МЕЖГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«БЕЛОРУССКО-РОССИЙСКИЙ УНИВЕРСИТЕТ»

Кафедра «Экономика и управление»

МНОГОМЕРНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ В ЭКОНОМИКЕ

Методические рекомендации к лабораторным работам для студентов направления подготовки 27.03.05 «Инноватика» дневной формы обучения

Могилев 2021

УДК 330.43
ББК 65.053
М73

Рекомендовано к изданию
учебно-методическим отделом
Белорусско-Российского университета

Одобрено кафедрой «Экономика и управление» «29» января 2021г., протокол № 6

Составитель ст. преподаватель Е. Г. Галкина

Рецензент канд. техн. наук, доц. Т. В. Пузанова

Методические рекомендации к лабораторным работам предназначены для студентов направления подготовки 27.03.05 «Инноватика» дневной формы обучения, изучающих дисциплину «Многомерный регрессионный анализ в экономике».

Учебно-методическое издание

МНОГОМЕРНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ В ЭКОНОМИКЕ

Ответственный за выпуск	И. В. Ивановская
Корректор	Е. А. Галковская
Компьютерная верстка	Н. П. Полевничая

Подписано в печать . Формат 60×16 /84. Бумага офсетная. Гарнитура Таймс.
Печать трафаретная. Усл. печ. л. .Уч.-изд. л. . Тираж 36 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, г. Могилев.

Содержание

Содержание и выполнение лабораторных работ

В результате выполнения лабораторной работы студенту необходимо научиться обоснованно и результативно применять методологические подходы и принципы аппарата эконометрического моделирования в прикладных исследованиях.

Ход выполнения лабораторной работы

1 Изучить теоретический материал по теме выполняемой лабораторной работы.

2 Получить индивидуальное задание у преподавателя.

3 Выполнить необходимые расчеты.

4 Сделать выводы по полученным результатам.

5 Оформить отчет.

Форма отчета

Результаты работы представить в виде:

– листа Excel с соответствующими формулами для их расчета. Сохранить на диске S в каталоге группы в своей папке файл с именем Лабораторная работа № X (X – номер лабораторной работы);

– отчета, оформленного произвольно, включающего:

а) цель работы;

б) постановку задачи;

в) краткое описание хода решения задачи (с пояснением применяемых условных обозначений, этапов расчета с приведением основных формул);

г) результаты расчета и выводы по полученным результатам.

К защите лабораторной работы допускаются только студенты, выполнившие работу и оформившие отчет. Защита проходит в форме устного и письменного собеседования, когда студент отвечает на вопросы преподавателя, примеры которых приведены в данных методических рекомендациях в списке контрольных вопросов к каждой лабораторной работе, а также дополняет свои ответы письменно примерами формул и расчетов по ним.

1 Лабораторная работа № 1. Предварительный анализ данных

Цель работы: исключение грубых ошибок измерений и проверка гипотезы о соответствии результатов измерений закону нормального распределения.

Задание к лабораторной работе.

Выполнить предварительную обработку результатов измерений, заключающуюся в исключении грубых ошибок и проверке гипотезы о соответствии результатов измерений закону нормального распределения.

Методические указания

1 Исключение грубых ошибок измерений.

Если число измерений n мало, то доверительный интервал широк, и даже значительные отклонения от среднего $\overline{x}$ в него укладываются. Если же n велико, то возрастает вероятность того, что хотя бы одно измерение x_i сильно отклонится от среднего на «законных основаниях», т. е. случайно.

Для больших выборок на практике используется следующий метод проверки однородности наблюдений.

Пусть произведено n независимых измерений и вычислены значения эмпирического среднего $\overline{x}$ и стандартного отклонения s. Сомнительный элемент выборки, резко отличающийся от других, будем обозначать через x_*. Это «крайний» элемент выборки, т. е. $x_* = x_{\max}$ или $x_* = x_{\min}$.

В основе рассматриваемого метода лежит тот факт, что критические значения максимального относительного отклонения

$$\tau = \frac{\left| x_* - \overline{x} \right|}{s} \tag{1}$$

выражаются через квантили распределения Стьюдента с $n-2$ степенями свободы:

$$\tau_{1-\alpha,n} = \frac{t_{1-\alpha,n-2}\sqrt{n-1}}{\sqrt{n-2+t^2_{1-\alpha,n-2}}}, \tag{2}$$

где $t_{1-\alpha,n-2}$ – t-критерий Стьюдента;

α – уровень значимости.

На практике обычно вычисляются два значения $\tau_{1-\alpha,n}$ при $\alpha = 0,05$ и $\alpha = 0,001$:

$$\tau_1 = \tau_{1-0,05;n}\,;\tau_2 = \tau_{1-0,001;n}. \qquad (3)$$

Этими значениями вся область изменения τ разбивается на три интервала: $\tau \le \tau_1$; $\tau_1 < \tau < \tau_2$; $\tau_2 \le \tau$. Наблюдения, попавшие в первый интервал, не рекомендуется отбрасывать ни в коем случае. Наблюдения, попавшие во второй интервал, можно исключить, если имеются какие-либо дополнительные соображения в пользу их ошибочности. Наконец, наблюдения, попавшие в третий интервал, всегда отбрасываются как грубо ошибочные.

2 Проверка гипотезы о нормальности распределения результатов измерения.

Приближенный метод проверки нормальности распределения основан на вычислении по результатам измерения эмпирических оценок коэффициентов асимметрии $\widehat{A}$, эксцесса $\widehat{E}$ и их дисперсий *D*(*A*) *и D*(*E*)*:*

$$\widehat{A} = \frac{\widehat{\mu}_3}{s^3} \approx \frac{1}{s^3(n-1)}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^3; \qquad (4)$$

$$\widehat{E} = \frac{\widehat{\mu}_4}{s^4} - 3 \approx \frac{1}{s^4(n-1)}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^4 - 3; \qquad (5)$$

$$D(A) = \frac{6(n-2)}{(n+1)(n+3)}; \qquad (6)$$

$$D(E) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}. \qquad (7)$$

Если выборочные асимметрия и эксцесс удовлетворяют неравенствам:

$$\left|\widehat{A}\right| \le 3\sqrt{D\left(\widehat{A}\right)}\,;\left|\widehat{E}\right| \le 5\sqrt{D\left(\widehat{E}\right)}\,, \qquad (8)$$

то гипотеза о нормальности наблюдаемого распределения принимается, в противном случае гипотеза отклоняется.

Если выборка достаточно велика, применяются иные критерии согласия, наиболее надежным и универсальным из которых является критерий Пирсона χ^2.

Контрольные вопросы

1 На основании чего выдвигается гипотеза о законе распределения?

2 Как описывается закон распределения?

3 Какой критерий используется для проверки гипотезы о нормальности распределения результатов измерения?

4 Как выбранный уровень значимости влияет на вывод?

2 Лабораторная работа № 2. Однофакторный и двухфакторный дисперсионный анализ

Цель работы: овладеть методикой анализа влияния одного или двух факторов на рассматриваемый признак.

Задание к лабораторной работе.

По данным индивидуального задания проверить нулевую гипотезу об отсутствии влияния фактора (уровней фактора) на результативный признак.

Методические указания

Однофакторный дисперсионный анализ (*ANOVA – analysis of variance*) используется для сравнения средних значений для трех и более выборок. *Фактором* называется *независимая* переменная, влияние которой изучается на *зависимую* переменную. Например, фактором может быть уровень образования, вид деятельности, возрастная группа респондентов, степень лояльности к торговой марке и т. д.

Анализ основан на расчете *F-статистики* (статистика Фишера), которая представляет собой отношение двух *дисперсий*: межгрупповой и внутригрупповой. *F*-тест в однофакторном дисперсионном анализе устанавливает, значимо ли отличаются средние нескольких независимых выборок.

Однофакторный дисперсионный анализ дает ответ на вопрос, влияет ли фактор на исследуемый показатель. Базовая идея состоит в том, что общая дисперсия признака раскладывается на составляющие, каждая из которых характеризует влияние того или иного фактора.

$$Q = Q_A + Q_o, \tag{9}$$

где Q – общая дисперсия;

Q_A – дисперсия (рассеяние характеризуется влиянием фактора A);

Q_o – остаточная дисперсия (рассеяние характеризуется влиянием других случайных факторов).

$$Q_A = n \cdot \sum_j^n (\bar{x}_j - \bar{\bar{x}})^2; \quad (10)$$

$$Q_o = \sum_j^n \sum_i^m (x_{ij} - \bar{x}_j)^2, \quad (11)$$

где m – количество групп;

n – количество единиц в каждой группе;

$\bar{x}$ – среднее значение признака.

Затем рассчитываются оценки дисперсий:

$$S_A^2 = \frac{Q_A}{m-1}; \quad (12)$$

$$S_o^2 = \frac{Q_o}{m \cdot (n-1)}. \quad (13)$$

На основе оценок дисперсий рассчитывают расчетное значение критерия Фишера, которое затем сравнивают с критическим (в MS Excel с *p*-level).

$$F_p = \frac{S_A^2}{S_o^2}. \quad (14)$$

Если $F_p > F_{кр}$, то делается вывод, что фактор влияет на исследуемый показатель.

Дисперсионный анализ можно проводить в MS Excel. Для этого нужно выбрать соответствующий вид дисперсионного анализа во вкладке *Данные* в группе *Анализ* / *Пакет анализа*.

Контрольные вопросы

1 С какой целью применяется дисперсионный анализ?

2 Какая нулевая и альтернативная гипотезы выдвигаются в случае однофакторного дисперсионного анализа?

3 Какая нулевая и альтернативная гипотезы выдвигаются в случае двухфакторного дисперсионного анализа?

4 Как выглядит правило сложения дисперсий в каждом случае?

5 Как определяется критическое значение?

3 Лабораторная работа № 3. Корреляционный анализ

Цель работы: овладеть методикой проведения парного и множественного корреляционного анализа.

Задание к лабораторной работе.

По данным индивидуального задания:

- выявить наличие взаимосвязи между признаками;
- определить формы связи;
- определить силы (тесноты) и направления связи.

Методические указания

1 Выявление наличия связи между признаками.

Простейшим визуальным способом выявить наличие взаимосвязи между количественными переменными является построение **диаграммы рассеяния** (*корреляционное поле*). Это график, на котором по горизонтальной оси (*X*) откладывается одна переменная, по вертикальной (*Y*) – другая. Каждому объекту на диаграмме соответствует точка, координаты которой равняются значениям пары выбранных для анализа переменных.

Направление связи. Связь может быть прямая и обратная. Связь прямая, если с увеличением одного признака второй возрастает (рисунок 1), и обратная, если с увеличением одного второй уменьшается (рисунок 2).

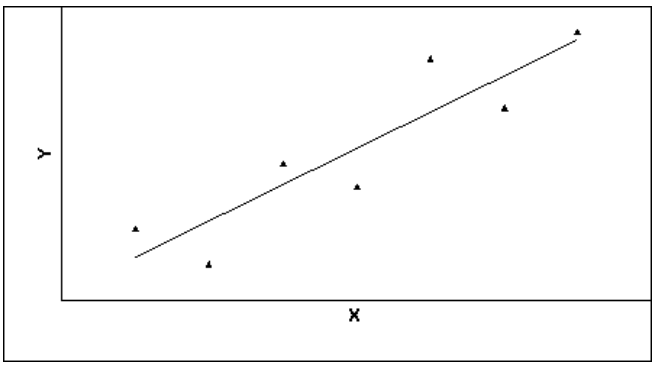

Рисунок 1 – Пример прямой связи

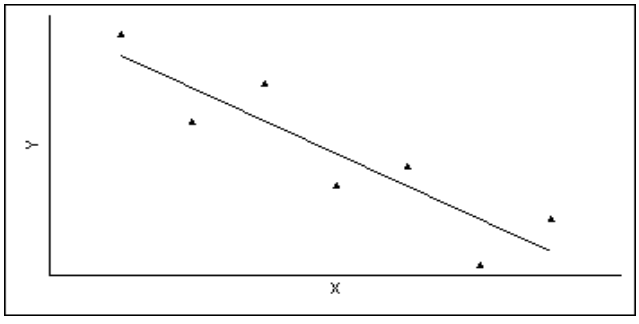

Рисунок 2 – Пример обратной связи

2 Подбор формы связи.

Если построенное облако точек напоминает очертания некоторой линии, то можно предполагать наличие зависимости, однако искаженную воздействием как случайных, так и неучтенных факторов, вызывающих отклонение точек от *теоретической* формы.

Поскольку наиболее простой формой зависимости в математике является прямая, то в корреляционном и регрессионном анализе наиболее популярны *линейные модели*.

Однако иногда расположение точек на диаграмме рассеяния показывает нелинейную зависимость либо вообще отсутствие связи между признаками.

3 Мера тесноты связи: ковариация и корреляция.

Выборочной ковариацией двух переменных *x*, *y* называется средняя величина произведения отклонений этих переменных от своих средних, т. е.

$$\operatorname{cov}(x, y) = \frac{1}{n}\sum (x_i - \bar{x})(y_i - \bar{y}), \tag{15}$$

где $\bar{x}, \bar{y}$ – выборочные средние переменных *x*, *y*.

Выборочная ковариация является *мерой взаимосвязи* между двумя переменными.

Более точной мерой зависимости между величинами является коэффициент корреляции. Различают *выборочный* и *теоретический* коэффициенты корреляции.

Выборочный коэффициент корреляции определяется выражением

$$r_{xy} = \frac{\operatorname{cov}(x, y)}{\sqrt{\operatorname{var}(x) \cdot \operatorname{var}(y)}};\ -1 \le r \le 1, \tag{16}$$

где $\operatorname{var}(x), \operatorname{var}(y)$ – выборочные средние квадратические отклонения величин *X,Y*.

Коэффициент является безмерной величиной и показывает *степень линейной связи* двух переменных. Выборочный коэффициент корреляции является случайной величиной.

Положительные значения коэффициента корреляции *r* свидетельствуют о прямой связи между признаками, отрицательные – об обратной связи.

Если *r* = 1, то между двумя переменными существует *функциональная прямая линейная связь*, т. е. на диаграмме рассеяния соответствующие точки лежат на одной прямой с положительным наклоном.

Если *r* = –1, то между двумя переменными существует *функциональная обратная линейная зависимость*, т. е. на диаграмме рассеяния соответствующие точки лежат на одной прямой с отрицательным наклоном.

Если *r* = 0, то рассматриваемые переменные линейно независимы, т. е. на диаграмме рассеяния облако точек «вытянуто по горизонтали».

Чем выше по модулю (по абсолютной величине) значение коэффициента корреляции, тем сильнее связь между признаками.

Принято считать, что коэффициенты корреляции, которые по модулю больше 0,7, говорят о сильной связи (при этом коэффициенты детерминации больше 50 %, т. е. один признак определяет другой более, чем наполовину).

Коэффициенты корреляции, которые по модулю меньше 0,7, но больше 0,5, говорят о связи средней силы (при этом коэффициенты детерминации меньше 50 %, но больше 25 %).

Наконец, коэффициенты корреляции, которые по модулю меньше 0,5, говорят о слабой связи (при этом коэффициенты детерминации меньше 25 %).

Коэффициент линейной корреляции (r) можно найти в MS Excel с помощью встроенной функции КОРРЕЛ().

Контрольные вопросы

1 Какова цель корреляционного анализа?

2 Как графически определить наличие/отсутствие связи между признаками?

3 Что такое ковариация?

4 По какой формуле вычисляется линейный коэффициент парной корреляции r_{xy}?

5 Как вычисляется индекс корреляции?

6 Как оценивается значимость парного линейного коэффициента корреляции?

7 Какой коэффициент определяет среднее изменение результативного признака при изменении факторного признака на 1 %?

4 Лабораторная работа № 4. Линейная парная регрессия

Цель работы: овладеть навыками определения параметров линейной регрессии.

Задание к лабораторной работе.

Построить уравнение линейной парной регрессии одного признака от другого согласно варианту индивидуального задания.

Методические указания

В экономических исследованиях одной из основных задач является анализ зависимостей между переменными.

Функциональная зависимость задается в виде точной формулы, в которой каждому значению одной переменной соответствует строго определенное значение другой; воздействием случайных факторов при этом пренебрегают.

Статистической зависимостью называется связь переменных, на которую накладывается воздействие случайных факторов.

Уравнение регрессии – это формула статистической связи между переменными. Формула статистической связи *двух* переменных называется парной регрессией, зависимость от *нескольких* переменных – множественной регрессией.

В *модели парной линейной регрессии* зависимость между переменными в генеральной совокупности представляется в виде

$$Y = \alpha + \beta \cdot X + \varepsilon, \tag{17}$$

где X – неслучайная величина;

Y, ε – случайные величины.

Величина Y называется объясняемой (зависимой) переменной, а X – объясняющей (независимой) переменной. Постоянные α, β – параметры уравнения.

На основе выборочного наблюдения оценивается выборочное уравнение регрессии

$$\hat{y} = a + bx, \tag{18}$$

где a, b – оценки параметров (α, β).

Величина $\hat{y}_i$ описывается как расчетное значение переменной y_i, соответствующее x_i.

Остаток e_i в i-м наблюдении определяется как разность между фактическим и расчетным значениями зависимой переменной, т. е.

$$e_i = y_i - \hat{y}_i. \tag{19}$$

Неизвестные значения a, b определяются *методом наименьших квадратов* (МНК):

$$b = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - (\bar{x})^2}; \quad a = \bar{y} - b\bar{x}. \tag{20}$$

Коэффициент b есть *угловой коэффициент регрессии*, он показывает, на сколько единиц в среднем изменяется переменная y при увеличении независимой переменной x на единицу.

Постоянная a дает *прогнозируемое* значение зависимой переменной при $x = 0$.

Коэффициентом детерминации R^2 называется отношение

$$R^2 = \frac{\mathrm{var}(\hat{y})}{\mathrm{var}(y)} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}, \tag{21}$$

причем $0 \le R^2 \le 1$, характеризующее долю вариации (разброса) зависимой переменной, *объясненную* с помощью уравнения регрессии.

Если $R^2 = 1$, то подгонка точная, т. е. все точки наблюдения лежат на одной

прямой. Если $R^2 = 0$, регрессия ничего не дает, т. е. переменная x не улучшает качества предсказания y по сравнению с горизонтальной прямой $\hat{y} = \bar{y}$. Чем ближе к единице R^2, тем лучше качество подгонки, т. е. $\hat{y}$ более точно аппроксимирует y.

Контрольные вопросы

1 Что изучает регрессионный анализ?
2 Что понимается под парной регрессией?
3 Какие задачи решаются при построении уравнения регрессии?
4 Какие методы применяются для выбора вида модели регрессии?
5 Какие функции чаще всего используются для построения уравнения парной регрессии?
6 Какой вид имеет уравнение парной линейной регрессии?
7 В чем суть метода наименьших квадратов (МНК)?
8 Какой смысл может иметь свободный член в парной линейной регрессии?
9 Как определяется коэффициент детерминации и каков его статистический смысл?
10 Как оценить значимость уравнения регрессии в целом?

5 Лабораторная работа № 5. Линейная множественная регрессия

Цель работы: научиться оценивать линейное уравнение множественной регрессии и проверять его качество.

Задание к лабораторной работе.
Построить уравнение множественной регрессии согласно варианту индивидуального задания. Пояснить смысл параметров уравнения.

Методические указания

Линейная модель множественной регрессии имеет вид:

$$Y_i = \alpha_0 + \alpha_1 \cdot x_{i1} + \alpha_2 \cdot x_{i2} + \alpha_m \cdot x_{im} + ... + \varepsilon_i. \quad (22)$$

Коэффициент регрессии α_j показывает, на какую величину в среднем изменится результативный признак Y, если переменную x_j увеличить на единицу измерения, т. е. α_j является нормативным коэффициентом.

Анализ уравнения множественной регрессии и методика определения параметров становиться более наглядными, а расчетные процедуры существенно упрощаются, если воспользоваться матричной формой записи уравнения

$$Y = X \cdot \alpha + \varepsilon, \tag{23}$$

где Y – вектор зависимой переменной размерности $n \times 1$, представляющий собой n наблюдений значений y_i;

X – матрица n наблюдений независимых переменных X_1, X_2, X_3, …, X_m, размерность матрицы X равна $n \times (m+1)$;

α – подлежащий оцениванию вектор неизвестных параметров размерности $(m+1) \times 1$;

ε – вектор случайных отклонений (возмущений) размерности $n \times 1$.

Таким образом,

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}; \quad X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1m} \\ 1 & x_{21} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{mn} \end{bmatrix}; \quad \alpha = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \dots \\ \alpha_m \end{bmatrix}.$$

Уравнение множественной регрессии содержит значения неизвестных параметров α_0, α_1, α_2, …, α_m. Эти величины оцениваются на основе выборочных наблюдений, поэтому полученные расчетные показатели представляют собой статистические оценки. Модель линейной регрессии, в которой вместо истинных значений параметров подставлены их оценки, имеет вид:

$$Y = Xa + e = \hat{Y} + e, \tag{24}$$

где α – вектор оценок параметров;

e – вектор «оцененных» отклонений регрессии, остатки регрессии $e = Y - Xa$;

$\hat{Y}$ – оценка значений Y, равная Xa.

Оценка параметров модели множественной регрессии с помощью МНК.

Формула для вычисления параметров регрессионного уравнения:

$$A = (X^T \cdot X)^{-1} X^T \cdot Y, \tag{25}$$

где X^T – транспонированная матрица X;

$(X^T \cdot X)^{-1}$ – обратная матрица $(X^T \cdot X)$.

Одним из условий регрессионной модели является предположение о линейной независимости объясняющих переменных. Линейная или близкая к ней связь между факторами называется *мультиколлинеарностью* и приводит к линейной зависимости нормальных уравнений, что делает вычисление параметров либо невозможным, либо затрудняет содержательную интерпретацию параметров модели. Чтобы избавиться от мультиколлинеарности, в модель включают лишь один из линейно связанных между собой факторов, причем тот, который в большей

степени связан с зависимой переменной.

В качестве критерия мультиколлинеарности может быть принято соблюдение следующих неравенств: $r_{yxi} > r_{xixk}$, $r_{yxk} > r_{xixk}$, $r_{xixk} < 0{,}8$, если приведенные неравенства (или хотя бы одно из них) не выполняется, то в модель включают тот фактор, который наиболее тесно связан с *Y*.

Проверка значимости модели регрессии.

Для проверки значимости модели регрессии используется *F*-критерий Фишера.

Контрольные вопросы

1 Что понимается под множественной регрессией?
2 Какие задачи решаются при построении уравнения регрессии?
3 Какие задачи решаются при спецификации модели?
4 Что понимается под ошибкой спецификации?
5 Какие требования предъявляются к факторам, включаемым в уравнение регрессии?
6 Как интерпретируются коэффициенты уравнения множественной регрессии?
7 Что понимается под коллинеарностью и мультиколлинеарностью факторов?
8 Как проверяется наличие коллинеарности и мультиколлинеарности?
9 Какие подходы применяются для преодоления межфакторной корреляции?
10 Что показывает значение коэффициента (индекса) множественной корреляции?
11 Как проверяется значимость уравнения регрессии и отдельных коэффициентов?
12 Опишите процедуру метода исключения переменных с использованием частных коэффициентов корреляции.

6 Лабораторная работа № 6. Подбор адекватной нелинейной зависимости

Цель работы: научиться подбирать лучшую из возможных нелинейных зависимостей в парной регрессии.

Задание к лабораторной работе.

Методом нелинейных преобразований исходных данных в варианте индивидуального задания подобрать лучшую из возможных нелинейных зависимостей в парной регрессии.

Методические указания

Если между экономическими явлениями существуют нелинейные соотношения, то они выражаются с помощью соответствующих нелинейных функций.

Уравнение линейной регрессии имеет вид:

$$\hat{y} = a + b \cdot x, \tag{26}$$

где a, b – оценки параметров (α, β), параметры (коэффициенты) регрессии.

Уравнение степенной модели имеет вид:

$$\hat{y} = a \cdot x^b. \tag{27}$$

Для построения этой модели необходимо произвести линеаризацию переменных. Для этого нужно произвести логарифмирование обеих частей уравнения.

Уравнение показательной кривой

$$\hat{y} = a \cdot b^x. \tag{28}$$

Для построения этой модели необходимо произвести линеаризацию переменных. Для этого нужно произвести логарифмирование обеих частей уравнения.

Уравнение гиперболической функции

$$\hat{y} = a + \frac{b}{x}. \tag{29}$$

Для построения этой модели нужно произвести линеаризацию путем замены $X = 1/x$.

Расчет средней относительной ошибки осуществляется по формуле

$$\overline{E}_{отн} = \frac{1}{n}\sum_{i=1}^{n}\frac{|e_i|}{y} \cdot 100\ \% = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y - \hat{y}}{y}\right| \cdot 100\ \%. \tag{30}$$

Контрольные вопросы

1 Перечислите основные виды нелинейных моделей в корреляционно-регрессионном анализе. Какова общая формула для нахождения коэффициента эластичности?

2 Как классический МНК применяется к нелинейным моделям регрессии?

3 Опишите метод линеаризации для сведения нелинейной модели к линейной для различных видов моделей.

7 Лабораторная работа № 7. Проверка выполнения предпосылок МНК: наличие и устранение гетероскедастичности и автокорреляции остатков

Цель работы: научиться оценивать наличие эффекта гетероскедастичности, автокорреляции и использовать взвешенный метод наименьших квадратов**.**

Задание к лабораторной работе.

Протестировать полученный в лабораторной работе № 5 вектор ошибок на нарушение предпосылок МНК.

Методические указания

Тест ранговой корреляции Спирмена.

Для обнаружения гетероскедастичности применяется тест ранговой корреляции Спирмена. Коэффициент ранговой корреляции рассчитывается по формуле

$$r_{x,\varepsilon} = 1 - 6 \cdot \frac{\sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)}, \tag{31}$$

где d_i – разность между рангами значений переменной X и модуля ошибки;

n – количество наблюдений.

$$d_i = R(x_i) - R(|\varepsilon_i|). \tag{32}$$

Проверка гипотезы осуществляется по критерию Стьюдента.

Тест Голдфельда-Квандта.

Для обнаружения гетероскедастичности выполняется тест Голдфельда-Квандта, включающий следующие этапы.

1 Вся совокупность наблюдений размерностью n упорядочивается по возрастанию значений фактора X (или $\hat{Y}$ для множественной регрессии).

2 Упорядоченная совокупность делится на три части размерностью k, $n - 2 \cdot k$, k соответственно; при этом k определяется из пропорции: при $n = 30$, $k = 11$.

3 Строятся отдельные уравнения регрессии для первой и третьей частей выборки и рассчитываются остаточные дисперсии для каждой из рассматриваемых частей.

4 Проверяется гипотеза о равенстве дисперсий двух совокупностей (большая дисперсия делится на меньшую) по формуле

$$F = \frac{S_б^2}{S_м^2}. \tag{33}$$

Критерий F имеет распределение Фишера с числами степеней свободы $\nu_1 = \nu_2 = k - m - 1$.

Тест Дарбина-Уотсона для обнаружения автокорреляции остатков.

Тест основан на вычислении DW-критерия по формуле

$$DW = \frac{\sum_{i=2}^{n} (\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^{n} \varepsilon_i^2}. \tag{34}$$

где ε_i – эмпирические случайные отклонения, упорядоченные по возрастанию значений фактора x_i (при множественной регрессии – по возрастанию прогнозируемых значений результирующего показателя $\hat{y}_i$).

Вычисленное значение DW-критерия сравнивается с критическими точками, dl и du, которые выбираются из таблицы Дарбина-Уотсона при заданном уровне значимости α в зависимости от числа наблюдений (объема выборки) n и числа факторов в регрессионной модели m (в парной регрессионной модели $m = 1$).

Вывод о наличии автокорреляции делается в зависимости от принадлежности критерия DW одному из интервалов:

– при $0 \le DW \le dl$ наблюдается положительная автокорреляция остатков;

– при $du \le DW \le (4 - du)$ автокорреляция остатков отсутствует;

– при $(4 - dl) \le DW \le 4$ наблюдается отрицательная автокорреляция остатков;

– оставшиеся интервалы являются областями неопределенности – однозначно сделать вывод о наличии или отсутствии автокорреляции остатков в этом случае невозможно.

При обнаружении автокорреляции остатков необходимо выяснить причины ее возникновения и предложить способ устранения.

Контрольные вопросы

1 Каково среднее значение случайного отклонения при выполнении предпосылок МНК?

2 Что такое гомоскедастичность и гетероскедастичность?

3 Что такое автокорреляция случайных отклонений?

4 Что означает несмещенность оценок параметров уравнения регрессии и их эффективность?

8 Лабораторная работа № 8. Построение моделей переменной структуры

Цель работы: научиться строить регрессионные модели с фиктивными переменными, а также модели переменной структуры.

Задание к лабораторной работе.

По данным индивидуального задания построить регрессионные модели для различных подынтервалов (подвыборок). Проверить гипотезу о структурной стабильности тенденции.

Методические указания

В практике эконометрики нередки случаи, когда имеются две выборки пар значений зависимой и объясняющих переменных x_i, y_i. Например, одна выборка пар значений переменных объемом n_1 получена при одних условиях, а другая, объемом n_2 – при несколько измененных условиях. Необходимо выяснить, действительно ли две выборки однородны в регрессионном смысле? Другими словами, можно ли *объединить* две выборки в одну и рассматривать единую модель регрессии Y по X?

При достаточных объемах выборок можно было, например, построить интервальные оценки параметров регрессии по каждой из выборок и в случае пересечения соответствующих доверительных интервалов сделать вывод о единой модели регрессии. Возможны и другие подходы.

В случае, если объем хотя бы одной из выборок незначителен, то возможности такого (и аналогичных) подхода резко сужаются из-за невозможности построения сколько-нибудь надежных оценок.

В *критерии* (*тесте*) *Грегори Чоу* эти трудности в существенной степени преодолеваются. По каждой выборке строятся две линейные регрессионные модели. Введем систему обозначений, приведенную в таблице 1.

Таблица 1 – Условные обозначения для алгоритма теста Чоу

Номер уравнения	Вид уравнения	Число наблюдений в совокупности	Остаточная сумма квадратов	Число параметров в уравнении	Число степеней свободы остаточной дисперсии
Кусочно-линейная модель					
1	$y^{(1)} = a_1 + b_1 x$	n_1	$C^1_{ост}$	k_1	$n_1 - k_1$
2	$y^{(2)} = a_2 + b_2 x$	n_2	$C^2_{ост}$	k_2	$n_2 - k_2$
Уравнение тренда по всей совокупности					
3	$y^{(3)} = a_3 + b_3 x$	n	$C^3_{ост}$	k_3	$n - k_3 = (n_1 + n_2) - k_3$

Выдвинем гипотезу H_0 о структурной стабильности тенденции изучаемого

временного ряда.

Остаточную сумму квадратов по кусочно-линейной модели $C^{к\text{-}л}{}_{ост}$ можно найти как сумму $C^{1}{}_{ост}$ и $C^{2}{}_{ост}$:

$$C_{ост}^{к-л} = C_{ост}^{1} + C_{ост}^{2}. \quad (35)$$

Соответствующее ей число степеней свободы составит:

$$(n_1 - k_1) + (n_2 - k_2) = n - k_1 - k_2. \quad (36)$$

Тогда сокращение остаточной дисперсии при переходе от единого уравнения тренда к кусочно-линейной модели можно определить следующим образом:

$$\Delta C_{ост} = C_{ост}^{3} + C_{ост}^{к-л}. \quad (37)$$

Число степеней свободы, соответствующее $\Delta C_{ост}$ с учетом соотношения будет равно:

$$n - k_3 - (n - k_1 - k_2) = k_1 + k_2 - k_3. \quad (38)$$

Далее, в соответствии с предложенной Г. Чоу методикой, определяется фактическое значение F-критерия по следующим дисперсиям на одну степень свободы вариации:

$$F_{факт} = \frac{\Delta C_{ост} : (k_1 + k_2 - k_3)}{C_{ост}^{к-л} : (n - k_1 - k_2)}. \quad (39)$$

Найденное значение $F_{факт}$ сравнивают с табличным, полученным по таблицам распределения Фишера для уровня значимости α и числа степеней свободы $(k_1 + k_2 - k_3)$ и $(n - k_1 - k_2)$.

Если $F_{факт} > F_{табл}$, то гипотеза о структурной стабильности тенденции отклоняется, а влияние структурных изменений на динамику изучаемого показателя признают значимым. В этом случае моделирование тенденции временного ряда следует осуществлять с помощью кусочно-линейной модели. Если $F_{факт} < F_{табл}$, то нет оснований отклонять ноль-гипотезу о структурной стабильности тенденции. Ее моделирование следует осуществлять с помощью единого для всей совокупности уравнения тренда.

Критерий Г. Чоу может быть использован при построении регрессионных моделей при воздействии *качественных* признаков, когда имеется возможность разделения совокупности наблюдений по степени воздействия этого фактора на отдельные группы и требуется установить возможность использования единой модели регрессии.

Оценивание регрессии с использованием *фиктивных переменных* более ин-

формативно в том отношении, что позволяет использовать *t*-критерий для оценки существенности влияния каждой фиктивной переменной на зависимую переменную.

Контрольные вопросы

1 В каком случае возникает необходимость вводить фиктивную переменную?

2 Как вводится фиктивная переменная, если у признака больше двух альтернатив?

3 Как интерпретируются коэффициенты при фиктивных переменных?

4 Когда используется тест Чоу и как он связан с применением фиктивных переменных?

9 Лабораторная работа № 9. Кластерный анализ

Цель работы: научиться выполнять группировку объектов, характеризующихся несколькими признаками.

Задание к лабораторной работе.

Решить задачу кластерного анализа согласно индивидуальному заданию.

Методические указания

Методы кластерного анализа позволяют выделить из исследуемой совокупности объектов *кластеры* – скопления объектов с близкими значениями параметров. Одной из проблем кластерного анализа является вычисление схожести объектов – обычно с этой целью применяются меры близости или расстояния в геометрическом смысле. Другой существенной проблемой является способ определения расстояния между кластерами, причём использование разных способов в одной и той же алгоритмической процедуре может привести к различным результатам. Для сравнения между собой различных кластерных решений используются т. н. *критерии качества*, основанные на подсчёте межкластерных и внутрикластерных расстояний. После выбора наилучшего решения полученные кластеры необходимо проинтерпретировать, исходя из средних значений параметров объектов, входящих в эти кластеры. *Интерпретация кластеров*, подобно интерпретации факторов, зависит от опыта и навыков исследователя и может давать неоднозначные конечные результаты.

Методы кластерного анализа можно разделить на две большие категории по алгоритму действия. Первая группа методов называется *иерархическими*, т. к. в процессе работы метода строится иерархия вложенности кластеров, обычно представляемая на графике – *дендрограмме*. На каждом шаге агломеративной иерархической процедуры объединяется пара ближайших кластеров. Методы второй категории называются *итерационными*, т. к. они основаны на поиске оптимального положения центров кластеров на каждой итерации – последователь-

ного рассмотрения всех объектов исходной выборки. Иерархические методы применяются для выборок небольшого объёма, т. к. их вычислительная эффективность резко снижается при увеличении числа объектов. Большинство итерационных методов зависит от значений некоторых параметров, например, предполагаемого числа кластеров, и хотя вычислительная эффективность позволяет обрабатывать большие выборки, вплоть до нескольких тысяч объектов, качество решений в некоторых случаях оказывается неудовлетворительным. Поэтому для получения заданного качества приходится применять такие методы несколько раз при различных значениях параметров.

В *иерархической агломеративной процедуре* на каждом шаге вычисляется матрица расстояния между всеми парами объектов и кластеров, если они уже были построены. По матрице расстояний находится пара ближайших кластеров, которые объединяются в кластер. Этот процесс продолжается до тех пор, пока все объекты не сольются в один кластер. Оптимальное число кластеров определяется по скачку *расстояния агломерации*, где под скачком подразумевается превышение расстояния на текущем шаге процедуры предыдущего расстояния в 1,5–2 раза. На практике такой скачок достигается редко, приходится иметь дело с превышением на 50 %...60 %. В случае нахождения шага r скачка оптимальное число кластеров N_k определяется по формуле

$$N_k = N - r + 1. \tag{39}$$

Чаще всего близость объектов i и j измеряется с помощью следующих метрик расстояния, если их характеристики измерены в интервальной шкале:

– дистанция Евклида:

$$d_{ij} = \sqrt{\sum_{g=1}^{l} \left(x_{ig} - x_{jg}\right)^2}, \tag{40}$$

где d_{ij} – расстояние между i-м и j-м объектами;

l – количество признаков;

x_{ig} – значение i-го признака g-го объекта;

x_{jg} – значение j-го признака g-го объекта;

– квадрат дистанции Евклида (для придания больших весов более отдаленным друг от друга объектам):

$$d_{ij} = \sum_{g=1}^{l} \left(x_{ig} - y_{jg}\right)^2; \tag{41}$$

– расстояние Чебышева. Это расстояние следует использовать, когда необходимо определить два объекта как «различные», если они сильно отличаются по какому-то одному измерению:

$$d_{ij} = \max\left|x_{ig} - x_{jg}\right|, i = \overline{1,k}, j = \overline{1,k}, \quad (42)$$

где k – количество объектов;

– расстояние Манхэттена (расстояние городских кварталов), также называемое «хеммингловым»

$$d_{ij} = \sum_{g=1}^{l} \left|x_{ig} - x_{ig}\right|. \quad (43)$$

В этом случае просто берутся абсолютные значения покоординатных расстояний и суммируются. Свое интересное название эта метрика получила из-за того, что моделирует расстояние, пройденное человеком в городе, когда перемещаться можно только по улицам, и нельзя, например, пересечь квартал по диагонали. Аналогия в декартовой плоскости приводит к перемещениям только по линиям, параллельным осям координат, и, соответственно, к манхэттенскому расстоянию. Для данного вида расстояния влияние имеющихся «выбросов» (больших отклонений) меньше, чем при использовании евклидова расстояния, поскольку в этом случае координаты не возводятся в квадрат;

Расстояние между кластерами определяется с помощью следующих основных методов:

– связь между группами – расстояние между двумя кластерами определяется как среднее значение расстояний между всеми возможными парами наблюдений, причём одно наблюдение берётся из одного кластера, а второе – из другого;

– связь внутри групп – расстояние между двумя кластерами определяется как среднее значение расстояний между всеми возможными парами наблюдений из обоих кластеров, включая пары наблюдений внутри кластеров;

– ближний сосед – расстояние между двумя кластерами определяется как минимальное из всех расстояний между всеми возможными парами наблюдений из разных кластеров;

– дальний сосед – расстояние между двумя кластерами определяется как максимальное из всех расстояний между всеми возможными парами наблюдений из разных кластеров;

– центроидная кластеризация – расстояние между двумя кластерами определяется как расстояние между центрами тяжести обоих кластеров;

– медианная кластеризация – расстояние между двумя кластерами определяется как взвешенное центроидное расстояние между кластерами, где веса соответствуют размеру каждого кластера;

– метод Варда – в этом методе объединяются только те два кластера, для которых прирост внутрикластерной дисперсии минимален.

Наиболее универсальными методами являются метод Варда и метод межгрупповой связи.

Контрольные вопросы

1 Что называется кластером?
2 Для чего используется кластерный анализ?
3 Результат кластерного анализа.
4 Назовите способы отображения результатов кластеризации.
5 Охарактеризуйте методы оценки схожести объектов.

10 Лабораторная работа № 10. Дискриминантный анализ

Цель работы: изучение основных процедур дискриминантного анализа: дискриминации и классификации, построение и определение количества дискриминантных функций и их разделительной способности.

Задание к лабораторной работе.
По данным индивидуального задания выполнить дискриминантный анализ.

Методические указания

Дискриминантный анализ используется для анализа данных в том случае, когда зависимая переменная категориальная, а предикторы (независимые переменные) интервальные. Дискриминантный анализ используется для принятия решения о том, какие переменные различают (дискриминируют) две или более возникающие совокупности (группы).

Дискриминантный анализ преследует следующие цели.

1 Определение дискриминантных функций (*discriminant functions*) или линейных комбинаций независимых переменных, которые наилучшим образом различают (дискриминируют) категории (группы) зависимой переменной.

2 Проверка существования между группами значимых различий с точки зрения независимых переменных.

3 Определение предикторов, вносящих наибольший вклад в межгрупповые различия.

4 Отнесение случаев к одной из групп (классификация), исходя из значений предикторов.

5 Оценка точности классификации данных на группы.

Метод дискриминантного анализа описывается числом категорий, имеющихся у зависимой переменной. Если она имеет две категории, то метод называют дискриминантным анализом для двух групп (*two-group discriminant analsysis*).

Если анализируют три или больше категорий, то метод называют множественным дискриминантным анализом (*multiple descriminant analysis*).

Главное отличие между ними заключается в том, что при наличии двух групп возможно вывести только одну дискриминантную функцию. Используя множественный дискриминантный анализ, можно вычислить несколько функций.

Модель дискриминантного анализа имеет следующий вид:

$$D = a_0 + a_1 \cdot x_1 + \ldots + a_k \cdot x_k, \quad (44)$$

где D – дискриминантный показатель (дискриминант);

a – дискриминантный коэффициент или вес;

x – предиктор или независимая переменная.

Дискриминантные переменные используются для того, чтобы отличать один класс (подмножество) от другого.

Коэффициенты или веса a определяют таким образом, чтобы группы максимально возможно отличались значениями дискриминантной функции. Это происходит тогда, когда отношение межгрупповой суммы квадратов к внутри групповой сумме квадратов для дискриминантных показателей максимально. Любая другая линейная комбинация предикторов приводит к меньшему значению этого отношения.

Контрольные вопросы

1 Для чего используется дискриминантный анализ?
2 Что такое дискриминантная функция?
3 Как определяются коэффициенты дискриминантной функции?
4 В чем заключается отличие дискриминантного от кластерного анализа?

Список литературы

1 **Басовский, Л. Е.** Эконометрика: учебное пособие / Л. Е. Басовский. – Москва: РИОР; ИНФРА-М, 2017. – 48 с.

2 **Горелов, Н. А.** Методология научных исследований : учебник и практикум для бакалавриата и магистратуры / Н. А. Горелов, Д. В. Круглов, О. Н. Кораблева. – 2-е изд., перераб. и доп. – Москва: Юрайт, 2017. – 365 с.

3 **Катаргин, Н. В.** Экономико-математическое моделирование: учебное пособие / Н. В. Катаргин. – Санкт-Петербург; Москва; Краснодар: Лань, 2018. – 256 с.

4 **Ниворожкина, Л. И.** Эконометрика: теория и практика: учебное пособие / Л. И. Ниворожкина, С. В. Арженовский, Е. П. Кокина. – Москва: РИОР; ИНФРА-М, 2018. – 207 с.

5 **Смирнов, В. А.** Математическое моделирование в машиностроении в примерах и задачах: учебное пособие / В. А. Смирнов. – Старый Оскол: ТНТ, 2019. – 364 с.