Министерство образования Республики Беларусь

Учреждение образования
«Гомельский государственный университет
имени Франциска Скорины»

Л. Н. Марченко,
Ю. Е. Дудовская,
Ю. В. Синюгина

ЭКОНОМЕТРИКА:

МОДЕЛЬ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

Практическое руководство

для студентов специальности
1-31 03 06-01 Экономическая кибернетика
(математические методы
и компьютерное моделирование в экономике)

Гомель
ГГУ им. Ф. Скорины
2016

УДК 330.43
ББК 65 я 631
Э40

Рецензенты:
кандидат физико-математических наук Ю. С. Крук
кандидат физико-математических наук В. В. Подгорная

Рекомендовано к изданию научно-методическим советом учреждения образования «Гомельский государственный университет имени Франциска Скорины»

Э40 **Эконометрика: модель множественной линейной регрессии** : практическое пособие / Л. Н. Марченко, Ю. Е. Дудовская, Ю. В. Синюгина ; М-во образования Республики Беларусь, Гомельский гос. ун-т им. Ф. Скорины. – Гомель : ГГУ им. Ф. Скорины, 2016. – 48 с.
ISBN 978-985-577-228-7

В практическом руководстве излагаются краткие теоретические сведения, решения типовых примеров, задания для лабораторных работ по дисциплине «Эконометрика». Рассматриваются вопросы, связанные с построением, анализом и корректировкой модели множественной линейной регрессии. Материал, представленный в доступной форме, позволяет студентам быстро освоить основные положения дисциплины по разделу «Модель множественной линейной регрессии».

Практическое руководство предназначено для студентов дневной формы обучения специальности 1-31 03 06-01 Экономическая кибернетика (математические методы и компьютерное моделирование в экономике).

УДК 330.43
ББК 65 я 631

ISBN 978-985-577-228-7

ОГЛАВЛЕНИЕ

Предисловие

Эконометрика представляет собой научную область на стыке экономической и математической наук, в рамках которой на основе установленных экономической теорией зависимостей между экономическими переменными с помощью статистических методов анализа реальных данных осуществляется разработка адекватных моделей исследуемых процессов.

Практическое руководство «Эконометрика: модель множественной линейной регрессии» предназначено для студентов специальности 1 31 03 06-01 Экономическая кибернетика (математические методы и компьютерное моделирование в экономике).

Темы практического руководства имеют идентичную структуру: краткие теоретические сведения, решение типовых примеров, задания для лабораторных работ, что позволяет использовать его как для проведения лабораторных занятий, так и для организации самостоятельной учебной работы студентов.

Практическое руководство является основой для подготовки и выполнения лабораторных работ, связанных с построением, анализом и корректировкой модели множественной линейной регрессии. Особенностью данного руководства является то, что вычислительные процедуры осуществляются детально с использованием в качестве инструмента расчетов MS Excel 2010. Такой подход позволяет «прочувствовать» процесс построения и анализа моделей, приобрести определенный опыт, способствующий в дальнейшем более глубокому пониманию и осмыслению результатов, получаемых при работе с эконометрическими пакетами прикладных программ.

Данное пособие не заменяет классических учебников по эконометрике, а лишь выступает в качестве руководства к выполнению лабораторных работ, а также служит для самостоятельного изучения отдельных вопросов дисциплины. В практическом руководстве предлагается список литературы для более глубокого изучения материала.

Авторы с благодарностью воспримут все критические замечания и пожелания, а также указания на возможные опечатки.

1 Модель множественной линейной регрессии

1.1 Краткие теоретические сведения

1.1.1 Модель множественной линейной регрессии в предположениях Гаусса-Маркова

Регрессионные модели применяются для исследования зависимости среднего значения переменной Y от объясняющих переменных (факторов) X_1, X_2, …, X_m. Модель множественной линейной регрессии (the multiple linear regression model) описывается соотношением

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots + \beta_m x_{tm} + \xi_t, \qquad (1.1)$$

где y_t – значение зависимой (эндогенной) переменной Y в наблюдении t;

x_{t1}, x_{t2}, …, x_{tm} – значения независимых (объясняющих, экзогенных) переменных X_1, X_2, …, X_m в наблюдении t;

ξ_t – случайная ошибка в наблюдении t, $t = 1, \ldots, n$.

Коэффициенты модели $\beta_0, \beta_1, \ldots, \beta_m$ называются *параметрами* модели (1.1), считаются неизвестными и подлежат оцениванию.

При $m = 1$ модель $y_t = \beta_0 + \beta_1 x_t$ называется *моделью парной линейной регрессии* Y на X.

Модель множественной линейной регрессии (1.1) в матричной форме имеет вид

$$Y = X \cdot \beta + \xi. \qquad (1.2)$$

Здесь $Y = (y_1\ y_2\ \ldots\ y_n)^T$ – вектор-столбец наблюдаемых значений зависимой переменной Y размерности $n \times 1$;

$X = \begin{bmatrix} 1\, x_{11}\ x_{12}\ \ldots\ x_{1m} \\ 1\, x_{21}\ x_{22}\ \ldots\ x_{2m} \\ \ldots \\ 1\, x_{n1}\ x_{n2}\ \ldots\ x_{nm} \end{bmatrix}$ – матрица наблюдаемых значений, объясняющих переменных размерности $n \times (m + 1)$, 1-й столбец которой является единичным, так как в модели регрессии (1.1) коэффициент β_0 умножается на 1;

$\beta = (\beta_0, \beta_1, \ldots, \beta_m)^T$ – вектор-столбец неизвестных параметров (коэффициентов) модели (1.1) размерности $(m + 1) \times 1$;

$\xi = (\xi_0, \xi_1, \ldots, \xi_m)^T$ – вектор-столбец случайных ошибок модели (1.1) размерности $n \times 1$.

Ожидаемое в соответствии с моделью (1.1) значение зависимой

переменной Y определяется как условное математическое ожидание случайной величины y_t при условии, что вектор объясняющих переменных $(X_1, X_2, \ldots, X_m)$ принимает фиксированное значение $x_t = (x_{t1}\ x_{t2}\ \ldots\ x_{tm})$:

$$M(y_t \mid x_t) = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots + \beta_m x_{tm},\ t = 1, \ldots, n.$$

Случайные величины $\{\xi_t\}$ характеризуют отклонение наблюдаемых значений зависимой переменной Y от ее ожидаемых значений в соответствии с моделью (1.1) и обусловлены действием неучтенных в модели случайных «нерегулярных» факторов.

Модель множественной линейной регрессии называется *нормальной*, если выполняются предпосылки Гаусса-Маркова.

1 Предположения относительно объясняющих переменных:

X.1 переменные $X_1, X_2, \ldots, X_m$ являются детерминированными;

X.2 число наблюдений n удовлетворяет условию $n > m + 1$;

X.3 матрица X является матрицей полного ранга $\text{rank}(X) = m + 1$, то есть столбцы матрицы X являются линейно независимыми векторами.

2 Предположения относительно случайных ошибок наблюдений:

ξ.1 математическое ожидание случайных ошибок $\{\xi_t\}$ равно нулю для всех наблюдений, то есть $M(\xi_t) = 0,\ t = 1, \ldots, n$;

ξ.2 (гомоскедастичность) дисперсия случайных ошибок $\{\xi_t\}$ является постоянной для всех наблюдений, то есть $M(\xi_t^2) = D(\xi_t) = \sigma^2,\ t = 1, \ldots, n$;

ξ.3 случайные ошибки некоррелированы для разных наблюдений

$$\text{cov}(\xi_t, \xi_s) = 0,\ t \neq s,\ t, s = 1, \ldots, n;$$

ξ.4 случайные ошибки $\{\xi_t\}$ имеют совместное нормальное распределение с нулевым математическим ожиданием и дисперсией σ^2, то есть $\xi_t \sim N(0,\sigma^2),\ t = 1, \ldots, n$.

1.1.2 Построение и оценка качества модели множественной линейной регрессии

Процедура построения и анализа качества модели множественной линейной регрессии включает следующие этапы.

Этап 1. Оценка неизвестных параметров модели. Оценка неизвестных параметров модели $\beta_0, \beta_1, \ldots, \beta_m$ при выполнении предпосылок Гаусса-Маркова осуществляется с помощью метода наименьших квадратов (МНК). МНК-оценка $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m)^T$ вектора $\beta = (\beta_0, \beta_1, \ldots, \beta_m)$ находится из условия минимума суммы квадратов отклонений наблюдаемых (фактических) значений y_t зависимой переменной Y от модельных (теоретических) значений $\hat{y}_t$, определяемых согласно (1.1):

$$\sum_{t=1}^{n} (y_t - \hat{y}_t)^2 \rightarrow \min,$$

и имеет вид

$$\hat{\beta} = (X^T X)^{-1} X^T Y. \qquad (1.3)$$

Оценка $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m)^T$ при выполнении предпосылок Гаусса-Маркова является несмещенной, эффективной и состоятельной.

Тогда оцененная модель запишется в виде

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_{t1} + \ldots + \hat{\beta}_m x_{tm}. \qquad (1.4)$$

Величины $e_t = y_t - \hat{y}_t$, $t = 1, \ldots, n$, называются *остатками* модели.

МНК-оценкой дисперсии $\sigma^2(\xi)$ случайной ошибки ξ является величина

$$S^2 = \frac{1}{n-m-1} \sum_{t=1}^{n} e_t^2 . \qquad (1.5)$$

Стандартной ошибкой регрессии называется величина

$$S = \sqrt{\frac{1}{n-m-1} \sum_{t=1}^{n} e_t^2} . \qquad (1.6)$$

Стандартной ошибкой оценки $\hat{\beta}_i$ называется величина

$$S_{\hat{\beta}_i} = S \sqrt{\left(X^T X \right)^{-1}_{ii}} , \; i = 0, \; 1, \; \ldots, \; m , \qquad (1.7)$$

где $(X^T X)^{-1}_{ii}$ – i-й диагональный элемент матрицы $(X^T X)^{-1}$.

Пусть $RSS = \sum_{t=1}^{n} (\hat{y}_t - \overline{Y})^2$ – сумма квадратов отклонений (regression sum of squares), обусловленная включенными в модель переменными $X_1, X_2, \ldots, X_m$,

$ESS = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2$ – сумма квадратов остатков (error sum of squares),

$TSS = \sum_{t=1}^{n} (y_t - \overline{Y})^2$ – полная сумма квадратов отклонений (total sum of squares) y_t от выборочного среднего значения $\overline{Y} = \frac{1}{n} \sum_{t=1}^{n} y_t$, при этом

$$TSS = ESS + RSS.$$

Коэффициент детерминации определяется выражением

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS} \qquad (1.8)$$

и показывает долю вариации зависимой переменной Y, обусловленную включенными в модель объясняющими переменными $X_1, X_2, \ldots, X_m$.

Коэффициент множественной корреляции равен $R = \sqrt{R^2}$.

Если $R^2 = 0$, то модель не улучшает качество предсказания y_t по сравнению с тривиальным $y_t = \overline{Y}$. Если $R^2 = 1$, то имеет место точная подгонка уравнения, то есть все $e_t = 0$.

На коэффициент R^2 нельзя ориентироваться как на главный критерий

при сравнении двух различных структур модели множественной линейной регрессии. Использовать R^2 целесообразно только совместно с дополнительным анализом модели, так как:

1) R^2 увеличивается при добавлении в модель объясняющих переменных;

2) при $m \to \infty$ значения R^2 есть «неправдоподобно» близкие к 1;

3) R^2 не может использоваться для выбора модели из набора альтернативных вариантов, получающихся при преобразовании переменной Y;

4) R^2 принимает значения близкие к 1 при построении регрессионных моделей по нестационарным временным рядам, что затрудняет распознавание ложных регрессионных зависимостей.

При сравнении альтернативных моделей множественной линейной регрессии, отличающихся количеством объясняющих переменных, целесообразно использовать *скорректированный* (adjusted) коэффициент детерминации:

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-m-1}. \qquad (1.9)$$

Свойства R^2_{adj}:

1) $R^2 \geq R^2_{adj}$ при $m > 1$;

2) $R^2_{adj} < 1$;

3) для некоторых моделей может быть $R^2_{adj} < 0$.

Коэффициенты R^2 и R^2_{adj} удобно рассчитывать с помощью дисперсионной таблицы 1.1.

Таблица 1.1 – Дисперсионная таблица

Источник вариации	Сумма квадратов (*SS*)	Число степеней свободы (*df*)	Дисперсия (*MS*)
Регрессия	$RSS = \sum_{t=1}^{n}(\hat{y}_t - \bar{Y})^2$	m	$MSR^2 = RSS/m$
Остатки	$ESS = \sum_{t=1}^{n}(y_t - \hat{y}_t)^2$	$n - (m + 1)$	$MSE^2 = ESS/(n-m-1)$
Полная вариация	$TSS = \sum_{t=1}^{n}(y_t - \bar{Y})^2$	$n - 1$	$MTS^2 = TSS/n$

Этап 2. Проверка гипотез о параметрах модели.

Проверка статистической значимости параметров модели. Возможна ситуация, когда часть оцененных коэффициентов модели не обладает необходимой степенью значимости. Тогда такие коэффициенты должны быть исключены из модели. Выдвигаются нулевые гипотезы о незначимости коэффициентов H_0: $\beta_i = 0$, $i = 0, 1, \ldots, m$, против альтернативных гипотез H_1: $\beta_i \neq 0$, $i = 0, 1, \ldots, m$.

Проверка гипотез осуществляется с помощью *t*-критерия со статистикой $t_i = \hat{\beta}_i / S_{\hat{\beta}_i}$, $i = 0, 1, \ldots, m$, имеющей асимптотически ($n \to \infty$) распределение Стьюдента (Student) с числом степеней свободы $df = n - m - 1$

$$t_i = \hat{\beta}_i / S_{\hat{\beta}_i} \sim T(n - m - 1). \qquad (1.10)$$

Гипотеза H_0 отклоняется на уровне значимости α, если $| t_i | \geq t_{кр}$ или выполняется неравенство $Pv = P\{T(n - m - 1) \geq t\} < \alpha$. Здесь t_i – статистика критерия для коэффициента β_i, $t_{кр} = t_{1-\alpha/2}(n - m - 1)$ – квантиль распределения Стьюдента. В этом случае оценка $\hat{\beta}_i$ значимо отличается от нуля, и, следовательно, X_i оказывает существенное влияние на Y, то есть существует статистическая линейная зависимость между X_i и Y. Если коэффициент β_i окажется незначимым, то X_i следует исключить из модели (при этом качество модели не ухудшится). Если же незначимым окажется коэффициент $\hat{\beta}_0$, то проводится пересчет оценок параметров модели в предположении $\beta_0 = 0$. Включение незначимой переменной в модель может определяться экономической целесообразностью.

Замечание. *p*-значение *Pv* (*p*-value) – это вероятность того, что случайная величина $K(\mathrm{df})$, имеющая заданное распределение со степенями свободы df принимает значение K, не меньшее, чем наблюдаемое значение статистики $Pv = P\{K(\mathrm{df}) \geq K\}$. Если $Pv < \alpha$, то гипотеза H_0 отклоняется, если $Pv > \alpha$, то H_0 не отклоняется. Если (в пределах округления) $Pv = \alpha$, то в отношении гипотезы H_0 можно принять любое из возможных решений.

Доверительный интервал, накрывающий с заданной вероятностью $P = 1 - \alpha$ истинное значение параметра β_i, $i = 0,1,\ldots, m$, имеет вид

$$(\hat{\beta}_i - S_{\hat{\beta}_i}\ t_{1-\alpha/2}(n - m - 1);\ \hat{\beta}_i + S_{\hat{\beta}_i}\ t_{1-\alpha/2}(n - m - 1)).$$

Проверка качества и значимости всего уравнения регрессии. Для проверки значимости модели проверяется гипотеза о значимости коэффициента детерминации $R^2 = 0$, которая означает, что все коэффициенты β_0, $\beta_1, \ldots, \beta_m$ одновременно равны нулю, то есть $\beta_0 = \beta_1 = \ldots = \beta_m = 0$.

На уровне значимости α выдвигается нулевая гипотеза H_0: $R^2 = 0$ против альтернативной H_1: $R^2 \neq 0$. Для проверки гипотезы используется *F*-критерий Фишера (Fisher) с $df_1 = m$ и $df_2 = n - m - 1$ степенями свободы и статистикой

$$F = \frac{R^2 / m}{(1 - R^2)/(n - m - 1)} \sim F\,(m, n - m - 1). \qquad (1.11)$$

Гипотеза H_0 отклоняется на уровне значимости α, если $F \geq F_{кр}$ или $Pv < \alpha$. Здесь $F_{кр} = F_{1-\alpha}(m, n - m - 1)$ – квантиль распределения Фишера, $Pv = P\{F(m, n - m - 1) \geq F\}$. Тогда среди объясняющих переменных есть

хотя бы одна, которая оказывает существенное влияние на зависимую переменную. Если гипотеза не отклоняется, то следуют признать модель неадекватной, то есть ни одна из переменных $X_1, X_2, \ldots, X_m$ не оказывает существенного влияния на Y.

Точечный и интервальный прогнозы по модели регрессии. Точечный прогноз есть значение $\hat{y}_{n+1}$, полученное постановкой в оцененную модель (1.4) вектора прогнозных значений $x_{n+1} = (1\ x_{(n+1)1}\ \ldots\ x_{(n+1)m})$:

$$\hat{y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{(n+1)1} + \ldots + \hat{\beta}_m x_{(n+1)m}.$$

Интервальный прогноз представляет собой интервал, в который с доверительной вероятностью $P = 1 - \alpha$ попадает фактическое значение переменной Y при заданных прогнозных значениях переменных $X_1, X_2, \ldots, X_m$. Доверительный интервал имеет вид

$$(\hat{y}_{n+1} - Se(y_{n+1})\, t_{1-\alpha/2}(n-m-1);\ \hat{y}_{n+1} + Se(y_{n+1})\, t_{1-\alpha/2}(n-m-1)),$$

где $Se(y_{n+1})$ – средняя ошибка прогноза, равная

$$Se(y_{n+1}) = S\sqrt{1 + x_{n+1}^T (X^T X)^{-1} x_{n+1}}\,. \qquad (1.12)$$

Этап 3. Оценка влияния объясняющих переменных на зависимую переменную на основе построенной модели. Разная степень влияния объясняющих переменных на зависимую переменную обусловлена различием их единиц измерения и разной степенью колеблемости.

Коэффициент эластичности $E_i = \hat{\beta}_i \overline{X}_i / \overline{Y}$, $i = 1,\ldots, m$, показывает, на сколько процентов изменяется Y при изменении X_i на 1 %. При этом коэффициент E_i не учитывает степень колеблемости остальных переменных.

Бета-коэффициент $\beta_i^* = \hat{\beta}_i S_{X_i} / S_Y$, $i = 1,\ldots, m$, показывает, на какую часть величины среднего квадратического отклонения изменится Y с изменением переменной X_i на величину своего среднеквадратического отклонения при фиксированных значениях остальных объясняющих переменных.

Дельта-коэффициент $\Delta_i = r_{YX_i} \beta_i^* / R^2$, $i = 1, \ldots, m$, показывает долю влияния X_i в суммарном влиянии всех переменных $X_1, X_2, \ldots, X_m$. Для нормальной модели все дельта-коэффициенты имеют положительные значения и их сумма равна 1. При достаточно сильной корреляции между объясняющими переменными некоторые дельта-коэффициенты могут оказаться отрицательными вследствие того, что соответствующий коэффициент модели имеет знак, противоположный парному коэффициенту корреляции этой переменной с Y. Интерпретация отрицательных дельта-коэффициентов лишена смысла, и при этом также искажаются выводы по дельта-коэффициентам других факторов.

Здесь

$$\overline{X}_i = \frac{1}{n}\sum_{t=1}^{n} x_{ti}\,,\ S_{X_i}^2 = \frac{1}{n-1}\sum_{t=1}^{n}(x_{ti} - \overline{X}_i)^2\,,$$

$$S_Y^2 = \frac{1}{n-1}\sum_{t=1}^{n}(y_t - \overline{Y})^2\,,\ r_{YX_i} = \frac{1}{n}\sum_{t=1}^{n}(x_{ti} - \overline{X}_i)(y_t - \overline{Y})\,,\ i = 1, .., m.$$

В качестве меры точности модели можно использовать показатели:

1) средняя абсолютная ошибка

$$\varepsilon = \frac{1}{n}\sum_{t=1}^{n} |\, e_t \,|, \tag{1.13}$$

которая показывает, насколько в среднем отклоняются фактические значения y_t от модельных $\hat{y}_t$;

2) стандартная ошибка модели

$$S = \sqrt{\frac{1}{n-m-1}\sum_{t=1}^{n} e_t^2}\,, \tag{1.14}$$

для которой чем меньше S, тем точнее модель;

3) средняя относительная ошибка аппроксимации

$$\overline{\varepsilon} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{e_t}{y_t}\right| \cdot 100\ \%, \tag{1.15}$$

допустимый предел которой составляет 8 – 15 %.

1.2 Решение типового примера

Пример 1.1 В таблице 1.2 приводятся статистические данные величины инфляции (Y), ВВП (X_1), денежной массы (X_2), импорта товаров и услуг в процентах от ВВП (X_3), индекса реального эффективного обменного курса (X_4), процентной ставки по кредитам (X_5) различных стран. Необходимо:

1 построить модель множественной линейной регрессии;

2 проверить значимость коэффициентов модели;

3 оценить качество и значимость модели;

4 оценить влияние объясняющих переменных на зависимую;

5 построить точечный и интервальный прогнозы переменной Y.

Таблица 1.2 – Исходные данные (в %) к примеру 1.1.

t	Страна	y_t	Объясняющие переменные				
			x_{t1}	x_{t2}	x_{t3}	x_{t4}	x_{t5}
1	Австралия	2,488	2,500	7,058	21,398	1,001	5,950
2	Алжир	2,916	3,800	16,666	31,981	1,054	8,000
3	Армения	2,981	3,500	8,336	46,905	1,018	16,409
4	Беларусь	16,200	1,588	23,867	57,929	0,980	18,742
5	Бразилия	6,332	0,145	13,531	14,274	0,870	32,008
6	Грузия	3,069	4,766	13,757	60,395	1,069	11,910
7	Израиль	0,476	2,554	10,950	30,637	1,034	3,907
8	Исландия	2,035	1,827	8,726	47,366	1,128	7,743
9	Малайзия	3,143	5,993	6,914	64,609	0,999	4,587
10	Мексика	4,019	2,231	12,193	33,465	1,025	3,552
11	Молдова	5,089	4,600	5,082	78,016	1,044	11,013
12	Россия	7,826	0,640	15,455	22,871	0,992	11,143
13	Румыния	1,069	2,777	8,359	41,042	1,015	8,466
14	Украина	12,188	-6,800	5,251	53,240	0,783	17,718
15	Чили	4,395	1,894	8,743	32,294	0,939	8,098

Решение. *1 Построение модели множественной линейной регрессии.* Модель множественной линейной регрессии (1.1) принимает вид

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 x_{t3} + \beta_4 x_{t4} + \beta_5 x_{t5} + \xi_t,\ t = \overline{1,15}. \qquad (1.16)$$

Матрица значений объясняющих переменных равна

$$X = \begin{pmatrix} 1 & 2,500 & 7,058 & 21,398 & 1,001 & 5,950 \\ 1 & 3,800 & 16,666 & 31,981 & 1,054 & 8,000 \\ 1 & 3,500 & 8,336 & 46,905 & 1,018 & 16,409 \\ 1 & 1,588 & 23,867 & 57,929 & 0,980 & 18,742 \\ 1 & 0,145 & 13,531 & 14,274 & 0,870 & 32,008 \\ 1 & 4,766 & 13,757 & 60,395 & 1,069 & 11,910 \\ 1 & 2,554 & 10,950 & 30,637 & 1,034 & 3,907 \\ 1 & 1,827 & 8,726 & 47,366 & 1,128 & 7,743 \\ 1 & 5,993 & 6,914 & 64,609 & 0,999 & 4,587 \\ 1 & 2,231 & 12,193 & 33,465 & 1,025 & 3,552 \\ 1 & 4,600 & 5,082 & 78,016 & 1,044 & 11,013 \\ 1 & 0,640 & 15,455 & 22,871 & 0,992 & 11,143 \\ 1 & 2,777 & 8,359 & 41,042 & 1,015 & 8,466 \\ 1 & -6,800 & 5,251 & 53,240 & 0,783 & 17,718 \\ 1 & 1,894 & 8,743 & 32,294 & 0,939 & 8,098 \end{pmatrix}.$$

Матрица, обратная к матрице $X^T X$, имеет вид

$$\left(X^T X\right)^{-1} = \begin{pmatrix} 33{,}285 & 0{,}545 & 0{,}067 & -0{,}004 & -33{,}379 & -0{,}151 \\ 0{,}545 & 0{,}021 & 0{,}000 & 0{,}000 & 0{,}577 & 0{,}000 \\ 0{,}067 & 0{,}000 & 0{,}004 & 0{,}000 & -0{,}104 & -0{,}001 \\ -0{,}004 & 0{,}000 & 0{,}000 & 0{,}000 & -0{,}008 & 0{,}000 \\ -33{,}379 & -0{,}577 & -0{,}104 & -0{,}008 & 34{,}602 & 0{,}143 \\ -0{,}151 & 0{,}000 & -0{,}001 & 0{,}000 & 0{,}143 & 0{,}002 \end{pmatrix}. \quad (1.17)$$

МНК – оценка вектора коэффициентов модели равна

$$\hat{\beta} = (X^T X)^{-1} X^T Y = (17{,}541 \quad -0{,}582 \quad 0{,}545 \quad 0{,}120 \quad -22{,}466 \quad -0{,}003)^T.$$

Тогда оцененная модель принимает вид

$$y_t = 17{,}541 - 0{,}582x_{t1} + 0{,}545x_{t2} + 0{,}120x_{t3} - 22{,}466x_{t4} - 0{,}003x_{t5}. \quad (1.18)$$

Подставляя в (1.18) наблюдаемые значения переменных X_1, X_2, X_3, X_4, X_5, получим вектор модельных значений зависимой переменной Y

$$\hat{Y} = (-0{,}009 \; 4{,}543 \; 2{,}749 \; 14{,}498 \; 6{,}911 \; 5{,}449 \; 2{,}452 \; 1{,}543 \; 3{,}1 \; 3{,}859 \; 3{,}489 \; 6{,}017 \; 2{,}57 \; 13{,}1 \; 3{,}954)^T.$$

Тогда вектор остатков есть

$$e = (2{,}497 \; -1{,}627 \; 0{,}232 \; 1{,}702 \; -0{,}579 \; -2{,}38 \; -1{,}976 \; 0{,}492 \; 0{,}043 \; 0{,}16 \; 1{,}6 \; 1{,}809 \; -1{,}501 \; -0{,}912 \; 0{,}441)^T.$$

Стандартная ошибка регрессии равна $S = 1{,}859$; стандартные ошибки оценок коэффициентов $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\beta}_4$, $\hat{\beta}_5$ равны соответственно

$$S_{\hat{\beta}_0} = 10{,}727, \; S_{\hat{\beta}_1} = 0{,}270, \; S_{\hat{\beta}_2} = 0{,}112,$$
$$S_{\hat{\beta}_3} = 0{,}029, \; S_{\hat{\beta}_4} = 10{,}938, \; S_{\hat{\beta}_5} = 0{,}090.$$

2 *Проверка значимости коэффициентов модели.* Выдвигаются нулевые гипотезы H_0: $\beta_i = 0$ против альтернативных H_1: $\beta_i \neq 0$, $i = 1, \ldots, 5$. На уровне значимости $\alpha = 0{,}05$ квантиль распределения Стьюдента равен

$$t_{кр} = t_{1-0{,}05/2}(9) = \text{СТЬЮДЕНТ.ОБР}(0{,}975;9) = 2{,}262,$$

p-значение t- статистики равно

$$Pv = \text{СТЬЮДЕНТ.РАСП.2Х}(t;15-5-1).$$

В таблице 1.3 представлены результаты проверки значимости коэффициентов модели. Сравнивая статистики t (столбец 2) с квантилем $t_{кр}$ и p-значение (столбец 3) с $\alpha = 0{,}05$, делаем вывод о значимости коэффициентов модели (столбец 4). В столбце 5 приведены доверительные интервалы коэффициентов модели.

Таблица 1.3 – Значимость коэффициентов модели

Коэффициент	Статистика критерия	*p* - значение	Значимость	Доверительный интервал
1	2	3	4	5
$\hat{\beta}_0$	1,635	0,136	Не значим	(–6,726; 41,808)
$\hat{\beta}_1$	–2,153	0,060	Не значим	(–1,194; 0,029)
$\hat{\beta}_2$	4,846	0,001	Значим	(0,291; 0,800)
$\hat{\beta}_3$	4,075	0,003	Значим	(0,053; 0,186)
$\hat{\beta}_4$	–2,054	0,070	Не значим	(–47,208; 2,277)
$\hat{\beta}_5$	–0,030	0,977	Не значим	(–0,207; 0,202)

3 Оценка качества и значимости модели. Рассчитаем суммы квадратов *TSS*, *ESS*, *RSS* и построим дисперсионную таблицу (таблица 1.4).

Таблица 1.4 – Дисперсионная таблица

Источник вариации	Сумма квадратов (*SS*)	Число степеней свободы (*df*)	Дисперсия (*MS*)
Регрессия	*RSS* = 223,666	5	223,666/5 = 44,733
Остатки	*ESS* = 31,116	9	31,116/9 = 3,457
Полная вариация	*TSS* = 254,782	14	254,782/14 = 18,199

Коэффициент детерминации равен R^2 = 223,666/254,782 = 0,878, скорректированный коэффициент детерминации равен R^2_{adj} = 0,810. Таким образом, 87,8 % вариации инфляции *Y* обусловлено изменчивостью ВВП (X_1), денежной массы (X_2), импорта товаров и услуг в процентах от ВВП (X_3), индекса реального эффективного обменного курса (X_4), процентной ставки по кредитам (X_5).

Для проверки значимости R^2 выдвигается нулевая гипотеза H_0: $R^2 = 0$ против альтернативной H_1: $R^2 \neq 0$. Статистика *F*-критерия равна *F* = 12,939. Квантиль распределения Фишера при α = 0,05 и *Pv* равны соответственно

$$F_{кр} = F_{1-0,05}(5;15 – 5 – 1) = \text{F.ОБР}(0,95;5;9) = 3,482,$$

$$Pv = \text{F.РАСП.ПХ}(12,939;15 – 5 – 1) = 0,0007.$$

Поскольку $F > F_{кр}$ ($Pv = 0,0007 < 0,05$), то гипотеза H_0 отклоняется, и R^2 признается статистически значимым.

4 Оценка влияния объясняющих переменных на зависимую переменную. В таблице 1.5 представлены коэффициенты эластичности, бета-коэффициенты и дельта-коэффициенты, позволяющие оценить влияние объясняющих переменных на зависимую переменную.

На основе коэффициентов эластичности можно сделать вывод, что при изменении объясняющих переменных на 1 % наибольшую изменчи-

вость инфляции (Y) обеспечивает индекс реального эффективного обменного курса (X_4). При изменении X_2 на величину своего среднеквадратического отклонения зависимая переменная Y увеличится на 64,6 % своего среднеквадратического отклонения (максимальное изменение). Поскольку коэффициент $\Delta_5 < 0$, то можно сделать вывод, что хотя бы одна из предпосылок Гаусса-Маркова не выполняется.

Таблица 1.5 – Коэффициенты эластичности, бета-коэффициенты, дельта-коэффициенты

	x_{t1}	x_{t2}	x_{t3}	x_{t4}	x_{t5}
Коэффициенты эластичности, E_j	– 0,251	1,211	1,027	– 4,525	– 0,006
Бета-коэффициенты, $в^*_j$	– 0,399	0,646	0,505	– 0,437	– 0,005
Дельта-коэффициенты, Δ_j	0,260	0,345	0,113	0,285	– 0,003

5 Построение точечного и интервального прогнозов переменной Y. Пусть прогнозные значения объясняющих переменных равны

$$x_{(16)1} = 2{,}370,\ x_{(16)2} = 10{,}115,\ x_{(16)3} = 43{,}007,\ x_{(16)4} = 0{,}978,\ x_{(16)5} = 12{,}309.$$

Подстановкой в (1.18) прогнозных значений объясняющих переменных получим точечный прогноз инфляции

$$\hat{y} = 17{,}541 - 0{,}582 \cdot 2{,}370 + 0{,}545 \cdot 10{,}115 + 0{,}120 \cdot 43{,}007 - 22{,}466 \cdot 0{,}978 -$$

$$- 0{,}003 \cdot 12{,}309 = 4{,}820.$$

Ошибка прогноза $Se(\hat{y}_{16}) = 0{,}083$. Тогда интервальный прогноз для $\hat{y}_{16}$ на уровне значимости $\alpha = 0{,}05$ есть интервал

$$(4{,}820 - 0{,}083 \cdot 2{,}262;\ 4{,}820 - 0{,}083 \cdot 2{,}262) = (4{,}631;\ 5{,}009).$$

Таким образом, с вероятностью 0,95 фактическое значение y_{16} попадет в полученный интервал.

1.3 Задания для лабораторной работы 1

Имеются статистические данные (таблица 1.6) уровня рентабельности торговой деятельности Y (%); среднемесячного товарооборота в расчете на душу населения X_1 (шт/чел), удельного веса продовольственных товаров в товарообороте X_2 (%), времени обращения товаров X_3 (дней), среднемесячной оплаты труда X_4 (ден. ед) и трудоемкости товарооборота X_5 (численности работников на 100000 единиц товарооборота). В таблице k – номер студента в журнале группы; $[z]$ – целая часть числа z. Требуется:

1) построить модель множественной регрессии Y от X_1, X_2, X_3, X_4, X_5;

2) проверить статистическую значимость коэффициентов модели, построить их доверительные интервалы;

3) для оценки качества построенной модели определить коэффициент детерминации, скорректированный коэффициент детерминации, проверить гипотезу о значимости коэффициента детерминации;

4) оценить влияние объясняющих переменных на зависимую переменную с помощью коэффициентов эластичности, бета-коэффициентов и дельта-коэффициентов;

5) построить точечный и интервальный прогнозы з Y для прогнозных значений переменных X_1, X_2, X_3, X_4, X_5 на уровне значимости $\alpha = 0{,}05$.

Таблица 1.6 – Исходные данные для лабораторной работы 1

Номер наблюдения	y_t	x_{t1}	x_{t2}	x_{t3}	x_{t4}	x_{t5}
1	2,62+k	27+[k/10]	74,2	25	1 560	11
2	2,8+k	29	73,5	27+[k/2]	1 620	12+k/2
3	2,77+k	28+[k/10]	70	25	1 490	13
4	2,92+k	23	64,3	29+[k/2]	1 330	14+k/2
5	3,33+k	35+[k/10]	87,3	31	1 970	10
6	3,21+k	33	86,1	26+[k/2]	1 820	15+k/2
7	3,02+k	22+[k/10]	63,1	39	1 270	18
8	3,75+k	28	67	33+[k/2]	1 490	13+k/2
9	3,12+k	21+[k/10]	64,3	41	1 335	17
10	4,33+k	25	77,3	35+[k/2]	1 965	9+k/2
11	4,01+k	33+[k/10]	80,1	38	1 850	10
12	4,02+k	22	63,1	39+[k/2]	1 270	18+k/2
13	4,73+k	28+[k/10]	67	33	1 480	19
14	4,12+k	21	64,3	42+[k/2]	1 335	17+k/2
15	4,31+k	25+[k/10]	78,3	47	1 967	12
Прогнозные значения						
16		33	70,1	31	1850	10

2 Анализ структурных изменений в модели множественной линейной регрессии

2.1 Краткие теоретические сведения

2.1.1 Фиктивные переменные

Фиктивные переменные (dummy variables) – это переменные, которые не имеют содержательной экономической интерпретации и используются в модели для учета качественных переменных, структурных или сезонных

изменений, аномальных наблюдений. Фиктивные переменные, как правило, являются бинарными переменными, то есть каждая переменная принимает только два значения 1 и 0:

$$d_t = \begin{cases} 0 \text{ при отсутствии признака,} \\ 1 \text{ при наличии признака.} \end{cases} \qquad (2.1)$$

Если качественная переменная имеет k альтернативных значений, то при моделировании используют только $(k-1)$ фиктивных переменных. В противном случае сумма фиктивных переменных равна константе, и, как следствие, становится затруднительным оценивание модели с помощью МНК (так называемая «ловушка фиктивной переменной» (dummy trap)).

Модель регрессии с переменной структурой – это модель регрессии, которая включает в качестве объясняющей переменной фиктивную переменную. В моделях регрессии применяются фиктивные переменные *сдвига* и фиктивные переменные *наклона.*

2.1.2 Модели с фиктивной переменной сдвига

Спецификация парной модели Y на X линейной регрессии с фиктивной переменной сдвига имеет вид

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 d_t + \xi_t, \qquad (2.2)$$

где d_t – качественная переменная бинарного типа, т.е. фиктивная переменная, определяемая (2.1) в наблюдении t.

Значение фиктивной переменной $d_t = 0$ называется *базовым.* Выбор базового значения определяется целями исследования или принимается произвольно. При замене базового значения переменной суть модели не меняется, а меняется знак параметра β_2 на противоположный.

Условное математическое ожидание зависимой переменной Y равно

$$M(Y \mid X, d_t = 0) = \beta_0 + \beta_1 x_t, \quad M(Y \mid X, d_t = 1) = \beta_0 + \beta_2 + \beta_1 x_t.$$

Величина β_2 представляет собой среднее изменение зависимой переменной Y при переходе из одной категории в другую при неизменных значениях остальных коэффициентов (рисунок 2.1).

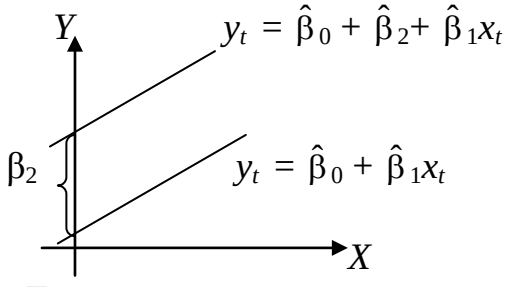

Рисунок 2.1 – Геометрическая интерпретация параметра β_2

Проверка статистической значимости коэффициента β_2 (определяемая при помощи *t*-критерия) показывает, влияет ли качественная переменная на переменную *Y*.

Замечание. Для оценки параметров модели исходные данные оформляются в виде таблицы 2.1.

Таблица 2.1 – Включение в модель фиктивной переменной сдвига

t	y_t	x_t	d_t
1	y_1	x_1	1
2	y_2	x_2	0
	…	…	…
n	y_n	x_n	1

2.1.3 Структурные изменения

Под *структурными изменениями* (structure breaks) модели понимаются скачкообразные изменения значений параметров, обусловленные изменением условий протекания моделируемого процесса или внешними «шоковыми» воздействиями. Пусть имеются два типа условий протекания процесса, и предполагается, что смена условий может являться причиной структурного изменения модели.

Для *k*-го ($k = 1, 2$) типа условий обозначим:

n_k – объем выборки для пространственных данных или длина временного ряда, $n_1 + n_2 = n$;

$Y^{(1)} = (y_1 \ y_2 \ \dots \ y_{n1})^T$ – вектор значений зависимой переменной размерности n_1;

$Y^{(2)} = (y_{n1+1} \ y_{n1+2} \ \dots \ y_{n1+n2})^T$ – вектор значений зависимой переменной размерности n_2;

$X^{(k)}$ – матрица значений объясняющих переменных $X_1, X_2, \dots, X_m$ размерности $n_k \times (m + 1)$, $k = 1, 2$ (первый столбец – единичный).

Наличие (отсутствие) в выборочных данных структурных изменений можно проверить с помощью теста Чоу (Chow), который исследует возможность использования единой модели регрессии. Для этого строятся следующие модели множественной линейной регрессии:

1) общая модель по объединенной выборке

$$y_t = \beta_0 + \sum_{k=1}^{m} \beta_k x_{tk} + \xi_t, \ t = 1, \dots, n_1 + n_2;$$

2) частные модели по каждой выборке в отдельности:

$$y_t = \beta_0^{(1)} + \sum_{k=1}^{m} \beta_k^{(1)} x_{tk} + \xi_t^{(1)}, \ t = 1, \dots, n_1,$$

$$y_t = \beta_0^{(2)} + \sum_{k=1}^{m} \beta_k^{(2)}{}_k x_{tk} + \xi_t^{(2)}, \ t = n_1 + 1, \dots, n_1 + n_2.$$

Пусть ESS_1, ESS_2 – суммы квадратов остатков для частных моделей; ESS – сумма квадратов остатков для общей модели. Равенство сумм ESS и $ESS_1 + ESS_2$ будет иметь место только при совпадении коэффициентов в общей и частных моделях. В других случаях для частных моделей выполняются условия: $ESS_1 < ESS$, $ESS_2 < ESS$ и $ESS_1 + ESS_2 < ESS$. Чем больше разница между двумя частями последнего неравенства, тем больше различие между двумя подвыборками с точки зрения параметров модели.

Выдвигается нулевая гипотеза H_0: $\beta_i^{(1)} = \beta_i^{(2)}$ об отсутствии структурных изменений против альтернативной H_1: $\beta_i^{(1)} \neq \beta_i^{(2)}$, $i = 0, 1, \ldots, m$. Определяемое гипотезой H_0 ограничение означает отсутствие структурных изменений. Для проверки гипотезы используется F-критерий Фишера со статистикой

$$F = \frac{(ESS - (ESS_1 + ESS_2))/(m+1)}{(ESS_1 + ESS_2)/(n - 2(m+1))} \sim F\,(m + 1,\, n - 2(m + 1)). \qquad (2.3)$$

Гипотеза H_0 отклоняется на уровне значимости α, если $F \geq F_{кр}$, или $Pv = P\{F(m + 1,\ n - 2(m + 1)) \geq F\} < \alpha$. Тогда есть структурные изменения в выборочных данных, и качество частных моделей регрессии превосходит качество общей модели без ограничений. Если $F < F_{кр}$, то нулевая гипотеза не отклоняется, и нет необходимости разбивать общую модель регрессии без ограничений на частные модели. Здесь $F_{кр} = F_{1-\alpha}(m + 1, n - 2(m + 1))$ – квантиль распределения Фишера.

2.1.4 Фиктивная переменная наклона

Фиктивная переменная наклона изменяет наклон линии регрессии. С помощью фиктивных переменных наклона можно построить кусочно-линейные модели, которые позволяют учесть *структурные изменения*. Например, спецификация модели парной регрессии имеет вид

$$y_t = \beta_0 + \beta_1 x_t + \beta_2\, x_t\, d_t + \xi_t, \qquad (2.4)$$

где

$$d_t = \begin{cases} 0 \text{ до структурных изменений,} \\ 1 \text{ после структурных изменений.} \end{cases}$$

Фиктивная переменная наклона d_t входит в уравнение в мультипликативной форме $x_t\, d_t$. Значимость коэффициентов β_1 и β_2 говорит о наличии структурных изменений.

2.2 Решение типовых примеров

Пример 2.1. В таблице 2.2 представлены квартальные данные о выпуске в сфере сельского хозяйства, охоты и лесного хозяйства (Y)

в Республике Беларусь в период с 2011 по 2013 гг. Требуется построить модель парной регрессии, учитывающую сезонные колебания на уровне значимости α = 0,05.

Таблица 2.2 – Выпуск в сфере сельского хозяйства, охоты и лесного хозяйства (в млрд. рублей)

Год	I квартал	II квартал	III квартал	IV квартал
2011	5 316,4	8 591,7	26 927,8	17 693,7
2012	12 497,5	18 368,4	42 424,8	28 800,4
2013	15 309,8	20 036,4	46 430,4	31 948,7

Решение. На рисунке 2.2 представлена зависимость выпуска y_t от времени t (кварталы). Изменение объема выпуска по кварталам имеет сезонный характер. В качестве базового периода рассмотрим I квартал года.

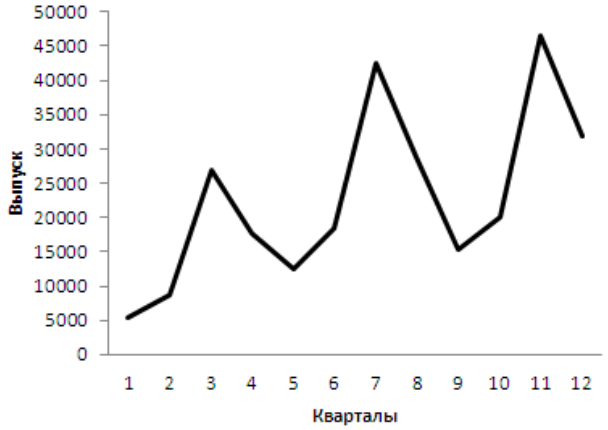

Рисунок 2.2 – Выпуск в сфере сельского хозяйства, охоты и лесного хозяйства

Для выявления фактора сезонности в квартале i, $i = 2, 3, 4$, в спецификацию модели введем следующие бинарные переменные:

$$d_i = \begin{cases} 1 & \text{для квартала } i, \\ 0 & \text{для остальных кварталов.} \end{cases}$$

Спецификация модели парной регрессии с фиктивными переменными, учитывающими сезонность, имеет вид

$$y_t = \beta_0 + \beta_1 t + \beta_2 d_{t2} + \beta_3 d_{t3} + \beta_4 d_{t4} + \xi_t.$$

Для построения модели данные удобно представить в виде таблицы 2.3.

Таблица 2.3 – Данные для построения модели к примеру 2.1

Кварталы	Выпуск	Кварталы	II квартал	III квартал	IV квартал
1	5 316,4	1	0	0	0
2	8 591,7	2	1	0	0
3	26 927,8	3	0	1	0
4	17 693,7	4	0	0	1
1	12 497,5	5	0	0	0
2	18 368,4	6	1	0	0
3	42 424,8	7	0	1	0
4	28 800,4	8	0	0	1
1	15 309,8	9	0	0	0
2	20 036,4	10	1	0	0
3	46 430,4	11	0	1	0
4	31 948,7	12	0	0	1

Коэффициенты модели оцениваются с помощью метода наименьших квадратов, и модель, учитывающая сезонные колебания, принимает вид

$$y_t = \underset{(1,029)}{2\,416,905} + \underset{(6,039)}{1\,724,866}t + \underset{(1,093)}{2\,899,401}d_{t2} + \underset{(8,929)}{24\,103,369}d_{t3} + \underset{(3,580)}{9\,931,770}d_{t4}, \quad (2.5)$$

$$R^2 = 0,959,\ F = 41,430,\ F_{кр} = F_{1-0,05}(4;7) = 4,120.$$

В круглых скобках под коэффициентами указываются значения t-статистик. На уровне значимости $\alpha = 0,05$, сравнивая t-статистики с квантилем $t_{кр} = t_{1-0,05/2}(7) = 2,365$, P-значения, равные 0,338 и 0,311 соответственно, с $\alpha = 0,05$, получим, что статистически незначимыми являются коэффициенты β_0 и β_2, остальные коэффициенты значимы. Так как $F > F_{кр}$ ($Pv = 0,00006 < 0,05$), то R^2 статистически значим, и в модели можно оставить все переменные.

Подставляя значения фиктивных переменных в модель регрессии (2.5), получим уравнения, соответствующие каждому кварталу:

– для I квартала $d_2 = d_3 = d_4 = 0$ и $y_t = 2\,416,905 + 1\,724,866t$;
– для II квартала $d_2 = 1$, $d_3 = d_4 = 0$ и $y_t = 5\,316,306 + 1\,724,866t$;
– для III квартала $d_3 = 1$, $d_2 = d_4 = 0$ и $y_t = 26\,520,274 + 1\,724,866t$;
– для IV квартала $d_4 = 1$, $d_2 = d_3 = 0$ и $y_t = 5\,190,845 + 1\,724,866t$.

Пример 2.2. В таблице 2.4 представлены курс доллара США (X) по отношению к белорусскому рублю, средняя стоимость квадратного метра общей площади жилья (Y) в Республике Беларусь за период с 2008 по 2015 год. Требуется построить модель парной регрессии.

Таблица 2.4 – Исходные данные к примеру 2.2

Год, t	Стоимость квадратного метра жилья, y_t (тыс. руб.)	Курс доллара США (на 31 декабря), x_t (\$/руб)
2008	2 138,8	2 200
2009	2 363,4	2 863
2010	2 681,8	3 000
2011	4 540,9	8 350
2012	7 667,0	8 570
2013	8 976,3	9 510
2014	1 1039,3	11 850
2015	13 821,6	18 569

Решение. На рисунке 2.3 изображен график зависимости стоимости квадратного метра площади жилья от курса доллара США.

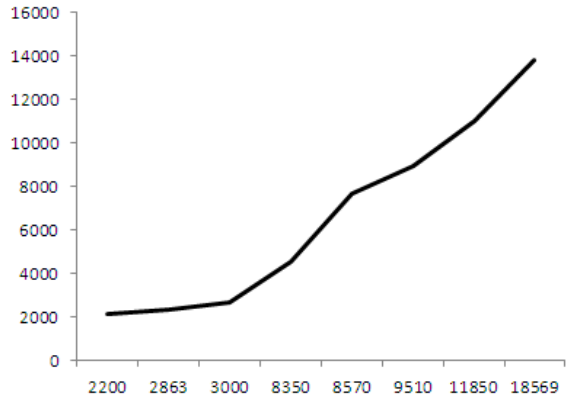

Рисунок 2.3 – График зависимости исследуемых показателей

По графику видно, что в выборочных данных могут быть структурные изменения. Проверим наличие (отсутствие) структурных изменений с помощью теста Чоу. Разобьем общую выборку на две части с 2008 по 2011 гг. и с 2012 по 2015 гг.

Оценим общую модель линейной регрессии:

$$y_t = 437{,}867 + 0{,}766 x_t, \qquad (2.6)$$

$$ESS = 10\,375\,813{,}068$$

и частные модели:

$$y_t = 1\,364{,}142 + 0{,}382 x_t,\ ESS_1 = 42\,868{,}145,$$

$$y_t = 3\,366{,}111 + 0{,}578 x_t,\ ESS_2 = 1\,194\,435{,}628.$$

Значение F-статистики по формуле (2.3) равно

$$F = \frac{10\,375\,813{,}068 - (42\,868{,}145 + 1194\,435{,}628)}{42\,868{,}145 + 1194\,435{,}628} \cdot \frac{8-2-2}{1+1} = 14{,}772\,.$$

Квантиль равен $F_{кр} = F_{1-0,05}(2;4) = 6,944$ на уровне значимости $\alpha = 0,05$. Поскольку $F > F_{кр}$ ($Pv = 0,014 < 0,05$), то гипотеза H_0 отклоняется, и в выборке есть структурные изменения.

Для учета структурных изменений в выборке введем в модель (2.6) фиктивную переменную в мультипликативной форме, оценим спецификацию, включающую как фиктивную переменную сдвига, так и фиктивную переменную наклона

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 d_t + \beta_3 x_t d_t + \xi_t.$$

Составим расчетную таблицу 2.5.

Таблица 2.5 – Расчетная таблица к примеру 2.2

t	y_t	x_t	d_t	$x_t d_t$
2008	2 138,8	2 200	0	0
2009	2 363,4	2 863	0	0
2010	2 681,8	3 000	0	0
2011	4 540,9	8 350	0	0
2012	7 667,0	8 570	1	8 570
2013	8 976,3	9 510	1	9 510
2014	11 039,3	11 850	1	11 850
2015	13 821,6	18 569	1	18 569

Оцененная модель имеет вид

$$y_t = \underset{(2,530)}{1\,364,141} + \underset{(3,393)}{0,382}\, x_t + \underset{(1,898)}{2\,001,969}\, d_t + \underset{(1,474)}{0,196}\, x_t d_t,$$

$$R^2 = 0,991,\ F = 145,291,\ F_{кр} = F_{0,95}(3;4) = 6,591.$$

Сравнивая t-статистики для коэффициентов модели с квантилем $t_{кр} = t_{1-0,05/2}(4) = 2,776$, получим, что статистически значимым является только коэффициент β_1, при этом $Pv = 0,027 < 0,05$. Однако R^2 статистически значим, так как $F > F_{кр}$ ($Pv = 0,000155 < 0,05$). Следовательно, целесообразно оставить фиктивные переменные в модели.

2.3 Задания для лабораторной работы 2

Задание 1. В таблице 2.6 представлены квартальные данные об объемах потребления (Y) и доходах домашних хозяйств (X) с I квартала 2010 г. по III квартал 2015 г. Требуется построить модель регрессии Y на X, учитывающую сезонные колебания.

Таблица 2.6 – Исходные данные к заданию 1

Год	Квартал	Объем потребления, y_t (ден. ед.)	Доход, x_t (ден. ед.)
2010	I	242,1	13,5
	II	269,4	16,3
	III	272,1	15,5
	IV	277,0	13,4
2011	I	247,1	9,3
	II	235,8	12,4
	III	271,0	13,2
	IV	281,3	14,2
2012	I	284,2	14,8
	II	307,6 + *a*	18,1 + *a*
	III	301,6 + *a*	16,0 + *a*
	IV	309,8 + *a*	15,6 + *a*
2013	I	311,5 + *a*	15,6 + *a*
	II	338,6 + *a*	19,7 + *a*
	III	331,7 + *a*	16,7 + *a*
	IV	346,2 + *a*	18,4 + *a*
2014	I	340,2 + *a*	16,0 + *a*
	II	377,5 + *a*	22,1 + *a*
	III	376,9 + *a*	20,4 + *a*
	IV	401,8 + *a*	22,6 + *a*
2015	I	406,2 + *a*	22,6 + *a*
	II	436,4 + *a*	26,8 + *a*
	III	437,5 + *a*	24,8 + *a*

Задание 2. Предприятие занимается продажей мясной продукции. В таблице 2.7 представлены объемы ежемесячных продаж Q (тонн) по ценам P (ден. ед. за 1 кг). Во время седьмого и восьмого месяцев на предприятии произошла забастовка. Проверить, произошли ли структурные изменения в выборочных данных, и построить модель регрессии зависимости объема продаж от цены.

Таблица 2.7 – Исходные данные к заданию 2

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Q	9	10	13	15	17	18	19	19	20 + *a*	21 + *a*	21 + *a*	23 + *a*	26 + *a*	28 + *a*
P	2	3	3	4	5	5,2	5	5,6	7 + *a*	7 + *a*	8 + *a*	8 + *a*	9 + *a*	9 + *a*

В таблицах 2.6 и 2.7 k – номер студента в журнале группы и

$$a = \begin{cases} 2k, & 1 \le k \le 10, \\ k, & 11 \le k \le 20. \end{cases}$$

3 Построение модели в условиях мультиколлинеарности факторов

3.1 Краткие теоретические сведения

3.1.1 Причины и последствия мультиколлинеарности

Мультиколлинеарность (multicollinearity) – это нарушение предпосылки Х.3 Гаусса-Маркова о независимости между собой объясняющих переменных $X_1, X_2, \ldots, X_m$, включенных в модель. В этом случае матрица X является вырожденной, поэтому матрица (X^TX) необратима, что порождает проблему идентифицируемости модели. *Нестрогая мультиколлинеарность* есть наличие сильной линейной корреляционной связи между объясняющими переменными. *Основная причина мультиколлинеарности* заключается в неправильном подборе переменных $X_1, X_2, \ldots, X_m$, включенных в модель: 1) ошибочное включение в модель множественной линейной регрессии двух или более линейно зависимых переменных; 2) объясняющие переменные, в нормальной ситуации слабо коррелированные, становятся в конкретной выборке сильно коррелированными; 3) в модель включается объясняющая переменная X_k, сильно коррелирующая с зависимой переменной Y; 4) использование в модели лаговых переменных. *Последствия мультиколлинеарности*: 1) основная гипотеза о незначимости коэффициентов в большинстве случаев не отклоняется, однако сама модель при проверке с помощью F-критерия оказывается значимой; 2) полученные оценки коэффициентов β_i неоправданно завышены или имеют неправильные знаки; 3) оценки коэффициентов β_i чувствительны к объему выборки; 4) наличие мультиколлинеарности может сделать модель непригодной для прогнозирования.

3.1.2 Обнаружение мультиколлинеарности

Для обнаружения мультиколлинеарности не существует точных критериев. Имеются некоторые признаки, по которым можно судить об ее наличии (отсутствии). *Внешним* признаком наличия мультиколлинеарности служат слишком большие значения элементов матрицы $(X^TX)^{-1}$.

Основной признак мультиколлинеарности: определитель корреляционной матрицы $Q = [r_{ij}]$, $i, j = 1, \ldots, m$, объясняющих переменных $X_1, X_2, \ldots, X_m$ близок к нулю, то есть $\det Q \approx 0$. Если все объясняющие переменные некоррелированы, то $\det Q = 1$. В противном случае $0 < \det Q < 1$.

Дополнительные признаки:

1) высокие R^2 и F-статистика, но некоторые (или даже все) коэффициенты β_i незначимы, то есть имеют низкие t-статистики;

2) высокие парные и частные коэффициенты корреляции между факторами;

3) знаки коэффициентов модели противоположны ожидаемым;

4) добавление (удаление) наблюдений выборки сильно изменяют значения оценок коэффициентов модели.

3.1.3 Методы устранения мультиколлинеарности

Если точность модели является удовлетворительной, то модель множественной линейной регрессии можно использовать и при наличии мультиколлинеарности. Если же точность модели получилась неудовлетворительной, то мультиколлинеарность необходимо устранять. Выделяют следующие методы устранения мультиколлинеарности.

1 Собрать дополнительные данные (на практике не всегда возможно).

2 Включить в модель более подходящие объясняющие переменные.

3 Преобразовать мультиколлинеарные переменные, используя линейные комбинации переменных, нелинейные формы модели, первые разности вместо самих переменных.

4 Исключить из модели одну (несколько) объясняющих переменных.

5 Использовать при оценке коэффициентов метод главных компонент или другие специальные процедуры расчета коэффициентов.

6 Использовать *ридж*-регрессию тогда, когда ни одну из переменных, включенных в модель регрессии, нельзя удалить. В этом случае ко всем диагональным элементам матрицы X^TX добавляется малое число τ, и оценка вектора параметров $\hat{\beta} = (\hat{\beta}_0\ \hat{\beta}_1\ \ldots\ \hat{\beta}_m)$ определяется по формуле $\hat{\beta} = (X^TX + \tau E)^{-1} X^TY$, где E – единичная матрица, $0{,}1 \le \tau \le 0{,}4$. Ридж-регрессия позволяет стабилизировать оценки коэффициентов.

7 *Метод пошагового включения* объясняющих переменных в модель, который предполагает определение из возможного набора переменных тех, которые улучшат качество модели регрессии.

Алгоритм метода пошагового включения

Шаг 1. Из исходного набора объясняющих переменных X_1, X_2, …, X_m выбирается переменная, имеющая наибольший по модулю коэффициент корреляции с зависимой переменной Y.

Шаг 2. Отбирается наиболее «информативная» пара переменных, одна из которых отобрана на шаге 1. Критерием отбора обычно служит R^2 или R^2_{adj}.

Шаг 3. Определяется тройка переменных, две из которых отобраны на шаге 2, с максимальным критерием отбора и т. д. Процесс повторяется до тех пор, пока включение очередных переменных не приводит к уменьшению критерия отбора или его стабилизации.

3.2 Решение типового примера

Пример 3.1 Для построенной в примере 1.1 модели множественной линейной регрессии требуется:

1) исследовать наличие (отсутствие) мультиколлинеарности;

2) уменьшить размерность модели методом пошагового включения.

Решение. 1 *Обнаружение мультиколлинеарности.* Для исследуемой модели среди элементов матрицы $(X^TX)^{-1}$ (1.17) есть очень большие и очень маленькие элементы, что может говорить о наличии мультиколлинеарности факторов. Корреляционная матрица Q объясняющих переменных $X_1, X_2, \ldots, X_5$ имеет вид

$$Q = \begin{pmatrix} 1 & 0{,}065 & 0{,}239 & 0{,}772 & -0{,}430 \\ 0{,}065 & 1 & -0{,}201 & 0{,}110 & 0{,}272 \\ 0{,}239 & -0{,}201 & 1 & 0{,}193 & -0{,}105 \\ 0{,}772 & 0{,}110 & 0{,}193 & 1 & -0{,}580 \\ -0{,}430 & 0{,}272 & -0{,}105 & -0{,}580 & 1 \end{pmatrix}.$$

Определитель матрицы $\det Q = 0{,}193$, среди элементов матрицы есть высокий парный коэффициент корреляции ($r_{X4X1} = 0{,}772$), что позволяет сделать предположение о наличии связи между объясняющими переменными.

Оцененная модель примера 1.1 имеет вид

$$y_t = \underset{(1{,}635)}{17{,}541} - \underset{(-2{,}153)}{0{,}582}x_{t1} + \underset{(4{,}846)}{0{,}545}x_{t2} + \underset{(4{,}075)}{0{,}120}x_{t3} - \underset{(-2{,}054)}{22{,}466}x_{t4} - \underset{(-0{,}030)}{0{,}003}x_{t5},$$

$$t_{\text{кр}} = 2{,}262,\ R^2 = 0{,}878,\ R^2_{adj} = 0{,}810,\ F = 12{,}939,\ F_{\text{кр}} = 7{,}471,\ Pv = 0{,}0007.$$

Видно, что среди коэффициентов модели есть незначимые, а R^2 и F-статистика имеют достаточно высокие значения, что также говорит о наличии мультиколлинеарности объясняющих переменных.

2 Устранение мультиколлинеарности методом пошагового включения.

Шаг 1. Из набора переменных $X_1, X_2, \ldots, X_5$ выберем переменную, имеющую наибольший по модулю коэффициент корреляции с зависимой переменной Y: $r_{YX1} = -0{,}572$, $r_{YX2} = -0{,}469$, $r_{YX3} = 0{,}196$, $r_{YX4} = -0{,}573$, $r_{YX5} = 0{,}543$, то есть переменную X_4.

Шаг 2. Выберем наиболее информативную пару переменных (одна из которых X_4), для которой значение скорректированного коэффициента детерминации является наибольшим (таблица 3.1, столбец «Информативные пары»). Максимальное значение R^2_{adj} соответствует включению в спецификацию модели переменных X_4 и X_2.

Шаг 3. Результаты вычисления коэффициентов детерминации моделей с тремя переменными представлены в таблице 3.1 (столбец «Инфор-

мативные тройки»). Максимальное значение R^2_{adj} соответствует включению в спецификацию модели переменных X_4, X_2 и X_3.

Шаг 4. Добавление четвертой переменной к тройке X_4, X_2, X_3 незначительно увеличивает значение R^2 (и соответственно R^2_{adj}) (таблица 3.1, столбец «Информативные четверки»), поэтому рекомендуется оставить в модели только информативную тройку. Тогда модель имеет вд

$$y_t = \underset{(4,782)}{31,947} + \underset{(4,812)}{0,544}x_{t2} + \underset{(3,340)}{0,111}x_{t3} - \underset{(-5,498)}{37,795}x_{t4},$$

$$t_{кр} = 2,201,\ R^2_{adj} = 0,765,\ F = 16,177,\ F_{кр} = 3,587,\ Pv = 0,0002.$$

Таблица 3.1 – Выбор информативных переменных к примеру 3.1

	Информативные пары				Информативные тройки			Информативные четверки	
	X_4, X_1	X_4, X_2	X_4, X_3	X_4, X_5	X_4, X_2, X_1	X_4, X_2, X_3	X_4, X_2, X_5	X_4, X_2, X_3, X_1	X_4, X_2, X_3, X_5
R^2	0,370	0,615	0,427	0,396	0,651	0,815	0,618	0,878	0,815
R^2_{adj}	0,265	**0,551**	0,331	0,295	0,555	**0,765**	0,514	**0,829**	0,742
F	3,523	9,602	4,463	3,929	6,826	16,177	5,933	17,962	11,040
Pv	0,063	0,003	0,036	0,049	0,007	0,0002	0,012	0,0001	0,001
	$F_{кр} = F_{0,95}(2;12) =$ = F.ОБР(0,95;2;12) = 3,885				$F_{кр} = F_{0,95}(3;11) =$ = F.ОБР(0,95;3;11) = 3,587			$F_{кр} = F_{0,95}(4;10) =$ = F.ОБР(0,95;4;10) = 3,478	

3.3 Задания для лабораторной работы 3

По статистическим данным (таблица 1.6) и результатам исследований лабораторной работы 1 требуется:

1) проверить модель на наличие (отсутствие) мультиколлинеарности объясняющих переменных;

2) уменьшить размерность модели методом пошагового включения.

4 Проблема автокорреляции остатков модели множественной линейной регрессии

4.1 Краткие теоретические сведения

4.1.1 Причины и последствия автокорреляции

Автокорреляция (avtocorrellation) – это корреляция между наблюдаемыми значениями во времени (временные ряды) или в пространстве (про-

странственные данные). *Автокорреляция остатков* характеризуется тем, что не выполняется предпосылка ξ.3 Гаусса-Маркова.

Лаг – величина сдвига между уровнями остатков модели регрессии. Величина лага определяет порядок q коэффициента автокорреляции ρ_q.

Виды автокорреляции: 1) *чистая* автокорреляция, обусловленная зависимостью случайных ошибок ξ_t от прошлых значений ξ_{t-1}, ξ_{t-2}, …; 2) *ложная* автокорреляция, вызванная неправильной спецификацией модели регрессии. *Причины чистой автокорреляции:* 1) инерция (изменение зависимой переменной под влиянием объясняющих происходит не мгновенно, а обладает определенной инертностью); 2) эффект паутины (многие экономические показатели реагируют на изменение экономических условий с запаздыванием); 3) сглаживание данных (усреднение данных по некоторому продолжительному интервалу времени).

Положительная автокорреляция ($\text{cov}(\xi_t, \xi_\tau) > 0$) чаще всего вызывается воздействием неучтенных в модели переменных. *Отрицательная* автокорреляция ($\text{cov}(\xi_t, \xi_\tau) < 0$) фактически означает, что за положительным отклонением ξ_t следует отрицательное и наоборот.

Последствия автокорреляции: 1) оценки модели множественной линейной регрессии остаются несмещенными и состоятельными, но теряется эффективность; 2) автокорреляция (особенно положительная) часто приводит к уменьшению стандартных ошибок коэффициентов, что влечет за собой увеличение t-статистик; 3) оценка дисперсии остатков является смещенной оценкой истинного значения, во многих случаях заниженной; 4) выводы по оценке качества коэффициентов и модели в целом могут быть неверными, что приводит к ухудшению прогнозных качеств модели.

4.1.2 Обнаружение автокорреляции

Обнаружение автокорреляции осуществляется путем анализа графика остатков модели, с помощью метода рядов и статистических тестов.

Оценивается модель

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_m x_{tm} + \xi_t, \qquad (4.1)$$

определяются $e_t = y_t - \hat{y}_t$ – остатки модели (оценки ошибки ξ_t), где y_t, $\hat{y}_t$ – соответственно наблюдаемые и модельные значения зависимой переменной Y, $t = 1, \ldots, n$.

Анализ графика остатков модели. Строится график зависимости остатков e_t от t, $t = 1, \ldots, n$. По виду корреляционного поля выдвигается предположение о наличии (отсутствии) автокорреляции. Наличие резких изменений знаков остатков может свидетельствовать об отрицательной автокорреляции; если же знаки остатков и их абсолютные значения изменяются не сильно, то можно ожидать положительную автокорреляцию.

По графику зависимости e_t от e_{t-1} выдвигается предположение о наличии (отсутствии) в остатках автокорреляции 1-го порядка.

Метод рядов (Runs Test, Ceary test). Последовательно определяются знаки остатков e_t, $t = 1, 2, \ldots, n$. Непрерывная последовательность одинаковых знаков называется *рядом* (*серией*), количество знаков в ряду – *длиной* ряда. Если рядов слишком мало по сравнению с количеством наблюдений n, то вполне вероятна положительная автокорреляция, если много – отрицательная автокорреляция.

Выдвигается нулевая гипотеза H_0: $\rho_1 = 0$ против альтернативной H_1: $\rho_1 \neq 0$. Для проверки гипотезы используется z-критерий, имеющий стандартное нормальное распределение, со статистикой

$$z = \left(k - \frac{2n^+n^-}{n} + 1 \right) \Big/ \sqrt{\frac{2n^+n^-(2n^+n^- - n)}{n^2(n-1)}} \sim N(0, 1). \qquad (4.2)$$

Здесь n^+ – количество знаков «+», n^- – количество знаков «–» при n наблюдениях, $n^+ + n^- = n$, k – количество рядов.

Гипотеза H_0 отклоняется на уровне значимости α, если $|z| \geq u_{1-\alpha/2}$ или $Pv = P\{u \geq z\} < \alpha$, то есть автокорреляция 1-го порядка в остатках присутствует. Здесь $u_{1-\alpha/2}$ – квантиль стандартного нормального распределения.

При малых n^+, n^- используются таблицы для k [3], определяющие нижнее k_1 и верхнее k_2 значения на уровне значимости α. Если $k_1 < k < k_2$, то говорят об отсутствии автокорреляции; $k \leq k_1$ – о положительной автокорреляции; $k \geq k_2$ – об отрицательной автокорреляции остатков.

Критерий Дарбина-Уотсона (Durbin-Watson). Тест Дарбина-Уотсона предназначен для обнаружения автокорреляции 1-го порядка, когда случайные ошибки ξ_t определяются по итерационной схеме $AR(1)$:

$$\xi_t = \rho_1 \xi_{t-1} + \eta_t, \qquad (4.3)$$

где η_t – независимые нормально распределенные случайные величины, $M(\eta_t) = 0$ и $D(\eta_t) = \sigma^2$, то есть $\eta_t \sim N(0; \sigma^2)$.

Выдвигается нулевая гипотеза H_0: $\rho_1 = 0$ против альтернативной H_1: $\rho_1 \neq 0$. Для проверки гипотезы используется DW-статистика

$$DW = \sum_{t=2}^{n}(e_t - e_{t-1})^2 \Big/ \sum_{t=1}^{n} e_t^2, \; DW \in [0, 4], \qquad (4.4)$$

с числом степеней свободы $df = m - 1$.

На уровне значимости α по таблицам критических значений DW-статистики [3] определяются значения верхней d_U и нижней d_L границ интервала. Возможны случаи:

1) гипотеза H_0 отклоняется при $0 \leq DW < d_L$ ($\rho_1 > 0$);

2) неопределенность при $d_L < DW_{\text{набл}} < d_U$ и $4 - d_U \leq DW < 4 - d_L$;

3) гипотеза H_0 не отклоняется при $d_U \leq DW < 4 - d_U$;

4) гипотеза H_0 отклоняется при $4 - d_L \leq DW \leq 4$ ($\rho_1 < 0$).

МНК-оценкой ρ

4.1.3 Корректировка спецификации модели

Рассмотрим авторегрессионное преобразование (autoregressive transformation). Пусть ошибки модели (4.1) есть результат модели AR(1), то есть

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} \ldots + \beta_m x_{tm} + \xi_t, \quad (4.10)$$

$$\xi_t = \rho_1 \xi_{t-1} + \eta_t,\ \eta_t \sim N(0; \sigma^2),\ t = 1, \ldots, n. \quad (4.11)$$

Для $(t-1)$-го наблюдения имеем по (4.10)

$$y_{t-1} = \beta_0 + \beta_1 x_{(t-1)1} + \beta_2 x_{(t-1)2} \ldots + \beta_n x_{(t-1)n} + \xi_{(t-1)}. \quad (4.12)$$

Тогда с учетом (4.10) и (4.12) получим

$$y_t - \rho_1 y_{t-1} = \beta_0(1-\rho_1) + \beta_1(x_{t1} - \rho_1 x_{(t-1)1}) + \ldots + \beta_m(x_{tm} - \rho_1 x_{(t-1)m}) + (\xi_t - \rho_1 \xi_{t-1}).$$

Положим

$$y_t^* = y_t - \rho_1 y_{t-1},\ \beta_0^* = \beta_0(1-\rho_1),\ x_{tj}^* = x_{tj} - \rho_1 x_{(t-1)j},\ j = 1, \ldots, m. \quad (4.13)$$

Тогда

$$y_t^* = \beta_0^* + \beta_1 x_{t1}^* + \ldots + \beta_m x_{tm}^* + \eta_t. \quad (4.14)$$

Случайная ошибка η_t модели (4.14) уже не подвержена автокорреляции, поэтому автокорреляцию остатков модели регрессии (4.10), (4.11) можно считать устраненной. МНК-оценки β_0^* и β_j^* являются несмещенными, состоятельными и эффективными.

Замечания. 1 Переход к y_t^* и x_t^* приводит к потере первого наблюдения и число степеней свободы уменьшается на 1, что при больших выборках не так существенно, но при малых выборках может привести к потере эффективности. Проблема потери первого наблюдения преодолевается с помощью поправки Прайса-Винстена (Price-Winsten):

$$y_1^* = \sqrt{1-\rho_1^2}\, y_1,\ x_{1j}^* = \sqrt{1-\rho_1^2}\, x_{1j},\ j = 1, \ldots, m \quad (4.15)$$

2 Авторегрессионное преобразование может быть обобщено на преобразования более высоких порядков.

4.1.4 Оценка ρ_1 и коэффициентов $\beta_0, \beta_1, \ldots, \beta_m$

На практике коэффициент автокорреляции 1-го порядка $\rho_1 = \rho$ обычно не известен. Существует несколько методов его оценивания.

1 При $n \to \infty$ на основе статистики DW из (4.6) определяется оценка

$$\hat{\rho} \approx 1 - DW/2, \quad (4.16)$$

которая далее используется в авторегрессионном преобразовании с поправкой Прайса-Винстена. Далее находятся МНК-оценки параметров β_0, $\beta_1, \ldots, \beta_m$ модели (4.1).

2 Метод Кохрейна-Оркатта (Cochrane-Orcutt), определяющий последовательно приближения параметра ρ путем итераций с точностью ε.

Алгоритм реализации метода Кохрейна-Оркатта

Шаг 1. Найдем МНК-оценки коэффициентов модели (4.1) и обозначим их $\beta_0^0, \beta_1^0, \ldots, \beta_m^0$.

Шаг 2. Для i - й итерации найдем остатки

$$e_t^i = y_t - (\beta_0^i + \beta_1^i x_{t1} + \ldots + \beta_m^i x_{tm}),\ i = 0, 1, 2, \ldots.$$

Шаг 3. По остаткам e_t^i оценим модель AR(1) вида (4.11), откуда найдем оценку параметра ρ (обозначим её $\hat{\rho}^i$). Если $|\hat{\rho}^i - \hat{\rho}^{i-1}| < \varepsilon$, то считаем, что $\hat{\rho}^i = \hat{\rho}$, процесс поиска прерывается, и переходим к шагу 5. В противном случае шаг 4.

Шаг 4. Получив $\hat{\rho}^i$, применим авторегрессионное преобразование с поправкой Прайса-Винстена. По преобразованным данным найдем МНК-оценки $\beta_0^{i+1}, \beta_1^{i+1}, \ldots, \beta_m^{i+1}$. Переходим к шагу 2 и находим остатки e_t^i с учётом новых оценок параметров $\beta_0^{i+1}, \beta_1^{i+1}, \ldots, \beta_m^{i+1}$.

Шаг 5. Используя последнюю полученную оценку $\hat{\rho}$, применяем авторегрессионное преобразование с поправкой Прайса-Винстена и находим МНК-оценки параметров $\beta_0, \beta_1, \ldots, \beta_m$ исходной модели.

4.2 Решение типового примера

Пример 4.1 Проверим на автокорреляцию остатков модели множественной линейной регрессии из примера 1.1.

Решение. Оцененная модель есть

$$y_t = 17{,}541 - 0{,}582x_{t1} + 0{,}545x_{t2} + 0{,}120x_{t3} - 22{,}466x_{t4} - 0{,}003x_{t5},$$

вектор остатков модели

$$e^T = (2{,}497\ -1{,}627\ 0{,}232\ 1{,}702\ -0{,}579\ -2{,}380\ -1{,}976\ 0{,}492\ 0{,}043\ 0{,}160\ 1{,}600\ 1{,}809\ -1{,}501\ -0{,}912\ 0{,}441)^T.$$

1 Анализ графика остатков. На рисунке 4.2 изображен точечный график остатков e_t от t, $t = 1, \ldots, n$.

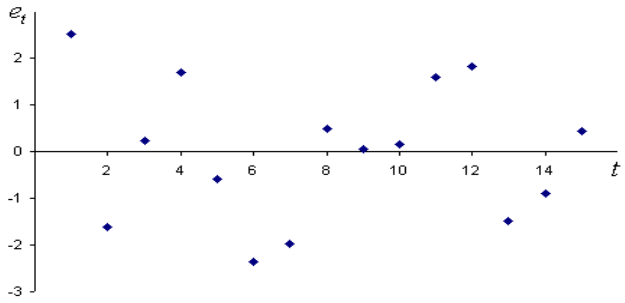

Рисунок 4.2 – График остатков модели

По виду графика можно предположить наличие положительной автокорреляции в остатках модели.

2 Метод рядов. Последовательность знаков компонент вектора e^T есть

(+) (–) (+ +) (– – –) (+ + + + +) (– –) (+).

Общее количество знаков «+» равно $n^+ = 9$, знаков «–» равно $n^- = 6$, количество рядов $k = 7$. По таблицам критических значений количества рядов [3] на уровне значимости $\alpha = 0{,}05$ определим нижнюю k_1 и верхнюю k_2 границы: $k_1 = 4$, $k_2 = 13$. В результате имеем $4 < 7 < 13$, что говорит об отсутствии автокорреляции.

3 Критерий Дарбина-Уотсона. Выдвигается нулевая гипотеза H_0: $\rho_1 = 0$ против альтернативной H_1: $\rho_1 \neq 0$. Для проверки нулевой гипотезы вычислим значение *DW*-статистики по формуле (4.4). Промежуточные расчеты представлены в таблице 4.1.

Значение статистики Дарбина-Уотсона равно $DW = 52{,}789 / 31{,}116 = 1{,}697$. Нижняя и верхняя границы d_L и d_U интервала критических значений DW для $n = 15$, $m = 5$, $\alpha = 0{,}05$ равны $d_L = 0{,}562$ и $d_U = 2{,}220$ [3]. Поскольку $d_L < DW < d_U$, то вопрос о принятии или отклонении гипотезы H_0 об отсутствии автокорреляции 1-го порядка остается открытым.

Таблица 4.1 – Промежуточные расчёты для *DW*-статистики

y_t	$\hat{y}_t$	e_t	e_{t-1}	e_t^2	$(e_t - e_{t-1})^2$
2,488	– 0,009	2,497	–	6,235	–
2,916	4,543	– 1,627	2,497	2,648	17,009
2,981	2,749	0,232	– 1,627	0,054	3,457
16,200	14,498	1,702	0,232	2,898	2,162
6,332	6,911	– 0,579	1,702	0,335	5,206
3,069	5,449	– 2,380	– 0,579	5,663	3,242
0,476	2,452	– 1,976	– 2,380	3,904	0,163
2,035	1,543	0,492	– 1,976	0,242	6,088
3,143	3,100	0,043	0,492	0,002	0,201
4,019	3,859	0,160	0,043	0,025	0,014
5,089	3,489	1,600	0,160	2,558	2,074
7,826	6,017	1,809	1,600	3,273	0,044
1,069	2,570	– 1,501	1,809	2,252	10,953
12,188	13,100	– 0,912	– 1,501	0,832	0,346
4,395	3,954	0,441	– 0,912	0,194	1,831

4 Критерий Бреуша-Годфри. Исследуем остатки модели на автокорреляцию до 2-го порядка ($q = 2$). Для этого построим зависимость e_t от объясняющих переменных, а также лаговых переменных e_{t-1} и e_{t-2}:

$$e_t = \lambda_0 + \lambda_1 x_{t1} + \lambda_2 x_{t2} + \lambda_3 x_{t3} + \lambda_4 x_{t4} + \lambda_5 x_{t5} + \delta_1 e_{t-1} + \delta_2 e_{t-2} + \eta_t.$$

Вспомогательная модель регрессии имеет вид

$$e_t = -5{,}345 - 0{,}156 x_{t1} + 0{,}013 x_{t2} + 0{,}02 x_{t3} + 4{,}387 x_{t4} + 0{,}025 x_{t5} + 0{,}341 e_{t-1} - 0{,}406 e_{t-2},$$

$$R^2 = 0{,}337.$$

Таблица 4.2 – Вычислительная процедура Кохрейна-Оркатта

	Нулевая итерация						Первая итерация						Вторая итерация					
№	Y	X_1	X_2	X_3	X_4	X_5	Y^*	X_1^*	X_2^*	X_3^*	X_4^*	X_5^*	Y^*	X_1^*	X_2^*	X_3^*	X_4^*	X_5^*
1	2,488	2,500	7,058	21,398	1,001	5,950	2,485	2,497	7,050	21,373	1,000	5,943	2,485	2,497	7,049	21,370	1,000	5,942
2	2,916	3,800	16,666	31,981	1,054	8,000	2,795	3,678	16,322	30,938	1,005	7,710	2,757	3,640	16,215	30,612	0,990	7,619
3	2,981	3,500	8,336	46,905	1,018	16,409	2,839	3,315	7,524	45,347	0,967	16,019	2,796	3,259	7,275	44,875	0,951	15,902
4	16,200	1,588	23,867	57,929	0,980	18,742	16,055	1,417	23,461	55,644	0,930	17,942	16,011	1,367	23,346	54,952	0,916	17,698
5	6,332	0,145	13,531	14,274	0,870	32,008	5,543	0,068	12,368	11,451	0,822	31,095	5,298	0,046	12,010	10,603	0,808	30,821
6	3,069	4,766	13,757	60,395	1,069	11,910	2,760	4,759	13,098	59,700	1,027	10,350	2,676	4,758	12,909	59,525	1,014	9,876
7	0,476	2,554	10,950	30,637	1,034	3,907	0,326	2,322	10,280	27,694	0,982	3,327	0,284	2,249	10,080	26,784	0,966	3,169
8	2,035	1,827	8,726	47,366	1,128	7,743	2,012	1,703	8,192	45,873	1,078	7,553	2,007	1,667	8,036	45,451	1,063	7,502
9	3,143	5,993	6,914	64,609	0,999	4,587	3,044	5,904	6,489	62,301	0,944	4,210	3,013	5,878	6,364	61,602	0,928	4,095
10	4,019	2,231	12,193	33,465	1,025	3,552	3,866	1,939	11,856	30,317	0,976	3,329	3,819	1,849	11,757	29,367	0,962	3,264
11	5,089	4,600	5,082	78,016	1,044	11,013	4,893	4,491	4,488	76,385	0,994	10,840	4,834	4,462	4,307	75,923	0,979	10,789
12	7,826	0,640	15,455	22,871	0,992	11,143	7,578	0,416	15,207	19,070	0,941	10,606	7,503	0,347	15,139	17,905	0,926	10,441
13	1,069	2,777	8,359	41,042	1,015	8,466	0,688	2,746	7,606	39,928	0,967	7,923	0,572	2,739	7,374	39,637	0,952	7,761
14	12,188	– 6,800	5,251	53,240	0,783	17,718	12,136	– 6,935	4,844	51,240	0,734	17,306	12,125	– 6,977	4,728	50,631	0,719	17,185
15	4,395	1,894	8,743	32,294	0,939	8,098	3,801	2,225	8,487	29,700	0,901	7,235	3,616	2,331	8,413	28,919	0,890	6,971
$\hat{\beta}$	17,533	– 0,581	0,545	0,120	– 22,459	– 0,002	12,378	– 0,661	0,526	0,116	–17,574	0,017	10,680	– 0,688	0,521	0,116	– 15,902	0,024
t_β	1,634	– 2,145	4,846	4,070	– 2,053	– 0,026	1,152	– 2,348	4,581	3,790	– 1,563	0,176	1,008	– 2,434	4,522	3,736	– 1,421	0,247
Pv	0,137	0,060	0,001	0,003	0,070	0,980	0,279	0,043	0,001	0,004	0,153	0,864	0,340	0,038	0,001	0,005	0,189	0,811
R^2	0,878						0,863						0,859					
F	12,939						11,337						10,977					
Pv	0,0007						0,001						0,001					
S	1,859						1,973						2,003					
DW	1,697						1,694						1,7					
$\hat{\rho}$	0,049						0,015						0,006					

Выдвигается нулевая гипотеза H_0: $\rho_1 = \rho_2 = 0$ против альтернативной: H_1: $\rho_i \neq 0$, $i = 1, 2$. Автокорреляция до 2-го порядка признается незначимой, поскольку

$$BG = (15 - 2)\cdot 0{,}337 = 4{,}381 < \chi^2_{1-0{,}05}(2) = 5{,}991,$$

$$Pv = \text{ХИ2.РАСП.ПХ}(4{,}381;2) = 0{,}11 > 0{,}05.$$

5 Корректировка автокорреляции методом Кохрейна-Оркатта. Согласно тесту рядов автокорреляция отсутствует. Тест Дарбина-Уотсона обозначил неопределенность относительно наличия в остатках автокорреляции 1-го порядка. В качестве примера продемонстрируем процедуру Кохрейна-Оркатта смягчения возможной автокорреляции. Зададим степень точности $\varepsilon = 0{,}01$. По остаткам регрессии оценим спецификацию AR(1) $e_t = 0{,}049e_{t-1}$, $t = 2, \ldots, 15$. Тогда $\hat{\rho}^0 = 0{,}049$. Выполним преобразование (4.13), результаты запишем в таблицу 4.2 (первая итерация). По преобразованным выборочным данным вычислим МНК-оценки коэффициентов, статистику *DW*, оценку $\hat{\rho}^1$. Так как $|\hat{\rho}^1 - \hat{\rho}^0| > 0{,}01$, продолжим итерационный процесс. Результаты второй итерации запишем в таблицу 4.2.

Заданная точность оценки достигнута, так как | ρ

5.1 Краткие теоретические сведения

5.1.1 Проблема гетероскедастичности

Гетероскедастичность (heteroskedasticity) – это предположение о неоднородности дисперсий случайных ошибок модели регрессии, то есть нарушение предпосылки ξ.3 Гаусса-Маркова. В качестве оценок неизвестных дисперсий σ_t^2 обычно используются квадраты остатков e_t^2.

Причинами гетероскедастичности могут быть неоднородность исследуемых объектов или характер наблюдений. *Последствия* гетероскедастичности: 1) оценки нормальной линейной модели регрессии остаются несмещенными и состоятельными, но теряется эффективность; 2) дисперсии оценок коэффициентов модели являются смещенными; 3) возможно неверное определение границ доверительных интервалов коэффициентов модели и прогнозных значений *Y*. Гетероскедастичность может быть: 1) *истинной* (обусловлена непостоянством дисперсии случайных ошибок, ее зависимостью от различных переменных); 2) *ложной* (обусловлена ошибочной спецификацией модели). Истинная гетероскедастичность возникает в пространственных данных при зависимости масштаба изменений от X_j, $j = 1, \ldots, m$ (наиболее распространенный случай: дисперсия растет с ростом переменной X_j). Во временных рядах она возникает тогда, когда зависимая переменная *Y* имеет большой интервал качественно неоднородных значений или высокий темп изменения.

5.1.2 Обнаружение гетероскедастичности

Гетероскедастичность может быть обнаружена посредством анализа графиков остатков модели и тестированием моделей гетероскедастичности.

Пусть оценивается модель

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_m x_{tm} + \xi_t,$$

определяются $e_t = y_t - \hat{y}_t$ – остатки модели (оценки ошибки ξ_t), где y_t, $\hat{y}_t$ – соответственно наблюдаемые и модельные значения зависимой переменной *Y*, $t = 1, \ldots, n$.

Анализ графиков остатков. Предположим, что дисперсии ошибок σ_t^2 связаны со значениями x_{tj} переменной X_j. По оси абсцисс откладываются значения x_{tj}, по оси ординат остатки e_t (или квадраты e_t^2), $j = 1, \ldots, m$, $t = 1, \ldots, n$. Анализ точечного графика (x_{tj}, e_t) позволяет выдвинуть предположение о наличии (отсутствии) функциональных зависимостей. График служит отправной точкой в исследовании гетероскедастичности.

Модели гетероскедастичности. *Модель межгрупповой гетероскедастичности.* Пусть гетероскедастичность обусловлена различными

условиями функционирования (временные ряды) или совместным анализом различных типов объектов (пространственные данные).

Предположения: задано число g интервалов постоянства дисперсии, $g \geq 2$. Разобьем всю совокупность данных на подвыборки, соответствующие интервалам постоянства, то есть $\{Y^{(l)}\}$, $\{X_j^{(l)}\}$, $l = 1, \ldots, g$. Построим модели регрессии для каждой подвыборки:

$$Y^{(l)} = \beta_0 + \beta_1 X_1^{(l)} + \ldots + \beta_m X_m^{(l)} + \xi^{(l)},$$

где $D(\xi^{(l)}) = \sigma_l^2$, $l = 1, \ldots, g$.

Выдвигается нулевая гипотеза H_0: $\sigma_1^2 = \ldots = \sigma_g^2$ против альтернативной H_1: $\sigma_1^2 \neq \ldots \neq \sigma_g^2$. Для проверки гипотезы используется критерий Бартлетта (Bartlett) со статистикой

$$V = 1/C \sum_{l=1}^{g} (n_l - 1) \ln(\bar{S}^2 / S_l^2) \sim \chi^2 (g - 1), \qquad (5.1)$$

$$C = 1 + \frac{1}{3(g-1)} \left(\sum_{l=1}^{g} \frac{1}{n_l - 1} - \frac{1}{n_1 + \ldots + n_g - g} \right), \quad \bar{S}^2 = \frac{\sum_{l=1}^{g} (n-1) S_l^2}{\sum_{l=1}^{g} (n_l - 1)}, \qquad (5.2)$$

где S_l^2 – выборочная несмещенная дисперсия остатков l-й подвыборки.

Гипотеза H_0 отклоняется на уровне значимости α, если $V \geq \chi^2_{1-\alpha}(g-1)$ или $Pv = P\{\chi^2(g-1) \geq V\} < \alpha$, где $\chi^2_{1-\alpha}(g-1)$ – квантиль распределения χ^2.

Тест ранговой корреляции Спирмена (Spearman). Предположение: дисперсии σ_t^2 монотонно изменяются с увеличением переменной X_j, то есть $|e_t|$ и x_{tj} коррелируют, $t = 1, \ldots, n$.

Пусть $R_{x_{tj}}$ – ранги значений x_{tj} (при этом не нарушается исходная нумерация данных), $R_{|e_t|}$ – ранги остатков $|e_t|$, $t = 1, \ldots, n$. *Выборочным коэффициентом ранговой корреляции Спирмена* называется величина

$$r^s_{e X_j} = \frac{(n^3 - n)/6 - \sum_{t=1}^{n} (R_{x_{tj}} - R_{|e_t|})^2 - T_{x_{tj}} - T_{|e_t|}}{\sqrt{\left((n^3 - n) - 2T_{x_{tj}}\right)\left((n^3 - n) - 2T_{|e_t|}\right)}\Big/6}. \qquad (5.3)$$

Здесь $T_{x_{tj}} = \sum_{k=1}^{m_j} ((n_{x_{tj}k})^3 - n_{x_{tj}k})/12$, m_j – число связок для j-й переменной X_j, $n_{x_{tj}k}$ – число рангов, входящих в данную связку. Аналогично для $T_{|e_t|}$.

Если связки отсутствуют, то $m_j = n$, $n_{x_{tj}k} = 1$, $T_{x_{tj}} = 0$ и

$$r^s_{e X_j} = 1 - \frac{6 \sum_{t=1}^{n} (R_{x_{tj}} - R_{|e_t|})^2}{n(n^2 - 1)}. \qquad (5.4)$$

Выдвигается нулевая гипотеза H_0: $\rho^{s}_{e X_j} = 0$ (отсутствие корреляции между фактором X_j и вектором остатков e) против альтернативной H_1: $\rho^{s}_{e X_j} \neq 0$. Для проверки гипотезы используется t-критерий со статистикой

$$t = \frac{r^{s}_{e X_j}\sqrt{n-2}}{\sqrt{1-(r^{s}_{e X_j})^2}} \sim T(n-2). \qquad (5.5)$$

Гипотеза H_0 отклоняется на уровне значимости α, если $|t| \geq t_{1-\alpha/2}(n-2)$ или $Pv = P\{t(n-2) \geq t\} < \alpha$, то есть гетероскедастичность присутствует.

Тест Голдфелда-Квандта (Goldfeld-Quandt). Предположения: 1) случайные ошибки ξ_t имеют нормальное распределение и автокорреляция остатков отсутствует; 2) дисперсия ошибок σ_t^2 связаны со значениями x_{tj} переменной X_j неубывающей функцией вида $\sigma_t^2 = \sigma^2 \cdot x_{tj}^2$, $t = 1, \ldots, n$.

Все наблюдения упорядочиваются по возрастанию переменной X_j. Упорядоченная выборка делится на три части так, чтобы в первой и третьей части было по n' элементов, $m < n' \leq n/2$. Для первой и третьей частей оцениваются две независимые модели множественной линейной регрессии:

$$Y^{(1)} = \beta_0^{(1)} + \beta_1^{(1)} X_1^{(1)} + \ldots + \beta_m^{(1)} X_m^{(1)} + \xi^{(1)},$$
$$Y^{(3)} = \beta_0^{(3)} + \beta_1^{(3)} X_1^{(3)} + \ldots + \beta_m^{(3)} X_m^{(3)} + \xi^{(3)}.$$

Выдвигается нулевая гипотеза H_0: $\sigma_1^2 = \sigma_3^2$ против альтернативной H_1: $\sigma_1^2 \neq \sigma_3^2$. Для проверки гипотезы используется GQ-критерий со статистикой

$$GQ = \max\{ESS^{(1)}, ESS^{(3)}\} / \min\{ESS^{(1)}, ESS^{(3)}\} \sim F(n'-m-1, n'-m-1), \qquad (5.6)$$

где $ESS^{(1)} = \sum_{t=1}^{n'} (e_t^{(1)})^2$, $ESS^{(3)} = \sum_{t=1}^{n'} (e_t^{(3)})^2$.

Гипотеза H_0 отклоняется, если $GQ \geq F_{1-\alpha}(n'-m-1,\ n'-m-1)$ или $Pv = P\{F(n'-m-1, n'-m-1) \geq GQ\} < \alpha$ на уровне значимости α, то есть гетероскедастичность присутствует.

Тест Глейзера (Glejser). Предположение: зависимость $|e_t|$ от значений x_{tj} переменной X_j имеет вид

$$|e_t| = \delta_0 + \delta_1 x_{tj}^{\gamma} + \nu_t, \quad \nu_t \sim N(0; \sigma^2), -2 \leq \gamma \leq 2. \qquad (5.7)$$

Оцениваются несколько моделей вида (5.7) при различных γ. Статистическая значимость коэффициента δ_1 означает наличие гетероскедастичности. Если для нескольких моделей получена значимая оценка δ_1, то характер гетероскедастичности определяется наиболее значимой.

Тест Парка (Park). Предположение: дисперсии ошибок связаны с переменной X_j зависимостью

$$\sigma_t^2 = \sigma^2 \cdot x_{tj}^{\delta} e^{\nu_t}, \quad \nu_t \sim N(0; \sigma^2). \qquad (5.8)$$

Прологарифмировав (5.8), получим

$$\ln \sigma_t^2 = \ln \sigma^2 + \delta \ln x_{tj} + \nu_t. \qquad (5.9)$$

По выборочным данным оценивается модель (5.9). Если коэффициент δ статистически значим, то гетероскедастичность присутствует.

Тест Уайта (White). Предположения: дисперсии ошибок модели связаны с переменными $X_1, X_2, \ldots, X_m$ зависимостью

$$\sigma_t^2 = f(x_{t1}, x_{t2}, \ldots, x_{tm}), \qquad (5.10)$$

где f – квадратичная функция.

Оценивается модель

$$e_t^2 = \delta_0 + \sum_{j=1}^{m} \delta_j x_{tj} + \sum_{j=1}^{m} \sum_{i=1}^{m} \delta_{ij} x_{ti} x_{tj} + \nu_t, \quad \nu_t \sim N(0; \sigma^2). \qquad (5.11)$$

Проверяется значимость вспомогательной модели регрессии (5.11). Выдвигается гипотеза H_0: $\delta_0 = \delta_1 = \ldots = \delta_m = \delta_{11} = \ldots = \delta_{mm} = 0$ против альтернативной H_1: $\delta_0 \neq \delta_1 \neq \ldots \neq \delta_m \neq \delta_{11} \neq \ldots \neq \delta_{mm} \neq 0$. Для проверки гипотезы используется U - критерий со статистикой

$$U = nR^2 \sim \chi^2 (Q - 1), \qquad (5.12)$$

где R^2, Q – коэффициент детерминации и число параметров модели (5.11).

Гипотеза H_0 отклоняется на уровне значимости α, если $U \geq \chi^2_{1-\alpha}(Q-1)$ или $P\nu = P\{\chi^2(Q-1) \geq U\} < \alpha$, то есть гетероскедастичность присутствует.

Тест Бреуш-Пагана-Годфри (Breusch-Pagan-Godfrey). Предположения: дисперсии ошибок модели связаны с p любыми переменными $X_{(1)}, X_{(2)}, \ldots, X_{(p)}$ зависимостью

$$\sigma_t^2 = \theta_0 + \theta_1 x_{t1} + \ldots + \theta_p x_{tp}. \qquad (5.13)$$

Оценивается модель

$$e_t^2/\sigma$$

5.1.3 Способы корректировки гетероскедастичности

При нарушении гомоскедастичности и наличии автокорреляции остатков обычно вместо традиционного МНК используется обобщенный МНК (ОМНК), который для случая устранения гетероскедастичности обычно называется методом взвешенных наименьших квадратов (МВНК).

Пусть модель множественной линейной регрессии имеет вид

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots + \beta_m x_{tm} + \xi_t,\ t = 1, \ldots, n, \qquad (5.16)$$

причем $\mathrm{cov}(\xi_t, \xi_t) = \sigma_t^2$, $\sigma_1^2 \neq \sigma_2^2 \neq \ldots \neq \sigma_n^2$.

Метод взвешенных наименьших квадратов при известных σ_t^2. Разделим наблюдаемые значения y_t и $x_{t1}, \ldots, x_{tm}$ на σ_t, $t = 1, \ldots, n$:

$$y_t/\sigma_t = \beta_0/\sigma_t + \beta_0 x_{t1}/\sigma_t + \ldots + \beta_m x_{tm}/\sigma_t + \xi_t/\sigma_t. \qquad (5.17)$$

Обозначим: $y_t^* = y_t/\sigma_t$, $z_t = 1/\sigma_t$, $x_{tj}^* = x_{tj}/\sigma_t$, $\nu_t = \xi_t/\sigma_t$. Тогда получим модель регрессии без свободного члена, но с дополнительной объясняющей переменной Z и с «преобразованной» случайной ошибкой ν:

$$y_t^* = \beta_0 z_t + \beta_1 x_{t1}^* + \ldots + \beta_m x_{tm}^* + \nu_t,\ t = 1, \ldots, n. \qquad (5.18)$$

Для модели (5.18) имеем

$$M(\nu_t) = M(\xi_t/\sigma_t) = M(\xi_t)/\sigma_t = 0,\ D(\nu_t) = D(\xi_t/\sigma_t) = D(\xi_t)/\sigma_t^2 = \sigma_t^2/\sigma_t^2 = 1,$$

то есть для модели (5.18) выполняются предпосылки Гаусса-Маркова, и МНК-оценки коэффициентов являются несмещенными, состоятельными и эффективными. Оценив их, перейдем к исходной модели (5.16).

Метод взвешенных наименьших квадратов при неизвестных σ_t^2. По корреляционному полю σ_t^2 и x_{tk} делается предположение о виде их функциональной зависимости. Рассмотрим случаи.

1 Дисперсии σ_t^2 пропорциональны x_{tk}

$$\sigma_t^2 = \sigma^2 x_{tk},\ t = 1, \ldots, n. \qquad (5.19)$$

Тогда уравнение (5.16) делится на $\sqrt{x_{tk}}$ и вводятся переменные

$$y_t^* = y_t/\sqrt{x_{tk}},\ x_{tj}^* = x_{tj}/\sqrt{x_{tk}},\ j \neq k,\ \nu_t = \xi_t/\sqrt{x_{tk}}. \qquad (5.20)$$

В результате имеем нормальную модель множественной линейной регрессии, МНК-оценки которой дают оценки исходной модели. После определения оценок коэффициентов возвращаются к исходной модели.

2 Дисперсии σ_t^2 пропорциональны x_{tk}^2

$$\sigma_t^2 = \sigma^2 x_{tk}^2,\ t = 1, \ldots, n. \qquad (5.21)$$

Тогда уравнение (5.16) делится на x_{tk} и вводятся переменные

$$y_t^* = y_t/x_{tk},\ x_{tj}^* = x_{tj}/x_{tk},\ \nu_t = \xi_t/x_{tk}. \qquad (5.22)$$

В результате имеем нормальную модель множественной линейной регрессии, МНК-оценки которой дают оценки исходной модели. После определения оценок коэффициентов возвращаются к модели (5.16).

5.2 Решение типового примера

Пример 5.1 Для построенной в примере 1.1 модели множественной линейной регрессии требуется: 1) проверить наличие (отсутствие) гетероскедастичности; 2) скорректировать гетероскедастичность методом взвешенных наименьших квадратов.

Решение. Оцененная модель имеет вид

$$y_t = \underset{(1,635)}{17,541} - \underset{(-2,153)}{0,582}x_{t1} + \underset{(4,846)}{0,545}x_{t2} + \underset{(4,075)}{0,120}x_{t3} - \underset{(-2,054)}{22,466}x_{t4} - \underset{(-0,030)}{0,003}x_{t5}, \qquad (5.23)$$

$$t_{\text{кр}} = 2,262,\ R^2 = 0,878,\ R^2_{adj} = 0,810,\ F = 12,939,\ Pv = 0,0007.$$

Вектор остатков модели есть

$$e^T = (2,497\ -1,627\ 0,232\ 1,702\ -0,579\ -2,380\ -1,976\ 0,492\ 0,043\ 0,160\ 1,600\ 1,809\ -1,501\ -0,912\ 0,441)^T.$$

1 Анализ графиков остатков. Точечные графики остатков e_t от значений x_{tj} переменных X_j, $j = 1, \ldots, 5$, $t = 1, \ldots, 15$, представлены на рисунке 5.1. Колеблемость остатков практически для всех переменных одинакова, поэтому явно гетероскедастичность выявить сложно. В этом случае целесообразно протестировать модели гетероскедастичности.

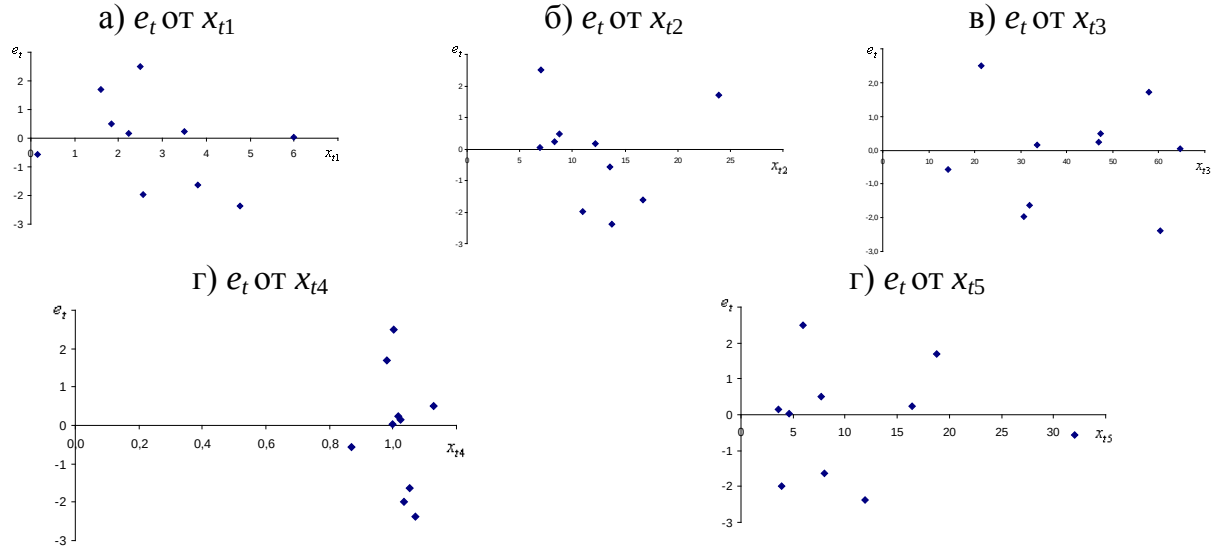

Рисунок 5.1 – Графики зависимости остатков e_t от объясняющих переменных

2 Тест ранговой корреляции Спирмена. Пусть, например, величины $|e_t|$ монотонно изменяются с увеличением x_{t5} переменной X_5, $t = 1, \ldots, 15$. В таблице 5.1 представлены промежуточные расчеты для рангового коэффициента корреляции Спирмена, $r^s_{e\,X_5} = 1 - 6 \cdot 492/15(15^2 - 1) = 0,121$.

Таблица 5.1 – Расчет рангового коэффициента корреляции Спирмена

x_{t5}	R_{Xt5}	$\|e_t\|$	$R_{\|et\|}$	$(R_{Xt5} - R_{\|et\|})$	$(R_{Xt5} - R_{\|et\|})^2$
5,950	12	2,497	1	11	121
8,000	10	1,627	6	4	16
16,409	4	0,232	13	–9	81
18,742	2	1,702	5	–3	9
32,008	1	0,579	10	–9	81
11,910	5	2,380	2	3	9
3,907	14	1,976	3	11	121
7,743	11	0,492	11	0	0
4,587	13	0,043	15	–2	4
3,552	15	0,160	14	1	1

Выдвигается нулевая гипотеза H_0: $\rho^s_{e\,X_5} = 0$ против альтернативной H_1: $\rho^s_{e\,X_5} \neq 0$. Значение t- статистики и квантиля на уровне значимости $\alpha = 0{,}05$ равны соответственно

$$t = 0{,}121 \cdot \sqrt{15-2} / \sqrt{1-0{,}121^2} = 0{,}441,\ t_{\text{кр}} = t_{1-0{,}05/2}(15-2) = 2{,}160.$$

Гипотеза H_0 не отклоняется, так как $|t| < t_{\text{кр}}$ ($Pv = 0{,}67 > 0{,}05$), то есть гетероскедастичность отсутствует.

3 Тест Голдфельда-Квандта. Пусть на изменчивость дисперсии ошибок влияет переменная X_2. Упорядочим все выборочные данные по величине модуля переменной X_2, $t = 1, \ldots, 15$ (таблица 5.2). И пусть число наблюдений в «частных» регрессиях равно 7, $n' = 7$ (условия теста Голдфельда-Квандта $m < n' \leq n/2$ и Гаусса-Маркова $n' > m + 1$ выполняются).

Таблица 5.2 – Упорядочение данных для теста Голдфельда-Квандта

№	Y	X_1	X_2	X_3	X_4	X_5
1	5,089	4,600	5,082	78,016	1,044	11,013
2	12,188	– 6,800	5,251	53,240	0,783	17,718
3	3,143	5,993	6,914	64,609	0,999	4,587
4	2,488	2,500	7,058	21,398	1,001	5,950
5	2,981	3,500	8,336	46,905	1,018	16,409
6	1,069	2,777	8,359	41,042	1,015	8,466
7	2,035	1,827	8,726	47,366	1,128	7,743
8	4,395	1,894	8,743	32,294	0,939	8,098
9	0,476	2,554	10,950	30,637	1,034	3,907
10	4,019	2,231	12,193	33,465	1,025	3,552
11	6,332	0,145	13,531	14,274	0,870	32,008
12	3,069	4,766	13,757	60,395	1,069	11,910
13	7,826	0,640	15,455	22,871	0,992	11,143
14	2,916	3,800	16,666	31,981	1,054	8,000
15	16,200	1,588	23,867	57,929	0,980	18,742

Оценки частных моделей регрессий

$$y_t^{(1)} = 15{,}245 - 0{,}477x_{t1} - 0{,}745x_{t2} + 0{,}047x_{t3} - 7{,}915x_{t4} + 0{,}070x_{t5},\ R^2 = 0{,}979,\ ESS^{(1)} = 1{,}814;$$

$y_t^{(3)} = 10{,}460 - 2{,}271x_{t1} + 0{,}677x_{t2} + 0{,}141x_{t3} + 5{,}569x_{t4} + 0{,}034x_{t5}$, $R^2 = 0{,}991$, $ESS^{(3)} = 1{,}418$.

Выдвигается нулевая гипотеза H_0: $\sigma_1^2 = \sigma_3^2$ против альтернативной H_1: $\sigma_1^2 \neq \sigma_3^2$. Значение GQ-статистики равно $GQ = 1{,}814 / 1{,}418 = 1{,}280$. На уровне значимости $\alpha = 0{,}05$ имеем

$$F_{\text{кр}} = F_{1-0{,}05}(7 - 5 - 1, 7 - 5 - 1) = 161{,}448.$$

Гипотеза H_0 не отклоняется, так как $GQ < F_{\text{кр}}$ ($Pv = 0{,}46 > 0{,}05$), то есть гетероскедастичность отсутствует.

4 Тест Глейзера. Предположим, что остатки e_t модели связаны со значениями x_{t3} переменной X_3 зависимостью (5.7), $t = 1, \ldots, 15$.

Полученные для $\gamma = \pm 2, \pm 1, \pm 0{,}5$ оценки коэффициентов моделей вида (5.7), t – статистики, коэффициенты детерминации и F - статистики представлены в таблице 5.3. Видно, что для всех γ коэффициент δ_1 статистически незначим ($t_{\text{кр}} = 2{,}16$), поэтому гетероскедастичность отсутствует.

Таблица 5.3 – Оценки моделей регрессий для теста Глейзера

γ	δ_0	$t(\delta_0)$	δ_1	$t(\delta_1)$	R^2	F
– 2	1,188	4,000	8,234	0,044	0,0 001	0,002
– 1	1,098	2,241	3,421	0,226	0,004	0,051
– 0,5	0,941	0,988	1,548	0,276	0,006	0,076
0,5	1,468	1,389	– 0,043	– 0,263	0,005	0,069
1	1,310	2,238	– 0,003	– 0,209	0,003	0,044
2	1,215	3,308	– 0,00 001	– 0,063	0,0 003	0,004

5 Тест Уайта. В предпосылках теста требуется оценить вспомогательную модель регрессии

$$e_t^2 = \delta_0 + \sum_{j=1}^{5} \delta_j x_{tj} + \sum_{j=1}^{5} \sum_{i=1}^{5} \delta_{ij} x_{ti} x_{tj} + \nu_t, \; \nu_t \sim N(0; \sigma^2). \qquad (5.24)$$

По условию примера 1.1 число наблюдений равно $n = 15$. Для МНК-оценки модели (5.24) не выполняется условие Гаусса-Маркова X.2, то есть $n > m + 1$, поскольку число параметров модели (5.24) равно $m = 20$. Поэтому для исходной модели примера 1.1 тест Уайта не применим. С целью демонстрации теста по данным таблицы 1.2 построим модель множественной линейной регрессии зависимости Y от X_2 и X_4, и протестируем остатки регрессии на гетероскедастичность.

Оцененная модель имеет вид

$$\underset{}{y_t} = \underset{(3{,}510)}{32{,}390} + \underset{(2{,}991)}{0{,}455}x_{t2} - \underset{(-3{,}514)}{32{,}547}x_{t4}, \qquad (5.25)$$

$t_{\text{кр}} = t_{1-0{,}05/2}(12) = 2{,}179$, $R^2 = 0{,}615$, $F = 9{,}602$, $F_{\text{кр}} = F_{1-0{,}05}(2;12) = 3{,}885$.

Вектор остатков

$$e^T = (-0{,}532 \; -2{,}748 \; -0{,}067 \; 4{,}853 \; -3{,}895 \; -0{,}784 \; -3{,}24 \; 2{,}389 \; 0{,}123 \; -0{,}555 \; 4{,}367 \; 0{,}694 \; -2{,}087 \; 2{,}894 \; -1{,}41)^T.$$

МНК-оценка вспомогательной модели регрессии вида (5.11) есть

$$e_t^2 = 202{,}670 + 8{,}697x_2 - 487{,}786x_4 + 0{,}125\, x_2^2 - 11{,}558\, x_2\, x_4 + 303{,}458\, x_4^2,$$
$$R^2 = 0{,}658.$$

Выдвигается нулевая гипотеза H_0: $\delta_0 = \delta_1 = \delta_2 = \delta_{11} = \delta_{12} = \delta_{22} = 0$ против альтернативной H_1: $\delta_0 \neq \delta_1 \neq \delta_2 \neq \delta_{11} \neq \delta_{12} \neq \delta_{22} \neq 0$. Значение статистики равно $U = 15 \cdot 0{,}658 = 9{,}872$. На уровне значимости $\alpha = 0{,}05$ имеем $\chi^2_{кр} = \chi^2_{1-0{,}05}(5-1) = 9{,}488$. Гипотеза H_0 отклоняется, так как $U > \chi^2_{кр}$ ($Pv = 0{,}043 < 0{,}05$), то есть гетероскедастичность присутствует.

6 Тест Парка. Пусть дисперсии ошибок модели связаны с переменной X_4 зависимостью (5.8). Для построения вспомогательной модели (5.9) составим расчетную таблицу 5.4.

Таблица 5.4 – Расчетная таблица для теста Парка

e^2	$\ln(e^2)$	$\ln(x_4)$
6,235	1,830	0,001
2,648	0,974	0,053
0,054	– 2,921	0,018
2,898	1,064	– 0,020
0,335	– 1,092	– 0,139
5,663	1,734	0,067
3,904	1,362	0,033
0,242	– 1,420	0,120
0,002	– 6,303	– 0,001
0,025	– 3,671	0,025
2,558	0,939	0,043
3,273	1,186	– 0,008
2,252	0,812	0,015
0,832	– 0,184	– 0,245
0,194	– 1,638	– 0,063

Вспомогательная модель принимает вид

$$\ln(e^2) = \underset{(-0{,}757)}{-\,0{,}478} + \underset{(0{,}223)}{1{,}644}\ln x_4,\ R^2 = 0{,}004,\ F = 0{,}05.$$

Значение t-статистики коэффициента δ при объясняющей переменной $\ln x_4$ вспомогательной модели

$$|\,t\,| = 0{,}223 < 2{,}160 = t_{кр} = t_{1-0{,}05/2}\,(15-1-1),\ Pv = 0{,}827 > 0{,}05,$$

поэтому коэффициент δ признается статистически незначимым, то есть гетероскедастичность отсутствует.

7 Тест Бреуш-Пагана-Годфри. Предположим, что дисперсии ошибок модели зависят от переменных $X_1, \ldots, X_5$, то есть $\sigma_t^2 = \theta_0 + \theta_1 x_{t1} + \ldots + \theta_5 x_{t5}$. Оценка дисперсии остатков равна σ

Выдвигается нулевая гипотеза H_0: $\theta_0 = \theta_1 = \ldots = \theta_5 = 0$ против альтернативной H_1: $\theta_0 \neq \theta_1 \neq \ldots \neq \theta_5 \neq 0$. Значение *BP*-статистики равно $BP = 1{,}809/2 = 0{,}904$, $\chi^2_{кр} = \chi^2_{1-0{,}05}(5) = 11{,}071$ на уровне значимости $\alpha = 0{,}05$. Гипотеза H_0 не отклоняется, так как $BP < \chi^2_{кр}$ ($Pv = 0{,}97 > 0{,}05$), то есть гетероскедастичность отсутствует.

8 Корректировка гетероскедастичности МВНК. Проиллюстрируем МВНК на примере оцененной модели (5.25), в которой согласно тесту Уайта присутствует гетероскедастичность. Предположим, что между переменной X_2 и остатками модели имеется зависимость вида (5.21). Тогда по формулам (5.22) получим значения новых переменных (таблица 5.5).

По преобразованным данным модель множественной линейной регрессии имеет вид

$$y_t = \underset{(0{,}080)}{0{,}016} + \underset{(6{,}170)}{35{,}827}\, x_{t2}^{*} - \underset{(-5{,}160)}{31{,}555}\, x_{t4}^{*},$$

$t_{кр} = t_{1-0{,}05/2}(12) = 2{,}179$, $R^2 = 0{,}795$, $F = 23{,}209$, $F_{кр} = F_{1-0{,}05}(2;12) = 3{,}885$.

Ожидается, что полученная модель не имеет гетероскедастичности.

Таблица 5.5 – Значения новых переменных для МВНК

Y	X_2	X_4	$Y^* = Y/X_2$	$X_2^* = 1/X_2$	$X_4^* = X_4/X_2$
2,488	7,058	1,001	0,353	0,142	0,142
2,916	16,666	1,054	0,175	0,060	0,063
2,981	8,336	1,018	0,358	0,120	0,122
16,200	23,867	0,980	0,679	0,042	0,041
6,332	13,531	0,870	0,468	0,074	0,064
3,069	13,757	1,069	0,223	0,073	0,078
0,476	10,950	1,034	0,043	0,091	0,094
2,035	8,726	1,128	0,233	0,115	0,129
3,143	6,914	0,999	0,455	0,145	0,144
4,019	12,193	1,025	0,330	0,082	0,084
5,089	5,082	1,044	1,001	0,197	0,205
7,826	15,455	0,992	0,506	0,065	0,064
1,069	8,359	1,015	0,128	0,120	0,121
12,188	5,251	0,783	2,321	0,190	0,149
4,395	8,743	0,939	0,503	0,114	0,107

5.3 Задания для лабораторной работы 5

По статистическим данным (таблица 1.6) и результатам исследований лабораторной работы 1 требуется:

1) проверить модель на наличие гетероскедастичности остатков (для переменной X_1 – тест ранговой корреляции Спирмена, для переменной X_2 – тест Голдфелда-Квандта, для переменной X_3 – тест Глейзера, для переменной X_4 – тест Парка, для переменных X_1, X_2 – тест Уайта, для всех переменных – тест Бреуш-Пагана-Голдфри);

2) скорректировать гетероскедастичность методом взвешенных наименьших квадратов.

Литература

1 Айвазян, С. А. Прикладная статистика. Основы эконометрики: учебное пособие: в 2 т. / С. А. Айвазян, В. С. Мхитарян. – М. : ЮНИТИ, 2002.

2 Бабешко, Л. О. Основы эконометрического моделирования: учебное пособие / Л. О. Бабешко. – М. : КомКнига, 2007.

3 Бородич, С. А. Вводный курс эконометрики: учебное пособие / С. А. Бородич. – Минск : БГУ, 2000.

4 Доугерти, К. Введение в эконометрику: учебное пособие / К. Доугерти. – М. : Инфра-М, 2004.

5 Елисеева, И. И. Эконометрика: учебное пособие / И. И. Елисеева. – М.: Финансы и статистика, 2004.

6 Магнус, Я. Р. Эконометрика: учебное пособие / Я. Р. Магнус [и др.]. – М. : Дело, 2004.

7 Практикум по эконометрике: учебное пособие / И. И. Елисеева [и др.], под ред. И. И. Елисеевой. – М. : Финансы и статистика, 2008.

8 Харин, Ю. С. Эконометрическое моделирование: учебное пособие / В. И. Харин [и др.]. – Минск : БГУ, 2003.

9 Харин, Ю. С. Математические и компьютерные основы статистического моделирования и анализа данных. / Ю. С. Харин, В. И, Малюгин, М. С. Абрамович – Минск : БГУ, 2008.

Производственно-практическое издание

Марченко Лариса Николаевна,
Дудовская Юлия Евгеньевна,
Синюгина Юлия Васильевна

Эконометрика:
модель множественной линейной регрессии

Практическое пособие

Редактор *В. И. Шкредова*
Корректор *В. В. Калугина*

Подписано в печать 11.11.2016. Формат 60x84 1/16.
Бумага офсетная. Ризография. Усл. печ. л. 2,8.
Уч.-изд. л. 3,1. Тираж 25 экз. Заказ 638.

Издатель и полиграфическое исполнение:
учреждение образования
«Гомельский государственный университет
имени Франциска Скорины».
Свидетельство о государственной регистрации издателя,
изготовителя, распространителя печатных изданий № 1/87 от 18.11.2013.
Специальное разрешение (лицензия) № 02330 / 450 от 18.12.2013.
Ул. Советская, 104, 246019, г. Гомель.