная зависимость.

3. Делим базу статистических данных на 2 части: 90% используем для нахождения коэффициентов регрессии и 10% оставляем для проверки качества построенной модели.

4. Используем инструмент MS Excel «данные – анализ данных – регрессия» для вывода результатов регрессионного анализа (рис. 2).

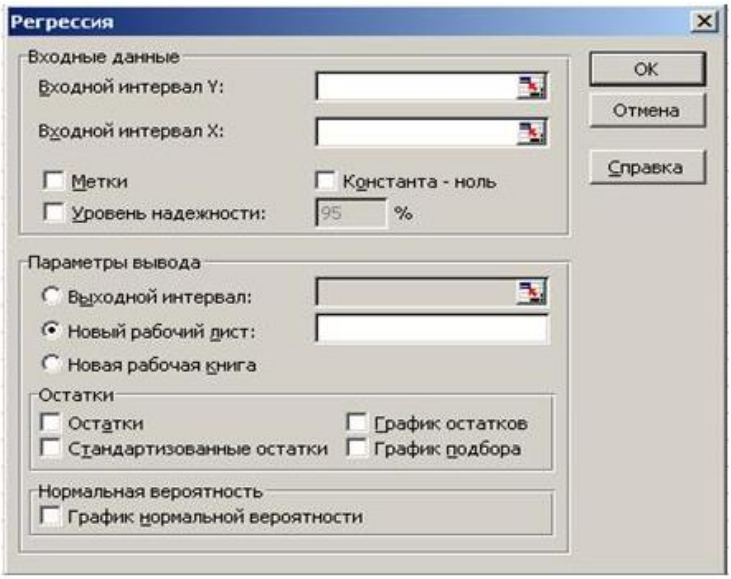

Рис. 2. Диалоговое окно инструмента анализа «Регрессия»

Диалоговое окно заполняется следующим образом:

- «входной интервал *y*» – выделяем столбец с результирующими данными о расходах на потребление продуктов питания вместе с заголовком из табл. 1;
- «входной интервал *x*» – выделяем столбец с результирующими данными о средней заработной плате вместе с заголовком из табл. 1;

- ставим значок в окне «метки» и «уровень надежности» (95%);
- «выходной интервал» – выделяем любую ячейку на свободном месте страницы;
- нажимаем «ОК» и на листе появятся таблицы с результатами расчетов (рис. 3).

5. Для записи уравнения регрессии и оценки ее качества необходимо посмотреть на ряд показателей (рис. 3).

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,91
R-квадрат	0,82
Нормированный R-квадрат	0,81
Стандартная ошибка	9,37
Наблюдения	19

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	6893,42	6893,42	78,47	0,00
Остаток	17	1493,41	87,85		
Итого	18	8386,83			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	-205,33	29,11	-7,05	0,00	-266,74	-143,92
x	2,08	0,23	8,86	0,00	1,58	2,57

Рис. 3. Результаты регрессионного анализа

Множественный *R* (коэффициент корреляции), равный 0,91, показывает, что между факторной и результирующей переменной существует очень сильная прямая связь. То есть, при увеличении

средней заработной платы доля расходов на потребление продуктов питания увеличивается.

R-квадрат (коэффициент детерминации), равный 0,82, означает, что увеличение расходов на потребление продуктов питания на 82% зависит от средней заработной платы. Чем выше значение коэффициента детерминации, тем выбранная модель считается более применимой для конкретной задачи. Считается, что она корректно описывает реальную ситуацию при значении *R*-квадрата выше 0,8.

Нормированный *R*-квадрат, объективно определяет достоверность связи, так как, в отличие от обычного коэффициента детерминации, он не зависит от числа наблюдений и числа факторов.

Число наблюдений 19.

Таблица «Дисперсионный анализ» включает в себя обусловленные регрессией («Регрессия»), необусловленные регрессией («Остаток») и суммарные:

- число степеней свободы *df*;
- сумму квадратов разностей (дисперсии *SS*);
- оценки дисперсий, приходящихся на одну степень свободы (*MS*).

Критерий Фишера *F* показывает правильность выбора формы модели. *F*-фактическое значение *F*-критерия Фишера, значимость *F*-табличное, т.к. *F* > значимости *F*, 78,47 > 0, то корреляционно-регрессионную модель можно считать адекватной.

Таблица результатов собственно регрессионного анализа (информация об уравнении регрессии) включает в себя:

Коэффициенты регрессии, по данным таблицы строим регрессионное уравнение

$$Y = -205{,}33 + 2{,}08x.$$

Отрицательное значение свободного члена регрессии не имеет экономического смысла. Коэффициент при x показывает, на сколько единиц изменится результат при увеличении факторной переменной на 1 единицу.

Критерий Стьюдента (t-статистика) показывает правильность расчета каждого коэффициента регрессии и, соответственно, правильность включения переменных в модель. Если расчетное значение t-статистики по модулю больше табличного, то коэффициент принимается. По данным таблицы (рис. 3):

$$ta = -7{,}05;\ tb = 8{,}86.$$

Табличные значения t-статистики определяются по таблице:

$$Tкрит. = 2{,}11.$$

Оба значения t-статистики по модулю больше критического значения, следовательно, оба коэффициента регрессии значимы.

P-значение – вероятность отказа от справедливой гипотезы. Если значение $p > 0{,}05$, то коэффициент регрессии считается равным 0.

В нашем случае оба p-значения равны нулю, что еще раз подтверждает значимость коэффициентов регрессии.

Нижние 95% и верхние 95% – это нижняя и верхняя границы значений коэффициентов. То есть, найденный коэффициент не является абсолютно точным, и его значение фактически может колебаться в каком-то интервале. Самое главное, чтобы обе границы интервала имели одинаковый знак.

Таким образом, построенная модель является качественной. Последняя проверка прогнозных качеств модели на 10% данных приведена в табл. 2.

Таблица 2. **Прогнозирование по построенной модели**

№	*y*	*x*	прогноз y
20	64,26	131,9	69,022
21	66,3	136,9	79,422

Можно отметить, что результат прогнозирования не обеспечивает 100% точность результата. Тем не менее, результат можно считать приемлемым.

Задание для самостоятельной работы

На сайте Росстат собрать статистическую информацию (по вариантам) и построить модель парной линейной регрессии.

1. Цена на зерно и курс доллара (по годам).
2. Курс рубля и ВВП России (по годам).
3. Курс рубля и средняя заработная плата (по годам).
4. Средняя заработная плата и величина потребительской корзины (по регионам).
5. Курс рубля и МРОТ (по регионам).
6. Курс рубля и цена за обучение в Вузе (по годам).
7. Средняя заработная плата и расходы на продукты питания (по регионам).
8. Ставка по кредиту и общая сумма выданных автокредитов (по годам).
9. Объем ипотечного кредитования и объем введенных жилых площадей (по годам).
10. Количество браков и количество рожденных детей (по годам).

Лабораторная работа 2
ПАРНАЯ НЕЛИНЕЙНАЯ РЕГРЕССИЯ

Цель работы: научиться строить модель парной нелинейной регрессии с помощью MS Excel и оценивать ее качество.

Пример выполнения работы:

Исходные данные в табл. 3.

Таблица 3

№	*y*	*x*
1	33,66	120,00
2	28,56	112,30
3	20,40	107,25
4	27,54	107,25
5	48,96	127,25
6	21,42	112,20
7	22,44	113,85
8	43,86	122,10
9	47,94	122,10
10	72,42	132,10
11	79,56	127,05
12	55,08	127,05
13	72,42	128,70
14	77,52	132,00
15	81,60	133,65
16	56,10	128,60
17	77,52	140,25
18	51,00	130,25
19	66,30	130,25
20	64,26	131,90
21	66,30	136,90

В табл. 3 приведена статистика распределения расходов на потребление продуктов питания *y* и средней заработной платы *x*.

1. Проверяем достаточность статистической информации – данных должно, быть как минимум, в 7 раз больше, чем факторных переменных.

В данном случае факторная переменная 1, число наблюдений 21 – условие выполняется.

2. Необходимо построить область распределения данных *y* от *x* на основе статистической информации.

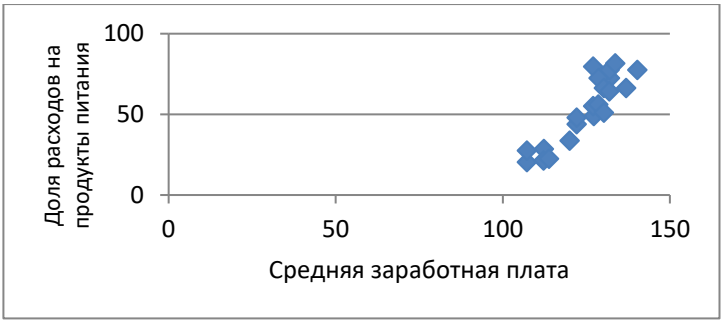

Рис. 4. Распределения расходов на потребление продуктов питания от средней заработной платы

Облако распределения имеет не строго линейную форму, поэтому попробуем построить другую модель – степенную регрессию.

Общий вид степенной функции $y = ax^b\varepsilon$.

Для нахождения коэффициентов регрессии в MS Excel используется метод наименьших квадратов, который применяется только для линейных моделей. Следовательно, если мы хотим воспользоваться этим инструментом для нахождения коэффициентов степенной функции, необходимо преобразовать ее в линейный вид.

Привести уравнение к линейному виду можно с помощью логарифмирования

$$Ln\ y = ln\ a + b \cdot ln\ x + ln\ \varepsilon,$$

далее можно найти коэффициенты регрессии, но сначала требуется преобразовать статистическую информацию (табл. 4).

3. Делим базу статистических данных на 2 части: 90% используем для нахождения коэффициентов регрессии и 10% оставляем для проверки качества построенной модели.

4. Используем инструмент MS Excel «данные – анализ данных – регрессия» (см. рис. 5).

В отличие от примера с парной линейной регрессией в ячейке «входной интервал *y*» – выделяем столбец «ln *y*» вместе с заголовком из табл. 4;

- «входной интервал *x*» – выделяем столбец «ln *x*» вместе с заголовком из табл. 4;
- ставим значок в окне «метки» и «уровень надежности» (95%);
- «выходной интервал» – выделяем любую ячейку на свободном месте страницы;
- нажимаем «ОК» и на листе появятся таблицы с результатами расчетов (рис. 6).

5. Для оценки качества построенной регрессии необходимо посмотреть на ряд показателей (рис. 6):

Множественный *R* (коэффициент корреляции) равный 0,99 показывает, что между факторной и результирующей переменной существует очень сильная прямая связь.

R-квадрат (коэффициент детерминации) равный 0,99 означает, что увеличение расходов на потребление продуктов питания на 99% зависит от средней заработной платы. Чем выше значение коэффициента детерминации, тем выбранная модель считается

более применимой для конкретной задачи. Считается, что она корректно описывает реальную ситуацию при значении *R*-квадрата выше 0,8.

Таблица 4. **Подготовка статистической информации**

№	*y*	*x*	ln *y*	ln *x*
1	33,66	120,00	3,52	4,79
2	28,56	112,30	3,35	4,72
3	20,40	107,25	3,02	4,68
4	27,54	107,25	3,32	4,68
5	48,96	127,25	3,89	4,85
6	21,42	112,20	3,06	4,72
7	22,44	113,85	3,11	4,73
8	43,86	122,10	3,78	4,80
9	47,94	122,10	3,87	4,80
10	72,42	132,10	4,28	4,88
11	79,56	127,05	4,38	4,84
12	55,08	127,05	4,01	4,84
13	72,42	128,70	4,28	4,86
14	77,52	132,00	4,35	4,88
15	81,60	133,65	4,40	4,90
16	56,10	128,60	4,03	4,86
17	77,52	140,25	4,35	4,94
18	51,00	130,25	3,93	4,87
19	66,30	130,25	4,19	4,87
20	64,26	131,90	4,16	4,88
21	66,30	136,90	4,19	4,92

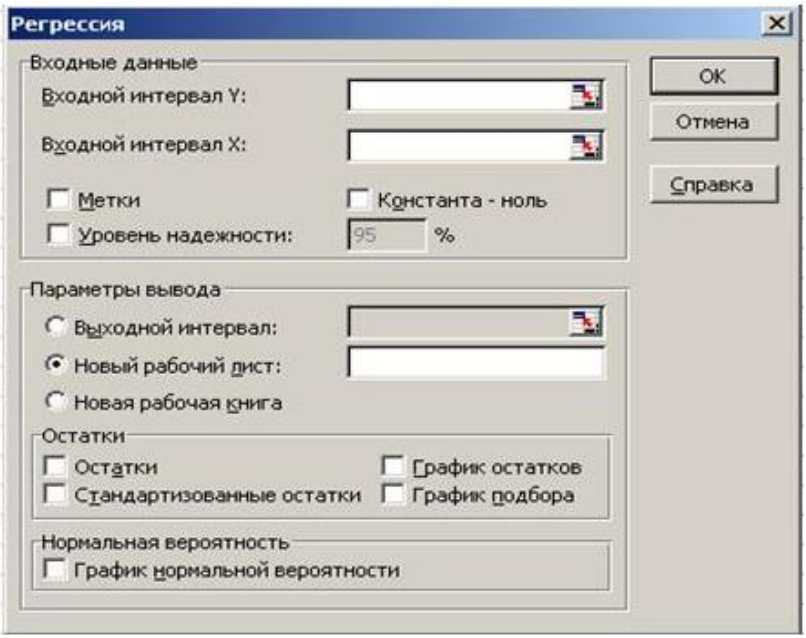

Рис. 5. Диалоговое окно инструмента анализа «Регрессия»

R-квадрат (коэффициент детерминации) равный 0,99 означает, что увеличение расходов на потребление продуктов питания на 99% зависит от средней заработной платы. Чем выше значение коэффициента детерминации, тем выбранная модель считается более применимой для конкретной задачи. Считается, что она корректно описывает реальную ситуацию при значении *R*-квадрата выше 0,8.

Нормированный *R*-квадрат, объективно определяет достоверность связи, так как в отличие от обычного коэффициента детерминации, он не зависит от числа наблюдений и числа факторов. 0,93 является хорошим результатом.

Число наблюдений 19.

Таблица «Дисперсионный анализ» включает в себя обусловленные регрессией («Регрессия»), необусловленные регрессией («Остаток») и суммарные:

- число степеней свободы *df*;
- сумму квадратов разностей (дисперсии *SS*);
- оценки дисперсий, приходящихся на одну степень свободы (*MS*).

ВЫВОД ИТОГОВ							
Регрессионная статистика							
Множественный R	0,99434						
R-квадрат	0,98872						
Нормированный R-квадрат	0,93316						
ошибка	0,42306						
Наблюдения	19						
Дисперсионный анализ							
	df	*SS*	*MS*	*F*	*Значимость F*		
Регрессия	1	282,348	282,348	1577,56	3,3E-18		
Остаток	18	3,22159	0,17898				
Итого	19	285,569					
	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	
Y-пересечение	0	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д	
ln x	0,80023	0,02015	39,7185	5,5E-19	0,7579	0,84255	

Рис. 6. Результаты регрессионного анализа

Критерий Фишера *F* показывает правильность выбора формы модели. *F*-фактическое значение *F*-критерия Фишера, значимость

F-табличное, т.к. *F* > значимости *F*, 1577,56 > 3,3E-18, то корреляционно-регрессионную модель считать адекватной.

Таблица результатов собственно регрессионного анализа (информация об уравнении регрессии) включает в себя:

Коэффициенты регрессии

$$ln\ a = 0;$$

$$b = 0{,}8;$$

$$lna=0.$$

Чтобы найти значение параметра *a*, необходимо провести процедуру потенциирования:

$$a = e^{\ln a};$$

$$a = 2{,}710;$$

$$a = 1.$$

Следовательно, степенная регрессия будет записана как

$$y = 1\ x0{,}8\ \varepsilon.$$

Отрицательное значение свободного члена регрессии не имеет экономического смысла. Коэффициент *b* показывает на сколько процентов изменится результат при увеличении факторной переменной на 1 процент.

Критерий Стьюдента (*t*-статистика) показывает правильность расчета каждого коэффициента регрессии и, соответственно, правильность включения переменных в модель. Если расчетное значение *t*-статистики по модулю больше табличного, то коэффициент принимается. По данным таблицы (рис. 6)

$$tb = 39{,}7.$$

Табличные значения **t**-статистики определяются по таблице критериев Стьюдента, учитывая число наблюдений = 19:

$$t\text{крит.} = 2{,}11.$$

Значение *t*-статистики по модулю больше критического значения, следовательно, оба коэффициента регрессии значимы.

P-значение – вероятность отказа от справедливой гипотезы. Если значение p > 0,05, то коэффициент регрессии считается равным 0. В нашем случае p-значение равно 5,5E-19, что еще раз подтверждает значимость коэффициентов регрессии.

Нижние 95% и верхние 95% – то нижняя и верхняя границы значений коэффициентов. То есть, найденный коэффициент не является абсолютно точным и его значение фактически может колебаться в каком-то интервале. Самое главное, чтобы обе границы интервала имели одинаковый знак.

Таким образом, построенная модель является качественной. Мы убедились в качестве построенной логарифмической модели, следовательно, исходная степенная модель тоже качественная.

Таблица 5. **Прогнозирование по построенной модели**

№	*y*	*x*	прогноз *y*
20	64,26	131,9	50
21	66,3	136,9	51

Можно отметить, что результат прогнозирования не обеспечивает 100% точность результата. Тем не менее, результат можно считать приемлемым.

Задание для самостоятельной работы

На сайте Росстат собрать статистическую информацию (по вариантам) и построить модель парной нелинейной регрессии.

1. Цена на зерно и курс доллара (по годам).
2. Курс рубля и ВВП России (по годам).
3. Курс рубля и средняя заработная плата (по годам).
4. Средняя заработная плата и величина потребительской корзины (по регионам).
5. Курс рубля и МРОТ (по регионам).
6. Курс рубля и цена за обучение в Вузе (по годам).
7. Средняя заработная плата и расходы на продукты питания (по регионам).
8. Ставка по кредиту и общая сумма выданных автокредитов (по годам).
9. Объем ипотечного кредитования и объем введенных жилых площадей (по годам).
10. Количество браков и количество рожденных детей (по годам).

Лабораторная работа 3
КЛАСТЕРИЗАЦИЯ

Цель работы: сформировать однородные группы объектов в целях выявления регрессионных зависимостей.

Пример выполнения работы:

Есть некоторое количество кроликов, описываемых такими параметрами, как цвет шерсти и длина ушей (табл. 6). Выделить группы сразу по двум признакам трудно, поэтому используем средства программного комплекса R.

Таблица 6. **Исходные данные для кластеризации**

Параметр	номер объекта					
	1	2	3	4	5	6
tsvet	Белый	Черн.	Черн.	Белый	Белый	Черн.
ush	Длин.	Кор.	Длин.	Кор.	Кор.	Длин.

Комплекс R устанавливается бесплатно. https://cran.r-project.org/bin/windows/base/

Нужно установить R, подходящий для Вашей версии компьютера, и R-studio.

ВАЖНО! Название файла и названия в таблице должны быть на английском языке.

Нумерацию строк можно не делать.

Начинаем с создания нового файла. Для этого в меню выбираем File – R script. В верхней левой четверти окна появляется свободное поле, где пишем код (рис. 7).

Перед началом работы необходимо установить основные библиотеки: ggplot2", psych, dplyr.

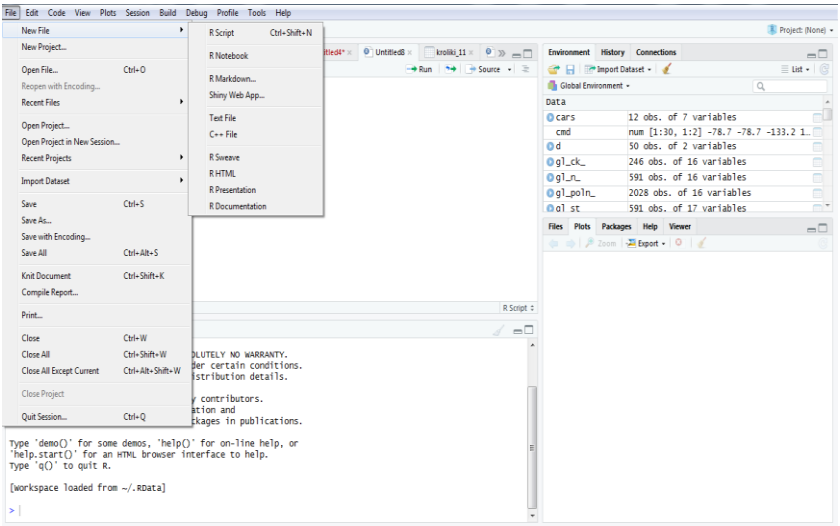

Рис. 7. Интерфейс программы R

Кластеризация методом K-means

При кластеризации по одному параметру: цвет (tsvet).

Кластеризация по одному параметру является самой простой и понятной. Здесь деление на объекты одного кластера будут содержать строго одинаковые параметры.

Подготавливаем таблицу с данными в формате MS Excel (важно, чтобы файл состоял из одной страницы, название файла латинскими буквами). Наш файл назван kroliki_12.

Качественные характеристики для удобства можно перевести в количественные, присвоив характеристике «Белый» значение 1 и характеристике «Черный» значение 2, «Длинные уши» – 3, «Короткие уши» – 4 (табл. 7).

Таблица 7. **Характеристики объектов наблюдения**

Параметр	номер объекта					
	1	2	3	4	5	6
tsvet	1	2	2	1	1	2
ush	4	3	4	3	3	4

Для того, чтобы работать с табличными данными необходимо загрузить таблицу. Для этого в левом верхнем углу окна ищем Import Dataset – From Excel – появляется новое окно. Там в строке File выбираем (прописываем) путь к файлу с исходными данными (рис. 8).

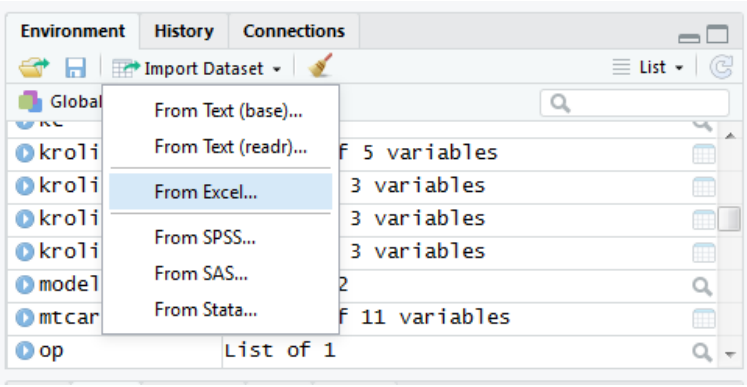

Рис. 8. Импорт данных в R

В поле Data Preview появляется таблица с исходными данными, для завершения процедуры загрузки данных необходимо нажать кнопку Import. После этого в левом верхнем окне появится вкладка с именем файла kroliki_11. Обратите внимание, при импорте название файла может немного измениться. Важно использовать в коде ту версию названия, которая отразилась на вкладке в программе R.

После установки библиотек и импорта данных прописываем, к какому файлу обращаемся за информацией (рис. 9).

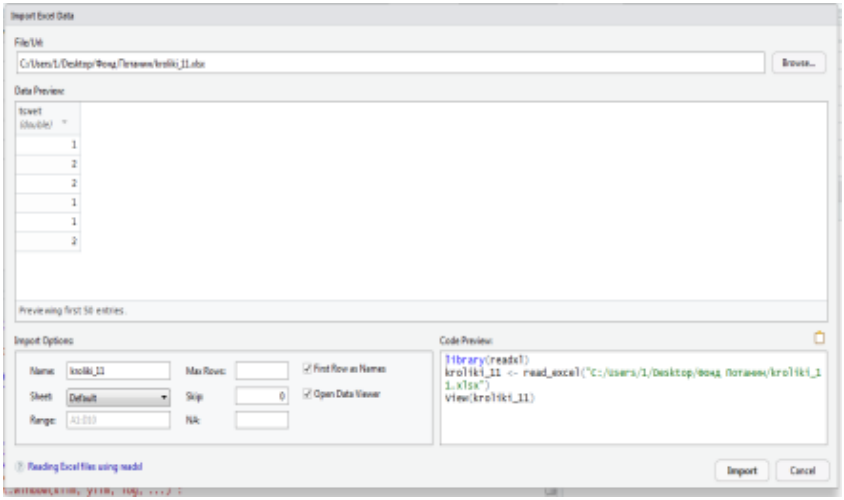

Рис. 9. Импорт данных в R

Ниже приведен код для кластеризации.

```
library("ggplot2")
library("psych")
library("dplyr")
kroliki_12<-kroliki_12[1:6, 1:2]
kroliki_12
k1 <- (nrow(kroliki_12)-1)*sum(apply(kroliki_12, 2, var)) for (i in 2:2) k1[i] <- sum(kmeans(kroliki_12,centers=i)$withinss)
plot(1:2, k1, type="b", x1ab="clusters number", ylab="The sum of squares of distances inside the clusters")
kc <- kmeans(kroliki_12, 2)
aggregate(kroliki_12, by=list(kc$cluster), FUN=mean)
kroliki_12 <- data.frame(kroliki_12, kc$cluster)
```

```
op <- par(mfrow = c(1, 2))
plot(kroliki_12[c("tsvet", "ush")], co1=kc$cluster)
points(kc$centers[, c("tsvet", "ush")], col=1:2, pch=8, cex=2)
```

Далее строим график (рис. 10).

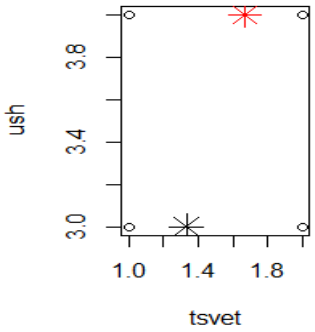

Рис. 10. Распределение объектов наблюдения в двухмерном пространстве

На рис. 10 снежинками отмечены центры кластеров.

R также фиксирует распределение объектов по кластерам (рис. 11).

	tsvet	ush	kc.cluster
1	1	4	2
2	2	3	1
3	2	4	2
4	1	3	1
5	1	3	1
6	2	4	2

Рис 11. Распределение объектов по кластерам

При распределении по кластерам в первом кластере будут кролики: 2, 4 и 5. Общим признаком для них будет длина ушей.

Также можно осуществить кластеризацию методом построения дендрограммы.

Кластеризация методом построения дендрограммы

Кластеризация по двум признакам. Исходные данные хранятся в файле kroliki_12.

```
library("ggplot2")
library("psych")
library("dplyr")
kroliki_12<-kroliki_12[1:6, 1:2]
kroliki_12
hc <- hclust(dist(kroliki_12))
plot(hc)
```

Результаты кластеризации представлены на рис. 12.

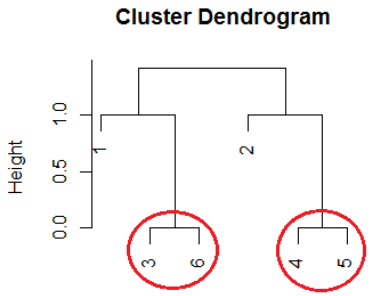

Рис. 12. Схема распределения объектов по кластерам

В данном методе кластеризации деление по группам идет «сверху вниз». Сначала по одному критерию, в данном случае

«ush» были сформированы две группы: 1, 3, 6 и 2, 4, 5. Но цвет шерсти в кластерах был смешанный. Следующим действием программа выделила более однородные кластеры, где совпадали и длина ушей, и цвет шерсти кроликов: 3, 6 и 4, 5. В этом случае объекты 1 и 2 оказались вне кластеров. Принять решение о том, выбрать кластеризацию по трем объектам или по двум, можно с помощью величины height.

Задание для самостоятельной работы

На основе данных сайта Росстат провести кластеризацию объектов.

1. Потребителей туристической фирмы для формирования продукта.

Параметры: направление тура, стоимость тура, тип отдыха, семейное положение/тип туриста.

2. Для риелторов – кластеризация квартир по типу.

Параметры: количество комнат, район, тип дома.

3. В рамках темы своей магистерской диссертации осуществить кластеризацию:

- товаров, в целях формирования товарных групп;
- потребителей, для формирования «идеального» продукта, разработки стратегии продвижения товара.

Лабораторная работа 4
ЛОГИТ И ПРОБИТ МОДЕЛИ

Цель работы: научиться строить модели для определения вероятности осуществления какого-либо события на основе качественных данных.

Задание: построить прогнозные модели логит и пробит на основе качественных данных.

Пример выполнения

Был проведен опрос на тему «Место женщины у плиты». В табл. 8 представлены характеристики респондентов.

Таблица 8. **Результаты опроса на тему согласия с тезисом «Место женщины у плиты»**

	agree	id	age	age2	Disagree	mw14	Meduc	adjinc	nsibs	fpro	cath	so	urb	Tradrole
1	0	1	20,33	413,4	1	0	8	13416	1	0	1	0	1	2
2	1	2	20	400	0	11	5	8944	8	0	1	0	1	4
3	0	3	17,42	303,3	1	0	10	10013	3	0	1	0	1	2
4	0	4	16,42	269,5	1	1	11	10013	3	0	1	0	1	1
5	0	8	20,5	420,2	1	1	9	4173	7	1	0	0	1	2
6	0	10	18,25	333,1	1	0	12	2048	3	0	1	0	1	2

Качественные показатели представлены в виде числовых значений.

Общий вид логистической функции:

$$p = F(Z) = \frac{1}{1+e^{-Z}},$$

где $Z = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k$.

Регрессионную функцию Z удобно строить с помощью инструмента MS Excel «данные-регрессия». На рис. 13 приведен вывод итогов.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множеств	1
R-квадрат	1
Нормиров	65535
Стандарт	0
Наблюден	6

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	13	0,83333	0,0641	#ЧИСЛО!	#ЧИСЛО!
Остаток	0	0	65535		
Итого	13	0,83333			

	Коэффициенты	*артная ошибка*	*татистика t*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*жние 95,0%*	*рхние 95,0%*
Y-пересеч	0,99051	0	65535	#ЧИСЛО!	0,990506037	0,990506037	0,99051	0,99051
id	-0,08189	0	65535	#ЧИСЛО!	-0,081888266	-0,081888266	-0,08189	-0,08189
age	0	0	65535	#ЧИСЛО!	0	0	0	0
age2	0,00019	0	65535	#ЧИСЛО!	0,000185628	0,000185628	0,00019	0,00019
disagree	0	0	65535	#ЧИСЛО!	0	0	0	0
mw14	0,08816	0	65535	#ЧИСЛО!	0,088162483	0,088162483	0,08816	0,08816
meduc	0	0	65535	#ЧИСЛО!	0	0	0	0
adjinc	-7,1E-05	0	65535	#ЧИСЛО!	-7,12726E-05	-7,12726E-05	-7,1E-05	-7,1E-05
nsibs	-0,02916	0	65535	#ЧИСЛО!	-0,029163235	-0,029163235	-0,02916	-0,02916
fpro	0	0	65535	#ЧИСЛО!	0	0	0	0
cath	0	0	65535	#ЧИСЛО!	0	0	0	0
so	0	0	65535	#ЧИСЛО!	0	0	0	0
urb	0	0	65535	#ЧИСЛО!	0	0	0	0
tradrole	0	0	65535	#ЧИСЛО!	0	0	0	0

Рис. 13. Вывод итогов регрессионной функции

Необходимо убрать из данных информацию о параметрах с нулевым коэффициентом. В результате получаем таблицу с коэффициентами регрессии (рис. 14).

При этом все показатели качества построенной регрессии подтверждают, что уравнение построено верно.

ВЫВОД ИТОГОВ

Регрессионная статистика

Множеств	1
R-квадрат	1
Нормиров	65535
Стандарт	0
Наблюде	6

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*ачимость F*
Регрессия	5	0,83333	0,166666667	#ЧИСЛО!	#ЧИСЛО!
Остаток	0	0	65535		
Итого	5	0,83333			

	Коэффициен	*артная ош*	*t-статистика*	*-Значение*	*ижние 95%*	*ерхние 95%*	*ижние 95,0*	*рхние 95,0%*
Y-пересеч	0,99051	0	65535	#ЧИСЛО!	0,99051	0,99051	0,99051	0,99051
id	-0,08189	0	65535	#ЧИСЛО!	-0,08189	-0,08189	-0,08189	-0,08189
age2	0,00019	0	65535	#ЧИСЛО!	0,00019	0,00019	0,00019	0,00019
mw14	0,08816	0	65535	#ЧИСЛО!	0,08816	0,08816	0,08816	0,08816
adjinc	-7,1E-05	0	65535	#ЧИСЛО!	-7,1E-05	-7,1E-05	-7,1E-05	-7,1E-05
nsibs	-0,02916	0	65535	#ЧИСЛО!	-0,02916	-0,02916	-0,02916	-0,02916

y= 0,991-0,08x1+0,0002x2+0,088x-0,000007x4-0,029x5

Рис. 14. Результаты регрессионного анализа

Подставляя регрессионное уравнение в логит-модель для имеющихся данных, проверяем ее адекватность в табл. 9.

Таблица 9. **Расчет вероятности осуществления события**

obs	id	age2	mw14	adjinc	nsibs	agree		P
1	1	413,4	0	13416	1	0		0,5
2	2	400	11	8944	8	1		0,7
3	3	303,3	0	10013	3	0		0,5
4	4	269,5	1	10013	3	0		0,5
5	8	420,2	1	4173	7	0		0,5
6	10	333,1	0	2048	3	0		0,5
Проба	12	3600	1	4173	3			0,4

Столбец P содержит значения вероятности реализации события. P меньше или равно 0,5 означает несогласие, P больше 0,5 согласие. Сравнивая полученные значения вероятности с фактическим согласием и несогласием респондентов, отмечаем верность прогноза.

Последняя строка «Проба» содержит новые значения параметров модели, для которых значение вероятности 0,4 прогнозирует отрицательный ответ на вопрос.

Задание для самостоятельной работы

1. Построить по данным примера Пробит-модель и сравнить результаты с Логит-моделью.

2. Провести статистический опрос на любую тему (не менее 10 респондентов), отобразить качественные характеристики респондентов и построить прогнозную Логит-модель. После чего проверить качество прогноза ее на трех респондентах.

ЗАКЛЮЧЕНИЕ

В результате выполнения лабораторных работ обучающимися освоены наиболее применяемые для выполнения магистерских диссертаций темы:

- парная линейная регрессия;
- парная нелинейная регрессия;
- кластеризация;
- логит и пробит модели.

У студентов сформированы способности составления прогнозных моделей основных социально-экономических показателей, способности оценивать эффективность проектов с учетом фактора неопределенности.

Изучены методы прогнозирования основных социально-экономических показателей; развиты навыки прогнозирования основных социально-экономических показателей.

Обучающиеся умеют применять инструменты для прогнозирования основных социально-экономических показателей и для построения моделей с качественными переменными.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Тимофеев В.С. Эконометрика: учебник / В.С. Тимофеев, А.В. Фадеенков, В.Ю. Щеколдин. – Новосибирск: НГТУ, 2014. – 345 с.
2. Герасимов А.Н. Эконометрика: продвинутый уровень: учеб. пособие / А.Н. Герасимов, Е.И. Громов, Ю.С. Скрипниченко. Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования Ставропольский государственный аграрный университет. – Ставрополь: Ставропольский государственный аграрный университет, 2016. – 272 с.
3. Котенко А.П. Эконометрика [Электронный ресурс]: интерактив. мультимед. пособие: система дистанц. обучения «Moodle». – Самара, 2012.
4. Красс, М.С. Математика для экономистов: учеб. пособие. – СПб., М., Нижний Новгород.: Питер, Питер Пресс, 2008. – 464 с.
5. Гладилин А.В. Эконометрика: учеб. пособие для вузов. – М.: КноРус, 2006. – 232 с.