**Владимирский государственный университет**

**А. М. ГУБЕРНАТОРОВ**

# АНАЛИЗ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ЭКОНОМЕТРИЧЕСКИХ МЕТОДОВ

**Учебное пособие**

**Владимир 2026**

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Владимирский государственный университет
имени Александра Григорьевича и Николая Григорьевича Столетовых»

А. М. ГУБЕРНАТОРОВ

# АНАЛИЗ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ЭКОНОМЕТРИЧЕСКИХ МЕТОДОВ

Учебное пособие

*Электронное издание*

Владимир 2026

ISBN 978-5-9984-1225-7

УДК 330.115(075.8)
ББК 65в631я73

Рецензенты:
Доктор экономических наук, профессор
зав. кафедрой бизнес-информатики и экономики
Владимирского государственного университета
имени Александра Григорьевича и Николая Григорьевича Столетовых
*И. Б. Тесленко*

Кандидат экономических наук, доцент
зав. кафедрой экономики и финансов Финансового университета
при Правительстве Российской Федерации (Владимирский филиал)
*Д. В. Кузнецов*

**Губернаторов, А. М.**

Анализ данных с использованием эконометрических методов [Электронный ресурс] : учеб. пособие / А. М. Губернаторов ; Владим. гос. ун-т им. А. Г. и Н. Г. Столетовых. – Владимир : Изд-во ВлГУ, 2026. – 128 с. – ISBN 978-5-9984-1225-7. – Электрон. дан. (2,51 Мб). – 1 электрон. опт. диск (CD-ROM). – Систем. требования: Intel от 1,3 ГГц ; Windows XP/7/8/10 ; Adobe Reader ; дисковод CD-ROM. – Загл. с титул. экрана.

Рассматриваются современные методы анализа данных, основанные на эконометрическом подходе. Подробно освещены вопросы спецификации моделей, оценки параметров (методами наименьших квадратов, максимального правдоподобия и др.), тестирования гипотез и прогнозирования социально-экономических явлений с использованием специализированного программного обеспечения.

Предназначено для студентов направлений подготовки 01.03.05 – Статистика, 38.03.05 – Бизнес-информатика и других экономических направлений всех форм обучения, аспирантов, а также аналитиков и специалистов, занимающихся обработкой и моделированием экономических процессов.

Рекомендовано для формирования профессиональных компетенций в соответствии с ФГОС ВО.

Ил. 15. Табл. 14. Библиогр.: 10 назв.

ISBN 978-5-9984-1225-7 

## ОГЛАВЛЕНИЕ

# ПРЕДИСЛОВИЕ

В современную эпоху, насыщенную информацией, умение извлекать ценные знания из данных и принимать на их основе обоснованные решения становится крайне важным навыком во всех профессиональных сферах – от экономики и финансов до маркетинга и социологии. В этом контексте на помощь приходит эконометрика – наука, предоставляющая исследователю мощный набор статистических инструментов для измерения, анализа и моделирования экономических и социальных процессов.

Учебное пособие «Анализ данных с использованием эконометрических методов» предназначено провести читателя от первых шагов в работе с данными до построения сложных многофакторных моделей и правильной интерпретации результатов. Цель пособия − не только познакомить с теоретическими аспектами эконометрики, но и развить практические навыки, необходимые для самостоятельного проведения прикладных исследований.

В отличие от чисто теоретических курсов в представленном издании автор делает особый акцент на практическом применении методов. Структура построена по принципу «от простого к сложному»: начиная с описательной статистики и проверки гипотез, переходя к регрессионному анализу, моделям временных рядов и современным подходам к обработке данных. Каждая глава сопровождается примерами реализации методов в популярных статистических программах (таких как Excel и R), что позволяет обучающемуся не только понять суть методов, но и сразу применять их на практике.

Пособие рассчитано на студентов экономических и социологических специальностей, магистрантов, аспирантов, а также аналитиков-практиков, желающих углубить свои знания в области количественного анализа данных и повысить качество своих исследований.

# Глава 1. ЭКСТРАПОЛЯЦИОННЫЕ МЕТОДЫ ПРОГНОЗИРОВАНИЯ

Экстраполяция – наиболее распространённый и хорошо разработанный метод среди всех методов прогнозирования. Его суть заключается в анализе динамических (временных) рядов, которые отражают изменение определенного показателя (параметра) во времени, выявлении существующих тенденций и их продолжении в будущее. Любое предполагаемое будущее состояние показателя рассматривается как результат предыдущих изменений. В настоящее время экстраполяционные методы широко применяются в стратегическом управлении и планировании. Основой их возможности является предположение о закономерности изменений различных показателей и инертности технико-экономических процессов. Формирование уровней динамических рядов обусловлено совокупным воздействием множества длительно и краткосрочно действующих факторов, включая случайные влияния. Динамический ряд показателей может быть представлен в следующем виде:

$$y_t = \hat{y}_t + E_t, \qquad (1.1)$$

где $\hat{y}_t$ – тренд (детерменирующая неслучайная компонента);

$E_t$ – стохастическая случайная компонента (помеха), отражающая случайные колебания и имеющая нормативный закон распределения.

Особенность прогнозной экстраполяции заключается в предварительной обработке числового ряда, целью которой является снижение воздействия случайных факторов (уменьшение случайных отклонений точек ряда от предполагаемой плавной кривой тренда). Иначе говоря, эта обработка направлена на приближение исходных данных к линии тренда.

**Задача прогноза** состоит в определении вида экстраполирующих функций $\hat{y}_t$ и $E_t$ на основе исходных данных.

Промежуток времени, на который разрабатывается прогноз, называется *периодом упреждения прогнозов*, а максимально возможный период упреждения – *горизонтом прогнозирования*.

По форме упреждения различают *точечные и интервальные прогнозы*. В первом случае прогноз задаётся одним числом, во втором указывается интервал, к которому с определённой вероятностью принадлежит прогнозируемая величина.

При прогнозировании ведётся наблюдение за процессом (рис. 1.1) и вычисляется его будущее значение в упреждённой точке, оценивается математическое ожидание процесса, величина интервала, в которой с заданной вероятностью попадает будущее значение прогноза.

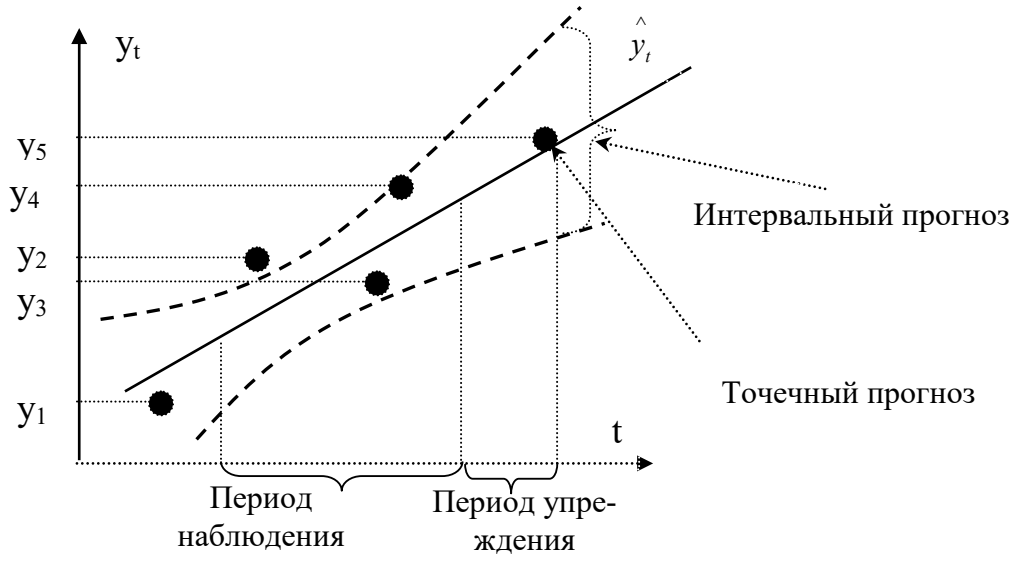



Рис. 1.1. Точечный и интервальный прогнозы

В данной лабораторной работе применяется алгоритм экстраполяции динамических рядов развития показателя прогнозируемого явления на основе ретроспективных данных с помощью методов статистического моделирования.

**Постановка задачи**

В лабораторной работе № 1 необходимо составить прогноз величины технико-экономического показателя $y_t$ производственной деятельности предприятия отрасли, предсказав его возможную величину на основе статистических данных о его изменении за несколько предыдущих лет.

Исходные данные для расчета даются в табл. П. 1.1 и представляют собой информацию о величине прогнозируемого показателя на интервале времени за несколько предыдущих лет $[t_1; t_2 \ldots t_n]$.

Совокупность числовых значений показателя $y_t$ образует динамический ряд $y_t = \{y_1, y_2 \ldots y_n\}$ на отрезки времени $T = \{t_1, t_2 \ldots t_n\}$. Количество числовых значений, необходимых для решения поставленной задачи, должно быть не менее 8.

Методические положения расчета перспективных технико-экономических показателей деятельности предприятия на основе экстраполяционного метода прогнозирования

**Этап 1.** Установление наличия и тесноты связи между величиной прогнозируемого показателя и фактором времени.

1.1. Определение точечной оценки коэффициента корреляции по формуле:

$$r_{yt} = \frac{\sum y_{t_i} t_i}{\sqrt{\sum t_i^2 \sum y_{t_i}^2}}, \qquad (1.2)$$

где $y_{ti}$ – текущее значение показателя $y_t$ ($t$ = 1…n),

$t_i$ – текущее значение показателя $t$ ($i$ = 1…n),

n – количество лет, за которые собраны статистические данные о значении показателя $y_t$.

По величине $r_{yt}$ определяется сила взаимосвязи $y_t$ и $t$ при наличии между ними линейной связи. Чем ближе $r_{yt}$ к «+1» или «–1», тем ближе связь между $y_t$ и $t$ к линейной. Наличие нелинейной связи определяется с помощью корреляционного отношения $\eta_{yt}$ (расчет см. далее).

1.2. Проверка значения рассчитанного коэффициента корреляции по критерию $t^* \rangle t\gamma, k$:

$$t^* = r_{yt}\sqrt{\frac{n-2}{1-r_{yt}^2}}, \qquad (1.3)$$

где $t_{\gamma,k}$ - коэффициент оценки достоверности гипотезы о значимости коэффициента парной корреляции (табл. П.1.1);

k = n – 2 – число степеней свободы (характеристика суммы квадратов (отклонений) показывает, сколько отклонений в сумме квадратов может изменяться "свободно");

$p = 1 - \frac{\gamma}{2}$ - вход в таблицу $t_{\gamma,k}$;

$\gamma$ - уровень значимости гипотезы.

При выполнении критерия (1.3) гипотеза о значимости коэффициента парной корреляции подтверждается, т. е. величина $y_t$ зависит от фактора времени.

**Этап 2.** Выбор вида математической модели, описывающей взаимозависимости $y_t$ и $t$.

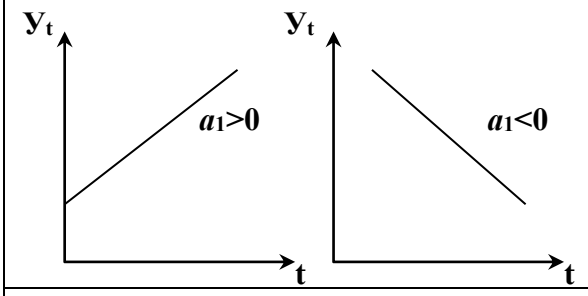
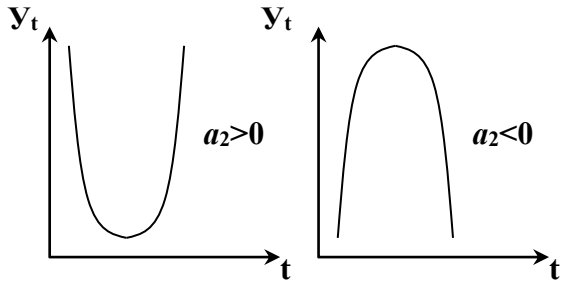
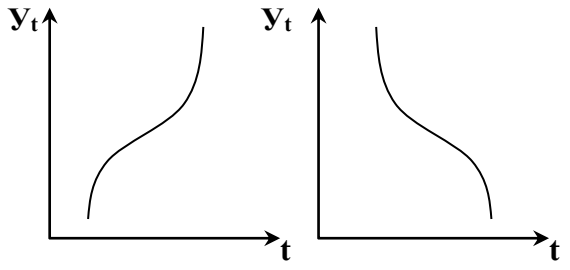
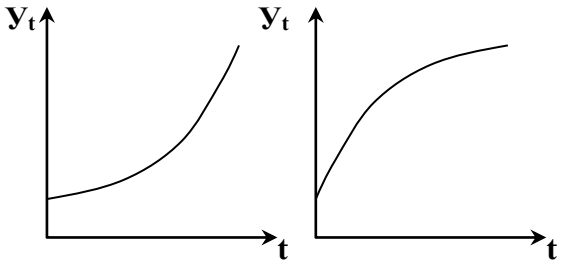
2.1. Построение графика изменения показателя $y_t$ на интервале $[t_1, t_n]$.

2.2. Выбор вида математической модели, описывающей взаимозависимость $y_t$ и $t$ .

Наиболее простой путь выбора формы кривой, описывающей динамику показателя $y_t$ , визуальный – выбор формы кривой на основе графически построенного ряда динамики (этап 2.1). В табл. 1.1 приводится перечень наиболее употребляемых видов кривых, на основе которых выбирают математическую модель и анализируют данные.

Таблица 1.1

Визуальный выбор формы взаимосвязи

| График функции | Вид зависимости | Уравнение тренда |
|---|---|---|
|  | Прямая | $\hat{y}_t = a_0 + a_1 \cdot t$ |
|  | Парабола 2-го порядка | $\hat{y}_t = a_0 + a_1 \cdot t + a_2 \cdot t^2$ |
|  | Парабола 3-го порядка | $\hat{y}_t = a_0 + a_1 \cdot t + a_2 \cdot t^2 + + a_3 \cdot t^3$ |
|  | Показательная кривая (экспонента) | $\hat{y}_t = a_0 \cdot a_1^t$<br>$\hat{y}_t = a_0 \cdot l_1^{a \cdot t}$ |

В большинстве случаев практически приемлемым является метод характеристик прироста, который основывается на сравнении характеристик изменения приростов исследуемого динамического ряда с соответствующими характеристиками кривых роста. Расчет количественных оценок приростов показателя, дополненный визуальным выбором формы взаимосвязи, уменьшает риск неправильного выбора модели для прогнозирования.

2.3. Расчет параметров тренда $\hat{y}_t = f(t)$.

1. Расчет параметров тренда, выбранного для экстраполяции, осуществляется по методу наименьших квадратов (МНК), сущность которого сводится к минимизации суммы квадратов отклонений фактических значений от расчетных (формула 1.4).

$$S(y_t - \hat{y}_t) \to \min \ . \qquad (1.4)$$

На основе МНК параметры уравнения тренда определяются с помощью системы нормальных уравнений.

Нормальные уравнения для расчета параметров прямой имеют вид

$$\begin{cases} na + b\sum t = \sum y_t, \\ a\sum t + b\sum t^2 = \sum t * y_t. \end{cases} \qquad (1.5)$$

Для параболических зависимостей параметры уравнения находят, решая соответствующие системы алгебраических уравнений, или используя встроенные функции Excel.

Следует учитывать, что в практических исследованиях в основном используются следующие функции: линейная, показательная (экспонента), параболическая (2-го и 3-го порядка), гипербола. Поэтому при форме связи следует отдавать предпочтение именно этим зависимостям (для упрощения расчетов). При выполнении вычислений на компьютере на практике осуществляется перебор всех поддающихся вычислению моделей $\hat{y}_t = f(t)$ и производится выбор наилучшей из них. Лучшей считается та модель, для которой приведенные критерии оценки точности принимают наименьшее значение.

Параметры тренда можно также определить в Excel, следуя следующему алгоритму действий:

1.Ввести исходные данные в таблицу.

2.В меню «Вставка» выбрать опцию «Диаграмма», затем выбрать тип «Точечная».

3.Щелкнуть правой кнопкой мыши по точкам графика и в контекстном меню выбрать «Добавить линию тренда».

4.Откроется диалоговое окно, изображенное на рис. 1.2.

5.Затем выбрать наиболее подходящий тип зависимости для данной кривой (например, линейную, логарифмическую и т.д.).

6.Перейти во вкладку «Параметры».

7.Установить галочку рядом с опцией «Показать уравнение на диаграмме» и подтвердить выбор нажатием ОК.

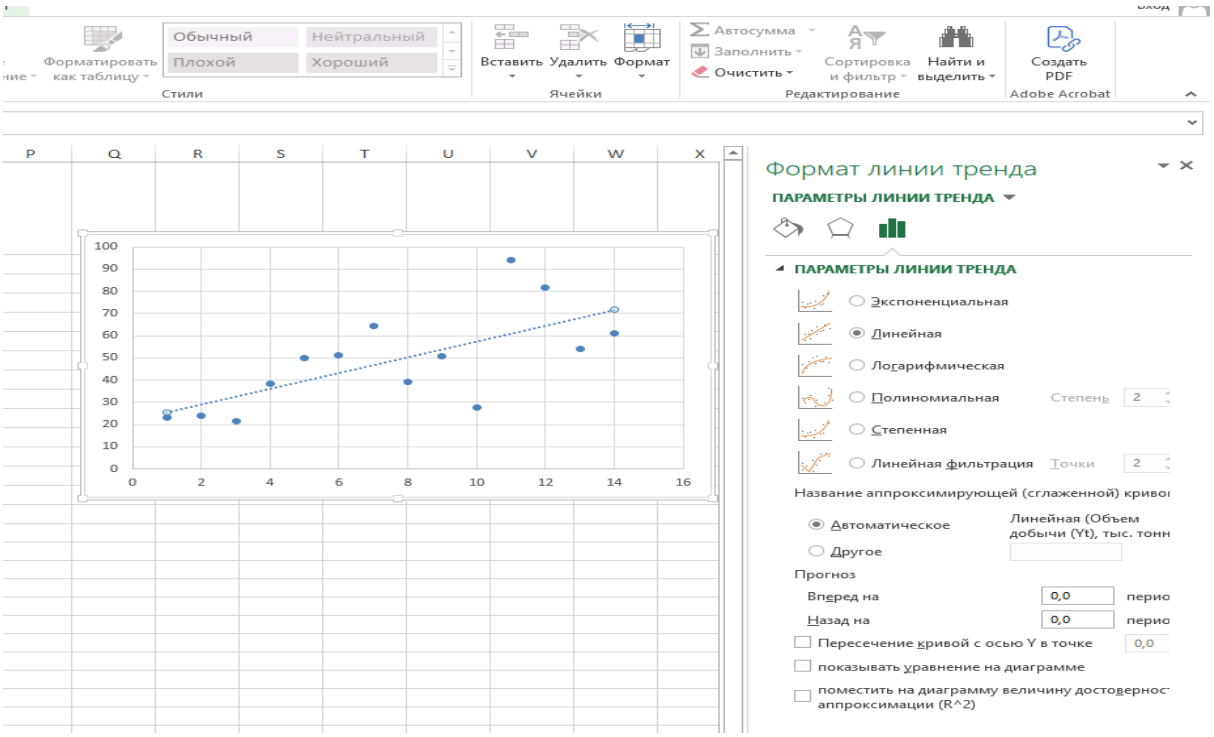



Рис. 1.2. Построение линии тренда и параметров уравнения тренда

Полученное уравнение можно в дальнейшем использовать для прогноза и получения его точечного значения.

**Этап 3.** Расчет критериев точности математической модели.

3.1. Определение расчетных значений моделируемого показателя $\hat{y}_t$, подставляя значения аргументов $(t)$ в полученное уравнение тренда.

3.2. Расчет отклонения фактических значений $y_t$ от расчетных $\hat{y}_t$ по формуле:

$$\Delta y = y_t - \hat{y}_t . \qquad (1.6)$$

3.3. Расчет среднего квадратического отклонения.

$$S_{\hat{y}_t} = \sqrt{\frac{\sum (y_t - \hat{y}_t)^2}{n - p}}, \qquad (1.7)$$

где n – число наблюдений,

p – количество расчетных коэффициентов в уравнении тренда.

3.4. Расчет средней относительной ошибки

$$\varepsilon = \frac{1}{n} \sum \left| \frac{y_t - \hat{y}_t}{y_t} \right| * 100 . \qquad (1.8)$$

Критерии (1.7) и (1.8) показывают степень точности воспроизведения моделью реального изменения моделируемого показателя.

3.5. Важным критерием надежности модели является эмпирическое корреляционное отклонение

$$\eta_{yt} = \sqrt{1 - \frac{S^2_{\hat{y}_e}}{S_2}}, \qquad (1.9)$$

где S2– общая дисперсия,

$S^2_{\hat{y}_t}$ – остаточная дисперсия.

$$\text{S2} = \sum_{i=1}^{n} \frac{(y_t - \bar{y}_t)^2}{n-1}, \qquad (1.10)$$

где $\bar{y}_t$ - математическое ожидание показателя (среднее значение), вычисленное по заданному динамическому ряду.

Корреляционное отношение характеризует тесноту связи между $y_t$ и t при нелинейных зависимостях, его значения находятся в пределах от 0 до 1. Если зависимость линейна, то $\eta_{yt} \approx |r_{yt}|$.

Поскольку можно утверждать, что если $r_{yt} = 0$, $\eta_{yt} \neq 0$, то в совокупности с графическим анализом зависимости $y_t$ и $t$ с помощью коэффициента парной корреляции можно оценивать наличие взаимосвязи как при линейной, так и нелинейной корреляционной зависимости.

Это условие может быть использовано в качестве критерия линейности модели. Если условие выполняется, то линейность регрессии подтверждается.

**Этап 4.** Прогнозирование показателя $y_t$.

1. Расчет точечной оценки прогноза показателя $\hat{y}_{n+1}$ осуществляется подстановкой величины $t_i$ = n+1 в полученное уравнение тренда.

2. Расчет интервальной оценки прогноза осуществляется по зависимости

$$y_{n+1} = \hat{y}_{n+1} \pm t_{jk} \times S^*_{yt}, \tag{1.11}$$

$$S^*_{yt} = S_{yt} \sqrt{1 + \frac{1}{n} + \frac{(t_{n+1} - \bar{t})^2}{\sum (t_i - \bar{t})^2}}^*, \tag{1.12}$$

где $t_{gk}$ – статистика Стьюдента, определяемая (прил., табл. П2) по выходам k= n-2 и p = 1-$\alpha$/2;

n -число наблюдений в динамическом ряду;

$t_{n+1}$ – величина t для прогноза года;

$\bar{t}$ – математическое ожидание t;

$S_{yt}$ - среднее квадратическое отклонение фактических наблюдений $y_t$ от расчетных $\hat{y}_t$ (см. формулу 1.7).

# Глава 2. МЕТОДЫ ЭКСПОНЕНЦИАЛЬНОГО СГЛАЖИВАНИЯ

В методе простого экспоненциального сглаживания используется взвешенное (экспоненциальное) скользящее усреднение всех ранее наблюдаемых данных. Эта модель обычно применяется к данным, в которых требуется определить наличие зависимости между анализируемыми показателями (например, тренда) или выявить взаимосвязь между данными. Основная задача экспоненциального сглаживания – оценить текущее состояние процесса, что служит основой для формирования всех последующих прогнозов.

*Экспоненциальное сглаживание предусматривает* постоянное обновление модели за счет наиболее свежих данных. Этот метод основывается на усреднении (сглаживании) временных рядов прошлых наблюдений в нисходящем (экспоненциально) направлении. То есть более поздним событиям присваивается больший вес. Вес присваивается следующим образом: для последнего наблюдения весом будет величина α, для предпоследнего – (1-α), для того, которое было перед ним, – $(1-\alpha)^2$ и т.д.

В сглаженном виде новый прогноз (для периода времени t+1) можно представлять как взвешенное среднее последнего наблюдения величины в момент времени t и ее прежнего прогноза на этот же период t. Причем вес α присваивается наблюдаемому значению, а вес (1-α) – прогнозу; при этом полагается, что $0< \alpha<1$. Это правило в общем виде можно записать следующим образом.

Новый прогноз = [α*(последнее наблюдение)]+[(1-α)*последний прогноз]

$$\hat{Y}_{t+1} = \alpha Y_t + (1-\alpha) * \hat{Y}_t \qquad (2.1)$$

где $\hat{Y}_{t+1}$ – прогнозируемое значение на следующий период;

α – постоянная сглаживания;

$Y_t$ – наблюдение величины за текущий период t;

$\hat{Y}_t$ - прежний сглаженный прогноз этой величины на период t.

Экспоненциальное сглаживание – это процедура для постоянного пересмотра результатов прогнозирования в свете самых последних событий.

Постоянная сглаживания α является взвешенным фактором. Ее реальное значение определяется тем, в какой мере текущее наблюде-

ние должно влиять на прогнозируемую величину. Если α близко к 1, значит в прогнозе существенно учитывается величина ошибки последнего прогнозирования. И наоборот, при малых значениях α прогнозируемая величина наиболее близка к предыдущему прогнозу. Можно представить $\hat{Y}_t$ как взвешенное среднее значение всех прошлых наблюдений с весовыми коэффициентами, экспоненциально убывающими с «возрастом» данных.

Таблица 2.1

Сравнение влияния разных значений постоянных сглаживания

| Период | α=0,1 | | α=0,6 | |
|---|---|---|---|---|
| | Расчет | Вес | Расчет | Вес |
| t | | 0.100 | | 0,600 |
| t-1 | 0,9*0,1 | 0,090 | 0,4*0,6 | 0,240 |
| t-2 | 0,9*0,9*0,1 | 0,081 | 0,4*0,4*0,6 | 0,096 |
| t-3 | 0,9*0,9*0,9*0,1 | 0,073 | 0,4*0,4*0,4*0,6 | 0,038 |
| t-4 | 0,9*0,9*0,9*0,9*0,1 | 0,066 | 0,4*0,4*0,4*0,4*0,6 | 0,015 |
| Остальные | | 0,590 | | 0,011 |
| | Всего | 1,000 | Всего | 1,000 |

Постоянная α является ключом к анализу данных. Если требуется, чтобы спрогнозированные величины были стабильны и случайные отклонения сглаживались, необходимо выбирать малое значение α. Большое значение постоянной α имеет смысл в том случае, если нужна быстрая реакция на изменения в спектре наблюдений.

1. Практический пример проведения экспоненциального сглаживания.

Логистический отдел торговой компании ведет учет поступления товара «А» на склад (в условных единицах) за период 28 месяцев. Для оперативного планирования закупок используется метод экспоненциального сглаживания. Необходимо подобрать оптимальный коэффициент адаптации (α), сравнив модели с параметрами α = 0,1 и α = 0,6.

В качестве начального условия (сглаженное значение на 1-й месяц) принято фактическое значение первого периода, равное 520 ед. Оценка качества моделей производится на ретроспективном участке

(последние 7 месяцев наблюдений) путем сравнения прогнозных ошибок.

Таблица 2.2

Исходные данные

| Период (Месяц) | Факт (поступление, ед.) | Модель α = 0,1 | Ошибка прогноза |
|---|---|---|---|
| 1 | 520 | 520,00 | 0,00 |
| 2 | 370 | 520,00 | -150,00 |
| 3 | 260 | 505,00 | -245,00 |
| 4 | 410 | 480,50 | -70,50 |
| 5 | 470 | 473,45 | -3,45 |
| 6 | 360 | 473,11 | -113,11 |
| 7 | 210 | 461,80 | -251,80 |
| 8 | 310 | 436,62 | -126,62 |
| 9 | 360 | 423,96 | -63,96 |
| 10 | 210 | 417,56 | -207,56 |
| 11 | 160 | 396,80 | -236,80 |
| 12 | 420 | 373,12 | 46,88 |
| 13 | 570 | 377,81 | 192,19 |
| 14 | 360 | 397,03 | -37,03 |
| 15 | 260 | 393,33 | -133,33 |
| 16 | 570 | 379,99 | 190,01 |
| 17 | 580 | 398,99 | 181,01 |
| 18 | 410 | 417,09 | -7,09 |
| 19 | 360 | 416,38 | -56,38 |
| 20 | 620 | 410,74 | 209,26 |
| 21 | 770 | 431,67 | 338,33 |
| 22 | 520 | 465,50 | 54,50 |
| 23 | 420 | 470,95 | -50,95 |
| 24 | 670 | 465,86 | 204,14 |
| 25 | 870 | 486,27 | 383,73 |

На рис. 2.1 изображен прогноз, построенный с использованием метода экспоненциального сглаживания с постоянным коэффициентом сглаживания, равным 0,1.


Рис. 2.1. Экспоненциальное сглаживание

***Решение в Excel***

1. Перейдите в меню «Анализ данных». В списке «Инструменты анализа» выберите «Экспоненциальное сглаживание». Если в меню отсутствует пункт «Анализ данных», необходимо установить «Пакет анализа». Для этого откройте «Параметры», выберите «Настройки» и в появившемся диалоговом окне установите флажок «Пакет анализа», затем нажмите ОК.

2. После этого на экране откроется диалоговое окно, аналогичное тому, что показано на рис. 2.2.

3. В поле «Входной интервал» введите диапазон исходных данных (оставьте одну свободную ячейку).

4. Установите флажок «Метки», если в диапазоне есть названия столбцов.

5. Введите значение «Фактор затухания» (1-α).

6. В поле «Входной интервал» укажите ячейку, в которой хотите получить результат.

7. Включите опцию «Вывод графика», установив соответствующий флажок, чтобы автоматически построить график.

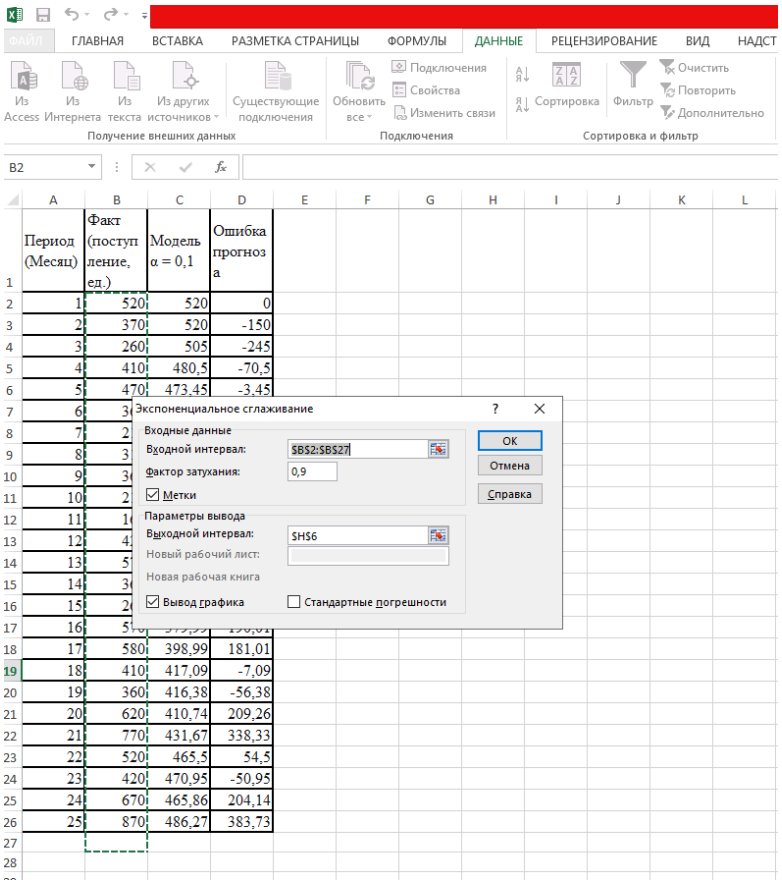



Рис. 2.2. Диалоговое окно для экспоненциального сглаживания

## Реализация экспоненциального сглаживания в R

### Подготовка данных и загрузка библиотек

```
# ================================================
# ЭКСПОНЕНЦИАЛЬНОЕ СГЛАЖИВАНИЕ ВРЕМЕННОГО РЯДА
# Практический пример для торговой компании "Восток-Импорт"
# ================================================

# Очистка рабочей области
rm(list = ls())

# Подключение необходимых библиотек
# install.packages(c("forecast", "ggplot2", "gridExtra", "TTR"))
library(forecast)  # для прогнозирования временных рядов
library(ggplot2)  # для визуализации
```

```r
library(gridExtra) # для компоновки графиков
library(TTR)    # для альтернативных методов сглажива-
ния
library(tidyr)   # для трансформации данных
library(dplyr)   # для манипуляции данными

# Ввод исходных данных
# Создаем вектор фактических значений временного ряда
actual_values <- c(
 500, 350, 250, 400, # 1 год
 450, 350, 200, 300, # 2 год
 350, 200, 150, 400, # 3 год
 550, 350, 250, 550, # 4 год
 550, 400, 350, 600, # 5 год
 750, 500, 400, 650, # 6 год
 850, 500, 450, 700  # 7 год
)

# Создаем временные метки
periods <- 1:28
years <- rep(1:7, each = 4)
quarters <- rep(1:4, times = 7)

# Формируем датафрейм с исходными данными
data <- data.frame(
 period = periods,
 year = years,
 quarter = quarters,
 actual = actual_values
)

# Просмотр первых строк данных
print("Первые 10 наблюдений временного ряда:")
head(data, 10)

# Проверка структуры данных
str(data)
```

**Функция для экспоненциального сглаживания**

Реализуем пользовательскую функцию для расчета экспоненциального сглаживания с заданным параметром α.

```
# ================================================
# Функция экспоненциального сглаживания
# ================================================

exp_smoothing <- function(y, alpha, initial_value =
NULL) {
 # y - вектор фактических значений
 # alpha - параметр сглаживания (0 < alpha < 1)
 # initial_value - начальное сглаженное значение (если
NULL, берется первое фактическое)

 n <- length(y)
 smoothed <- numeric(n)

 # Установка начального значения
 if (is.null(initial_value)) {
  smoothed[1] <- y[1]
 } else {
  smoothed[1] <- initial_value
 }

 # Расчет сглаженных значений по формуле St = α·Yt +
(1-α)·St-1
 for (t in 2:n) {
  smoothed[t] <- alpha * y[t] + (1 - alpha) *
smoothed[t-1]
 }

 return(smoothed)
}

# Функция для расчета ошибок прогноза
calc_errors <- function(actual, smoothed) {
 errors <- actual - smoothed
```

```
 return(errors)
}

# Функция для расчета метрик точности
calc_accuracy_metrics <- function(actual, smoothed) {
 errors <- actual - smoothed
 abs_errors <- abs(errors)
 squared_errors <- errors^2
 percent_errors <- (abs_errors / actual) * 100

 # Исключаем первую точку (где нет ошибки прогноза)
 valid_idx <- 2:length(actual)

 metrics <- data.frame(
  MAE = mean(abs_errors[valid_idx]),           # Mean
Absolute Error
  MSE = mean(squared_errors[valid_idx]),         # Mean
Squared Error
  RMSE = sqrt(mean(squared_errors[valid_idx])),     #
Root Mean Squared Error
  MAPE = mean(percent_errors[valid_idx], na.rm = TRUE)
# Mean Absolute Percentage Error
 )

 return(metrics)
}
```

**Расчет экспоненциального сглаживания для α = 0,1**

```
# ================================================
# Расчет модели с параметром α = 0,1
# ================================================

alpha1 <- 0.1
initial_val <- 500 # начальное значение согласно усло-
вию
```

```r
# Расчет сглаженных значений
smoothed_alpha1 <- exp_smoothing(data$actual, alpha =
alpha1, initial_value = initial_val)

# Расчет ошибок прогноза
errors_alpha1 <- calc_errors(data$actual,
smoothed_alpha1)

# Добавление результатов в датафрейм
data$smoothed_01 <- smoothed_alpha1
data$error_01 <- errors_alpha1

# Создание таблицы результатов для α = 0,1 (аналог
таблицы в Excel)
results_alpha1 <- data.frame(
 Год = data$year,
 Квартал = data$quarter,
 Период = data$period,
 Фактическое = data$actual,
 Сглаженное_01 = round(data$smoothed_01, 2),
 Ошибка_01 = round(data$error_01, 2)
)

print("РЕЗУЛЬТАТЫ ЭКСПОНЕНЦИАЛЬНОГО СГЛАЖИВАНИЯ (α =
0,1):")
print(results_alpha1)

# Расчет метрик точности для α = 0,1
metrics_alpha1 <- calc_accuracy_metrics(data$actual,
data$smoothed_01)
print("Метрики точности для модели с α = 0,1:")
print(metrics_alpha1)
```

**Расчет экспоненциального сглаживания для α = 0,6**

```r
# ================================================
# Расчет модели с параметром α = 0,6
# ================================================
```

```r
alpha2 <- 0.6
# Расчет сглаженных значений
smoothed_alpha2 <- exp_smoothing(data$actual, alpha =
alpha2, initial_value = initial_val)
# Расчет ошибок прогноза
errors_alpha2 <- calc_errors(data$actual,
smoothed_alpha2)

# Добавление результатов в датафрейм
data$smoothed_06 <- smoothed_alpha2
data$error_06 <- errors_alpha2

# Создание таблицы результатов для α = 0,6
results_alpha2 <- data.frame(
 Год = data$year,
 Квартал = data$quarter,
 Период = data$period,
 Фактическое = data$actual,
 Сглаженное_06 = round(data$smoothed_06, 2),
 Ошибка_06 = round(data$error_06, 2)
)

print("РЕЗУЛЬТАТЫ ЭКСПОНЕНЦИАЛЬНОГО СГЛАЖИВАНИЯ (α =
0,6):")
print(results_alpha2)

# Расчет метрик точности для α = 0,6
metrics_alpha2 <- calc_accuracy_metrics(data$actual,
data$smoothed_06)
print("Метрики точности для модели с α = 0,6:")
print(metrics_alpha2)
```

**Сравнительная таблица результатов**

```r
# ================================================
# Сравнительная таблица для обеих моделей
# ================================================
```

```
comparison_table <- data.frame(
 Период = data$period,
 Год = data$year,
 Квартал = data$quarter,
 Факт = data$actual,
 Сглаж_01 = round(data$smoothed_01, 2),
 Ошибка_01 = round(data$error_01, 2),
 Сглаж_06 = round(data$smoothed_06, 2),
 Ошибка_06 = round(data$error_06, 2)
)

print("СРАВНИТЕЛЬНАЯ ТАБЛИЦА РЕЗУЛЬТАТОВ:")
print(comparison_table)

# Сохранение таблицы в CSV (опционально)
# write.csv(comparison_table, "exponen-
tial_smoothing_results.csv", row.names = FALSE)
```

**Визуализация результатов**

```
# ================================================
# ВИЗУАЛИЗАЦИЯ РЕЗУЛЬТАТОВ ЭКСПОНЕНЦИАЛЬНОГО СГЛАЖИВА-
НИЯ
# ================================================

# График 1: Сравнение фактических значений и сглажен-
ных рядов
p1 <- ggplot(data, aes(x = period)) +
 geom_line(aes(y = actual, color = "Фактические значе-
ния"), size = 1) +
 geom_line(aes(y = smoothed_01, color = "Сглаживание α
= 0,1"), size = 1, linetype = "dashed") +
 geom_line(aes(y = smoothed_06, color = "Сглаживание α
= 0,6"), size = 1, linetype = "dotted") +
 labs(title = "Экспоненциальное сглаживание временного
ряда продаж",
```

```
    subtitle = "Сравнение моделей с различными пара-
метрами сглаживания",
    x = "Период (квартал)", y = "Объем продаж (тыс.
усл. ед.)") +
 scale_color_manual(values = c("Фактические значения"
= "black",
                    "Сглаживание α = 0,1" = "blue",
                    "Сглаживание α = 0,6" = "red")) +
 theme_minimal() +
 theme(legend.position = "bottom")

# График 2: Ошибки прогноза для обеих моделей
p2 <- ggplot(data, aes(x = period)) +
 geom_line(aes(y = error_01, color = "Ошибки (α =
0,1)"), size = 0.8) +
 geom_line(aes(y = error_06, color = "Ошибки (α =
0,6)"), size = 0.8) +
 geom_hline(yintercept = 0, linetype = "dashed", color
= "gray50") +
 labs(title = "Ошибки прогноза для моделей экспоненци-
ального сглаживания",
    x = "Период (квартал)", y = "Ошибка прогноза (тыс.
усл. ед.)") +
 scale_color_manual(values = c("Ошибки (α = 0,1)" =
"blue",
                    "Ошибки (α = 0,6)" = "red")) +
 theme_minimal() +
 theme(legend.position = "bottom")

# Компоновка графиков
grid.arrange(p1, p2, ncol = 1, heights = c(1.2, 1))
```

**Дополнительная визуализация с разделением по годам**

```
# ================================================
# Детальный график с разбивкой по годам
# ================================================
```

```r
# Создание меток для оси X
data$time_label <- paste(data$year, "Q", data$quarter,
sep = "")

# График с точками и линиями
p3 <- ggplot(data, aes(x = period)) +
 geom_point(aes(y = actual, color = "Факт"), size = 2)
+
 geom_line(aes(y = actual, color = "Факт"), size =
0.8, alpha = 0.5) +
 geom_line(aes(y = smoothed_01, color = "α = 0,1"),
size = 1.2) +
 geom_line(aes(y = smoothed_06, color = "α = 0,6"),
size = 1.2) +
 scale_x_continuous(breaks = seq(1, 28, by = 2),
           labels = data$time_label[seq(1, 28, by =
2)]) +
 labs(title = "Детальный анализ экспоненциального
сглаживания",
    subtitle = "Поквартальные данные за 7 лет",
    x = "Временной период", y = "Объем продаж (тыс.
усл. ед.)") +
 scale_color_manual(name = "Ряд данных",
           values = c("Факт" = "black",
                "α = 0,1" = "blue",
                "α = 0,6" = "red")) +
 theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust =
1),
    legend.position = "bottom")

print(p3)
```

**Использование встроенной функции HoltWinters (для сравнения)**

```r
# ================================================
# Использование встроенной функции HoltWinters
```

```r
# ================================================
# Преобразование в объект временного ряда
ts_data <- ts(data$actual, start = c(1, 1), frequency
= 4)

# Простое экспоненциальное сглаживание с помощью
HoltWinters
hw_alpha1 <- HoltWinters(ts_data, alpha = alpha1, beta
= FALSE, gamma = FALSE)
hw_alpha2 <- HoltWinters(ts_data, alpha = alpha2, beta
= FALSE, gamma = FALSE)

print("Результаты HoltWinters (α = 0,1):")
print(hw_alpha1)

print("Результаты HoltWinters (α = 0,6):")
print(hw_alpha2)

# Сравнение сглаженных значений, полученных разными
способами
# (наша функция и встроенная должны давать одинаковые
результаты)
comparison_hw <- data.frame(
 Период = data$period,
 Факт = data$actual,
 Наша_01 = data$smoothed_01,
 HW_01 = round(as.numeric(fitted(hw_alpha1)[,1]), 2),
 Наша_06 = data$smoothed_06,
 HW_06 = round(as.numeric(fitted(hw_alpha2)[,1]), 2)
)

print("Сравнение результатов ручного расчета и
HoltWinters:")
head(comparison_hw, 10)
```

**Прогнозирование на будущие периоды**

```
# ================================================
# Прогнозирование на следующие 4 квартала
# ================================================

# Построение прогноза с помощью HoltWinters
hw_model <- HoltWinters(ts_data, alpha = 0.6, beta =
FALSE, gamma = FALSE)
forecast_hw <- forecast(hw_model, h = 4)

print("Прогноз на следующие 4 квартала:")
print(forecast_hw)

# Визуализация прогноза
p_forecast <- autoplot(forecast_hw) +
 labs(title = "Прогноз продаж на основе экспоненциаль-
ного сглаживания",
     subtitle = "Модель с параметром α = 0,6",
     x = "Время (годы)", y = "Объем продаж (тыс. усл.
ед.)") +
 theme_minimal()

print(p_forecast)

# Таблица с прогнозными значениями
forecast_table <- data.frame(
 Период = c("8 год, I кв", "8 год, II кв", "8 год, III
кв", "8 год, IV кв"),
 Прогноз = round(as.numeric(forecast_hw$mean), 2),
 Нижняя_граница_80 =
round(as.numeric(forecast_hw$lower[,1]), 2),
 Верхняя_граница_80 =
round(as.numeric(forecast_hw$upper[,1]), 2),
 Нижняя_граница_95 =
round(as.numeric(forecast_hw$lower[,2]), 2),
```

```
 Верхняя_граница_95 =
round(as.numeric(forecast_hw$upper[,2]), 2)
)

print("ПРОГНОЗНЫЕ ЗНАЧЕНИЯ С ДОВЕРИТЕЛЬНЫМИ ИНТЕРВАЛА-
МИ:")
print(forecast_table)
```

**Оценка качества моделей и выбор оптимальной**

```
# ================================================
# СРАВНИТЕЛЬНЫЙ АНАЛИЗ КАЧЕСТВА МОДЕЛЕЙ
# ================================================

# Сбор метрик для обеих моделей
metrics_comparison <- rbind(
 data.frame(Модель = "α = 0,1 (сильное сглаживание)",
metrics_alpha1),
 data.frame(Модель = "α = 0,6 (слабое сглаживание)",
metrics_alpha2)
)

print("СРАВНИТЕЛЬНЫЕ МЕТРИКИ КАЧЕСТВА МОДЕЛЕЙ:")
print(metrics_comparison)

# Визуализация сравнения метрик
metrics_long <- metrics_comparison %>%
 pivot_longer(cols = c(MAE, MSE, RMSE, MAPE), names_to
= "Метрика", values_to = "Значение")

p_metrics <- ggplot(metrics_long, aes(x = Метрика, y =
Значение, fill = Модель)) +
 geom_bar(stat = "identity", position = "dodge") +
 labs(title = "Сравнение метрик качества моделей экс-
поненциального сглаживания",
     x = "Метрика", y = "Значение метрики") +
 theme_minimal() +
```

```
 scale_fill_manual(values = c("α = 0,1 (сильное сгла-
живание)" = "lightblue",
                "α = 0,6 (слабое сглаживание)" =
"lightcoral"))

print(p_metrics)

# Статистический тест для сравнения точности прогнозов
# (парный t-тест для абсолютных ошибок)
abs_errors_01 <- abs(data$error_01[2:28])
abs_errors_06 <- abs(data$error_06[2:28])

t_test_result <- t.test(abs_errors_01, abs_errors_06,
paired = TRUE)

cat("\nСТАТИСТИЧЕСКОЕ СРАВНЕНИЕ МОДЕЛЕЙ:")
cat("\nПарный t-тест для абсолютных ошибок прогноза:")
cat("\nt-статистика =", round(t_test_result$statistic,
4))
cat("\np-value =", format.pval(t_test_result$p.value,
digits = 4))

if (t_test_result$p.value < 0.05) {
 cat("\nВЫВОД: Различие в точности моделей статистиче-
ски значимо.")
} else {
 cat("\nВЫВОД: Различие в точности моделей статистиче-
ски не значимо.")
}
```

3. Задание для выполнения лабораторной работы.

Горнодобывающее предприятие «Северные месторождения» ведет мониторинг ежемесячных объемов добычи нефти. В табл. 2.3 представлены фактические данные о добыче за период с мая первого года по июнь второго года. Всего в выборке представлено 14 наблюдений.

Таблица 2.3

Исходные данные

| Год | Месяц | Период (t) | Объем добычи (Yt), тыс. тонн |
|---|---|---|---|
| 1 | Май | 1 | 23,15 |
| 1 | Июнь | 2 | 23,77 |
| 1 | Июль | 3 | 21,47 |
| 1 | Август | 4 | 38,37 |
| 1 | Сентябрь | 5 | 50,00 |
| 1 | Октябрь | 6 | 51,12 |
| 1 | Ноябрь | 7 | 64,33 |
| 1 | Декабрь | 8 | 39,35 |
| 2 | Январь | 9 | 50,60 |
| 2 | Февраль | 10 | 27,50 |
| 2 | Март | 11 | 94,00 |
| 2 | Апрель | 12 | 81,90 |
| 2 | Май | 13 | 54,20 |
| 2 | Июнь | 14 | 61,10 |

Проведите экспоненциальное сглаживание рядов. Коэффициент экспоненциального сглаживания примите равным 0,1; 0,2; 0,3. Полученные результаты прокомментируйте. Можно использовать статистические данные, представленные в прил., табл. П1.

# Глава 3. ОЦЕНКА СОГЛАСОВАННОСТИ МНЕНИЙ ЭКСПЕРТОВ С ПРИМЕНЕНИЕМ КОЭФФИЦИЕНТА КОНКОРДАЦИИ

При проведении различных выборочных исследований, таких как экономические или социологические, часто возникает необходимость оценки согласованности мнений экспертов. Это позволяет, во-первых, определить различные подходы экспертов к оценке различных явлений, признаков или критериев, а во-вторых, провести более глубокий анализ ситуации и принять обоснованные решения.

Для этого необходимо провести экспертизу, направленную на выявление степени влияния различных факторов на финансово-экономическую деятельность предприятия. В рамках задания студенты делятся на две группы, выступая в роли экспертов.

Первая группа определяет факторы, отражающие влияние внутренней среды предприятия, а вторая – внешней среды. Каждая группа формирует матрицу исходных данных, в которой в качестве объектов выступают соответствующие факторы, влияющие на финансово-экономическую деятельность. Каждый студент-эксперт оценивает влияние каждого фактора, присваивая ему балл по десятибалльной шкале.

После этого обе группы оценивают согласованность своих мнений с помощью коэффициента конкордации и делают соответствующие выводы.

1. Методические положения оценки согласованности мнения экспертов с применением коэффициента конкордации.

В связи с тем, что данные, полученные в результате экспертных оценок (обычно в виде баллов), имеют характер ранговых (непараметрических) показателей, для их анализа обычно применяются соответствующие ранговые методы.

Чтобы определить степень согласованности мнений двух экспертов при оценке ряда признаков или объектов (что важно при решении задач ранжирования), используют коэффициенты корреляции Спирмена или Кендалла.

Если же экспертов больше, например, при оценке мнений целой группы специалистов, то для оценки согласованности применяют дисперсионный коэффициент конкордации:

$$W = \frac{12 * S}{m^2 * (n^3 - n)}, \tag{3.1}$$

где
$$S = \sum_{j=1}^{n} (\sum_{i=1}^{m} R_{ij} - \frac{m(n+1)}{2})^2, \tag{3.2}$$

n – количество анализируемых объектов,

m – количество экспертов,

$R_{ij}$ – ранг *j*-го объекта, который присвоен ему *i*-ым экспертом.

Дисперсионный коэффициент конкордации рассчитывают по матрице ранжировок *n* объектов группой из *m* экспертов, где $r_{ij}$ – ранг, присвоенный *j*-ым экспертом *i*-ому объекту.

Следует обратить внимание на отличие значений коэффициента конкордации от коэффициента корреляции, так как он существует в пределах от 0 до 1. Если мнения экспертов полностью противоположны, коэффициент конкордации равен нулю ($W = 0$), а коэффициент корреляции в этом случае будет равен -1.

При наличии одинаковых связанных рангов формула (3.1) приобретает следующий вид:

$$W = \frac{12 \cdot S}{m^2 \cdot (n^3 - n) - m \cdot \sum_{j=1}^{m} T_j}, \tag{3.3}$$

$$T_j = \sum_{k=1}^{H_j} (h_k^3 - h_k), \tag{3.4}$$

$$S = \sum_{i=1}^{n} (\sum_{j=1}^{m} r_{ij} - \bar{r})^2, \tag{3.5}$$

где $T_j$ – показатель связанных (одинаковых) рангов в *j*-ой ранжировке,

$H_j$ – число групп равных рангов в *j*-ой ранжировке;

$h_k$ – число равных рангов в *k*-ой группе связанных рангов при ранжировке *j*-ым экспертом,

*n* – число объектов,

*m* – число экспертов

$r_{ij}$- ранг, присваиваемый *j*-ым экспертом *i*-ому объекту;

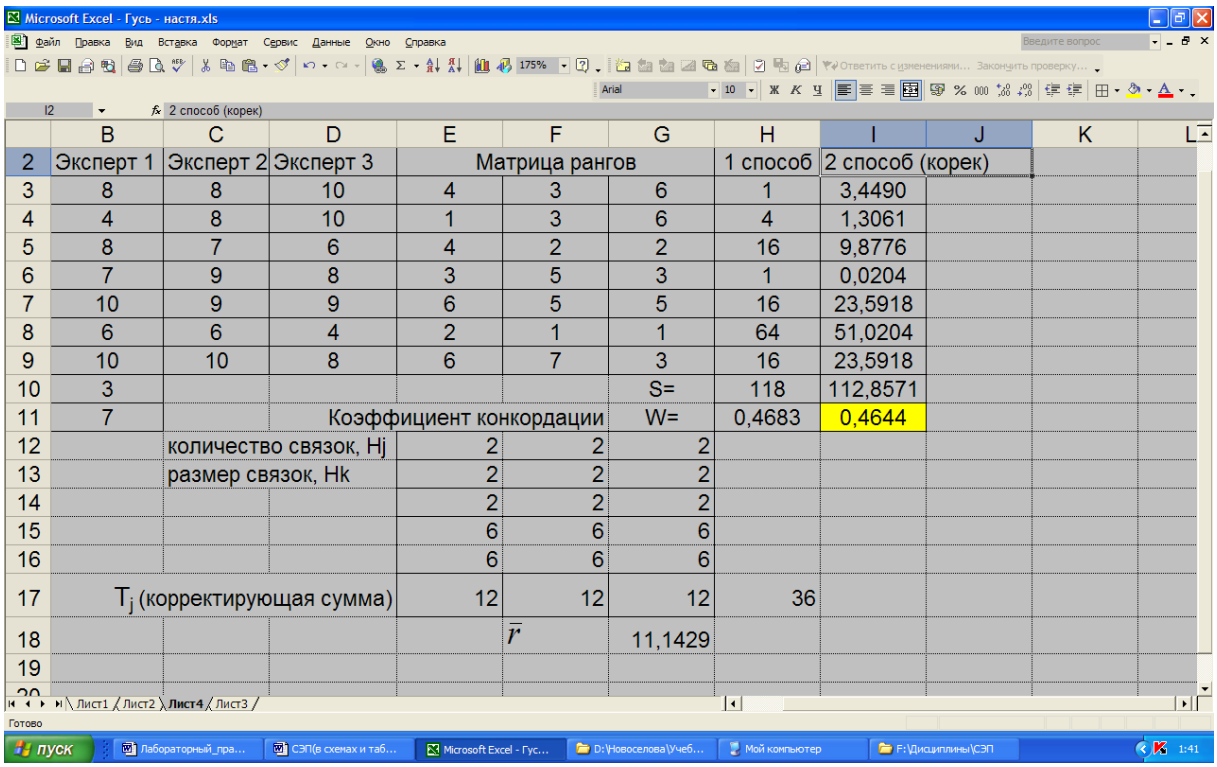
$\bar{r}$ – средний ранг, равный $\bar{r} = \frac{1}{n} \cdot \sum_{i=1}^{n} r_i$.

Если коэффициент конкордации равен 1, то все ранжировки экспертов одинаковы; а $W = 0$, если все ранжировки различны, то совершенно нет совпадений. Мнения экспертов согласованны, если $W >$ 0,6. Если $W <$ 0,6, анализируют ответы на согласованность мнений,

выявляют дополнительные факторы, которые необходимо учесть экспертам, определяют экспертов, мнение которых максимально расходится с общим мнением.

2. Практический пример оценки согласованности мнения экспертов с применением коэффициента конкордации.

Имеется 7 объектов, каждый из которых оценивается независимо тремя экспертами по десятибалльной шкале (см. рис. 3.1, ячейки диапазона A2:D9). Необходимо определить степень согласованности мнений экспертов – коэффициент конкордации.




| | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | Эксперт 1 | Эксперт 2 | Эксперт 3 | Матрица рангов | | | 1 способ | 2 способ (корек) | | | |
| 3 | 8 | 8 | 10 | 4 | 3 | 6 | 1 | 3,4490 | | | |
| 4 | 4 | 8 | 10 | 1 | 3 | 6 | 4 | 1,3061 | | | |
| 5 | 8 | 7 | 6 | 4 | 2 | 2 | 16 | 9,8776 | | | |
| 6 | 7 | 9 | 8 | 3 | 5 | 3 | 1 | 0,0204 | | | |
| 7 | 10 | 9 | 9 | 6 | 5 | 5 | 16 | 23,5918 | | | |
| 8 | 6 | 6 | 4 | 2 | 1 | 1 | 64 | 51,0204 | | | |
| 9 | 10 | 10 | 8 | 6 | 7 | 3 | 16 | 23,5918 | | | |
| 10 | 3 | | | | | S= | 118 | 112,8571 | | | |
| 11 | 7 | | Коэффициент конкордации | | | W= | 0,4683 | 0,4644 | | | |
| 12 | | количество связок, Hj | | 2 | 2 | 2 | | | | | |
| 13 | | размер связок, Hk | | 2 | 2 | 2 | | | | | |
| 14 | | | | 2 | 2 | 2 | | | | | |
| 15 | | | | 6 | 6 | 6 | | | | | |
| 16 | | | | 6 | 6 | 6 | | | | | |
| 17 | $T_j$ (корректирующая сумма) | | | 12 | 12 | 12 | 36 | | | | |
| 18 | | | | | $\bar{r}$ | 11,1429 | | | | | |
| 19 | | | | | | | | | | | |

Рис. 3.1. Исходные данные и результаты расчета коэффициента конкордации

***Решение в Excel***

Промежуточные и конечные результаты расчетов приведены на рис. 3.1. Расчет коэффициента конкордации проводится по следующим этапам.

**1 этап.** Формируется таблица (матрица) рангов для исходных данных. Данные ранжируются по экспертам (по столбцам). Для этого в ячейку E3 помещается формула =Ранг(B3;B$3:B$9;1) и согласованно копируется во все ячейки диапазона E3:G9.

**2 этап.** В ячейки B10 и B11 вводится число экспертов m =ЧИСЛСТОЛБ(C3:E3) и число объектов n =ЧСТРОК(C3:C9). В указанные ячейки можно прямо ввести соответствующие значения: в B10 – 3 (количество экспертов), а в B11 – 7(количество объектов).

**3 этап.** Формируется столбец вспомогательных результатов для расчета величины *S*. Для этого суммируется построчно (в нашем случае) ранги для *i*-го объекта, из суммы вычитается m(n+1)/2 и результат возводится в квадрат. С этой целью в ячейку H3 помещается формула: = (СУММ(E3:G3)-($B$10*($B$11+1))/2)^2. После этого согласованно копируется во все ячейки диапазона H3:H9.

**4 этап**. Находится значение величины *S*, просуммировав значения по столбцу, в ячейку H10 вводится формула =СУММ(H3:H9).

**5 этап.** Помещая в ячейку H11 формулу = 12*H10/(B10^2*(B11^3- -B11)), рассчитывается значение коэффициента конкордации. Рассчитанный коэффициент конкордации (0,4683) требует корректировки, так как в нем не учтено то, что в полученных ранжировках имеются одинаковые значения. Для того чтобы учесть их, необходимо рассчитать коэффициент конкордации по формуле 3.3. и выполнить дл этого следующие действия.

**6 этап**. Визуально определяется для каждой ранжировки Hj – количество связок и поместить их в соответствующие ячейки диапазона E12:G12; также определить $h_k$ – размер связок в *k* –ой группе, поместив значения в диапазон ячеек E13:G14.

**7 этап**. Вычисляется выражение $T_j$, которое используется для корректировки коэффициента конкордации в случае наличия связок. Формулы, которые необходимо ввести в соответствующие ячейки, приведены в табл. 3.1.

Таблица 3.1

Формулы рабочего листа для расчета корректирующего выражения

| | **E** | **F** | **G** |
|---|---|---|---|
| **15** | =E13^3-E13 | =F13^3-F13 | =G13^3-G13 |
| **16** | =E14^3-E14 | =F14^3-F14 | =G14^3-G14 |
| **17** | =СУММ(E15:E16) | =СУММ(F15:F16) | =СУММ(G15:G16) |

**8 этап.** Определяется средний ранг по формуле =СУММ(E3:G9)/B11, результат помещается в ячейку G18. Также находится значение величины *S* сначала по объектам, например, по 1-

му ячейка I3 =(СУММ(E3:G3)-$G$18)^2 , далее значения суммируются, определяется S в ячейке I10.

Значение коэффициента конкордации с учетом связок рассчитывается в ячейке I11 по формуле = 12*I10/(B10^2*(B11^3-B11)-B10*СУММ(E17:G17)). В результате коэффициент конкордации равен Wкор = 0,4644.

Учитывая вышеизложенные данные, строится график согласованности мнений экспертов (рис 3.2).

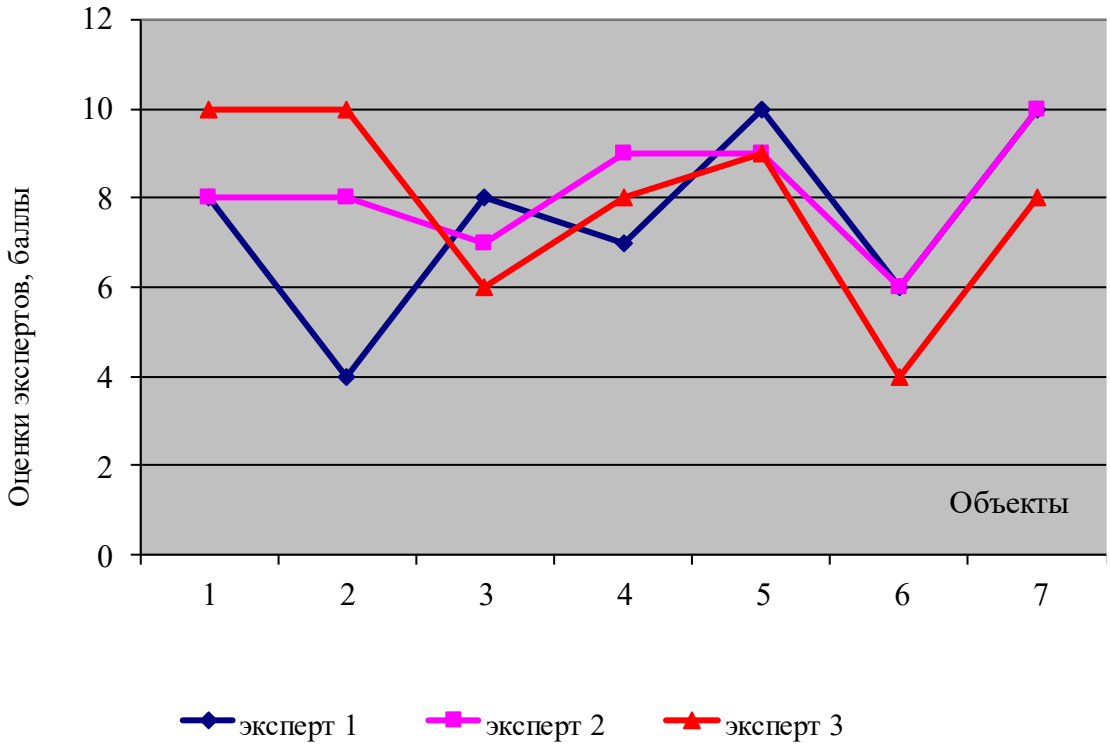



Рис. 3.2. Иллюстрация несогласованности мнений экспертов

По рисунку видно, что мнения экспертов не согласуются, так как по объектам значительно варьируются мнения экспертов и Wкор = 0,464<0,6.

# Глава 4. РЕГРЕССИОННЫЙ АНАЛИЗ

В экономических исследованиях нередко возникает необходимость описать характер взаимосвязи между различными показателями. В отличие от жесткой детерминированной (функциональной) связи, при которой каждому значению переменной X соответствует строго определенное значение Y, в реальных экономических системах преобладают вероятностные (стохастические) связи.

Регрессионный анализ – это раздел математической статистики, предназначенный для изучения именно таких связей. Его основная задача – построить математическую модель, которая отражает зависимость среднего значения зависимой переменной Y от одного или нескольких факторов (независимых переменных $X_1$, $X_2$, ..., $X_n$).

Особенность регрессионной модели заключается в учете случайной составляющей. Это означает, что при фиксированном значении X значение Y может варьировать из-за влияния неучтенных факторов, ошибок измерения или внутренней случайной природы процесса. Например, зная уровень дохода домохозяйства (X), можно предсказать его средние расходы (Y), но конкретные расходы конкретного домохозяйства могут отличаться из-за личных предпочтений или внешних условий.

Математически модель записывается так:

$Y = f(X) + \varepsilon$

где f(X) – систематическая часть (функция регрессии), а ε – случайная ошибка или возмущение.

Основные задачи регрессионного анализа:

1. Выявить факт наличия зависимости между переменными и определить ее форму (линейную или нелинейную).

2. Оценить силу связи (насколько изменение X влияет на Y).

3. Построить регрессионное уравнение для оценки средних значений Y.

4. Проверить адекватность модели и значимость ее параметров.

Регрессия – это зависимость среднего значения какой-либо случайной величины от некоторой другой величины или нескольких величин. При регрессионной связи одному и тому же значению величины *X* (в отличие от функциональной связи) могут соответствовать разные случайные значения величины *Y*. Основное отличие от экстраполяции в том, что последняя является определением будущих,

ожидаемых значений экономических величин, показателей на основе имеющихся данных об их изменении в прошлые периоды; перенесением прошлого на будущее, исходя из выявленных в прошлом тенденций изменения. Математически экстраполяция сводится к продолжению кривой, характеризующей предыдущее изменение экономического показателя.

В ходе изучения лабораторной работы необходимо рассмотреть теоретический материал проведения регрессионного анализа, решить задачу согласно выбранному варианту и составить отчет по установленным требованиям, содержащий пояснения результатов выполненного прогноза. Решение комплексной задачи проводится на основе представленного практического примера.

Методические положения проведения регрессионного анализа.

**1 этап.** Первым этапом составления прогноза проводится анализ зависимости между двумя переменными с помощью метода наименьших квадратов. Для наглядного изображения исходных данных, дальнейшего анализа и прогнозирования составляется диаграмма рассеивания исходных данных. Оценивается выборочный коэффициент корреляции, по результатам расчетов необходимо сделать соответствующие выводы.

**2 этап**. Построение прямой регрессии с помощью метода наименьших квадратов.

Для набора пар данных *X – Y* в качестве *прямой наилучшего приближения* будет выбираться такая, для которой наименьшее значение принимает сумма квадратов расстояний от точек *(x, y)* из заданного набора данных до этой прямой, измеренных в вертикальном направлении (по оси *Y)*. Эта прямая называется прямой регрессии, а ее уравнение – уравнением регрессии.

Уравнение прямой приближения имеет вид $\hat{Y} = b_0 + b_1 X$. Первый параметр $b_0$ называется свободным членом, а второй $b_1$ – угловым коэффициентом, отражающим величину, на которую изменяется значение *Y* при увеличении *X* на единицу. Таким образом, необходимо определить данные параметры.

Построение прямой регрессии проводится с помощью критерия наименьших квадратов.

$$SSE = \sum (Y2 - \hat{Y})^2 = \sum (Y - b_0 - b_1 X)^2 \quad (4.1)$$

$$b_1 = \frac{n * \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2} = \frac{\sum (X - \overline{X}) \sum (Y - \overline{Y})}{\sum (X - \overline{X})^2}, \quad (4.2)$$

$$b_0 = \frac{\sum Y}{n} - \frac{b_1 \sum X}{n} = \overline{Y} - b_1 \overline{X}, \quad (4.3)$$

где $b_0$ - свободный член;

$b_1$ - угловой коэффициент;

SSE – сумма квадратов ошибок.

Как можно предположить, значение углового коэффициента $b_1$ связано с выборочным коэффициентом корреляции. В данном случае получается следующее:

$$b_1 = \frac{\sqrt{\sum (Y - \overline{Y})^2}}{\sqrt{(X - X)^2}} r \cdot \quad (4.4)$$

Значит $b_1$ и $b_0$ пропорциональны друг другу и имеют один и тот же знак.

Разности между фактически полученными значениями Y и вычисленными по уравнению регрессии соответствующими значениями прогнозов $\hat{Y}$ называются *отклонениями.* Отклонения – это расстояния по вертикали (положительные или отрицательные) от точек, отмеченных по исходным данным, до прямой регрессии.

Можно сказать, что величины прогноза являются моделируемыми значениями рассматриваемых данных, а отклонения показывают отличие от ожидаемой модели. Разделение на прогноз и отклонение применяется и в тех ситуациях, когда рассматривается модель, отличная от прямой линии.

В модели простой линейной регрессии зависимая величина Y является суммой ее математического ожидания и случайного отклонения ε. Значения ε отражают возможную вариацию величин Y, в них скрыто влияние различных ненаблюдаемых факторов.

**3 этап.** Определение стандартной ошибки оценки.

Имея прямую регрессии, можно определить, насколько сильно точки исходных данных отклоняются от прямой регрессии. Можно выполнить оценку разброса, аналогичную стандартному отклонению

выборки. Этот показатель, называемый *стандартной ошибкой оценки,* измеряет степень отличия реальных значений *Y* от оцененной величины $\hat{Y}$ . Она обозначается через $S_{y*x}$ и вычисляется по следующей формуле:

$$S_{y*x} = \sqrt{\frac{\Sigma(Y-\bar{Y})^2}{n-2}} \ . \qquad (4.5)$$

Стандартная ошибка оценки подобна стандартному отклонению. Ее можно использовать для оценки стандартного отклонения совокупности. Фактически $S_{y*x}$ оценивает стандартное отклонение σ слагаемого ошибки в статистической модели простой линейной регрессии. Другими словами $S_{y*x}$ оценивает общее стандартное отклонение σ нормального распределения значений *Y*, имеющих математические ожидания $\mu_y = \beta_0 + \beta_1 X + \varepsilon$ для каждого *X*.

Если стандартная ошибка оценки велика, точки данных могут значительно удаляться от прямой.

Для удобства вычислений уравнение (4.5) можно привести к следующему виду:

$$S_{y*x} = \sqrt{\frac{\Sigma Y^2 - b_0 \Sigma Y - b_1 \Sigma XY}{n-2}} \ . \qquad (4.6)$$

**4 этап.** Прогнозирование величины *Y*.

Регрессионную прямую можно использовать для оценки величины переменной *Y* при данных значениях переменной X. Чтобы получить *точечный прогноз,* или предсказание для данного значения *X*, необходимо вычислить значение найденной функции регрессии в точке *X*.

Есть два источника неопределенности в точечном прогнозе, использующем уравнение регрессии.

1. Неопределенность, обусловленная отклонением точек данных от выборочной прямой регрессии.

2. Неопределенность, обусловленная отклонением выборочной прямой регрессии от регрессионной прямой генеральной совокупности.

*Интервальный прогноз* значений переменной *Y* можно построить так, что при этом будут учтены оба источника неопределенности.

*Стандартная ошибка прогноза* $S_f$ дает меру вариативности предсказанного значения *Y* около истинной величины *Y* для данного значения *X*. Стандартная ошибка прогноза равна следующему:

$$S_f = \sqrt{S^2 y*x + S^2 y*x(\frac{1}{n} + \frac{(X-\bar{X})^2}{\sum(X-\bar{X})^2})} \; ; \tag{4.7}$$

$$S_f = S_{y*x}\sqrt{1+\frac{1}{n} + \frac{(X-\bar{X})^2}{\sum(X-\bar{X})^2})} \quad . \tag{4.8}$$

Первое слагаемое $S^2 y*x$ под первым радикалом в уравнении 5.7 дает меру отклонения точек данных от выборочной прямой регрессии (первый источник неопределенности). Второе слагаемое $S^2 y*x$ измеряет отклонение выборочной прямой регрессии от регрессионной прямой генеральной совокупности (второй источник неопределенности). Отметим, что стандартная ошибка прогноза зависит от значения *X*, для которого прогнозируется величина *Y.* Также следует отметить, что $S_f$ минимально, когда $X = \bar{X}$ , поскольку тогда числитель в третьем слагаемом под корнем в уравнении 4.7 будет $(X-\bar{X})^2 = 0$ . При прочих неизменных величинах большему отличию *X* от $\bar{X}$ соответствует большее значение стандартной ошибки прогноза.

Если статистическая модель простой линейной регрессии соответствует действительности, границы интервала прогноза величины *Y* равны следующему:

$$\hat{Y} \pm ts_f \quad , \tag{4.9}$$

где *t* – квантиль распределения Стьюдента с *n-2* степенями свободы *(df=n-2).*

Если выборка велика (n≥30), этот квантиль можно заменить соответствующим квантилем стандартного нормального распределения. Например, для большой выборки 95%-ный интервал прогноза задается следующими значениями:

$$\hat{Y} \pm 2s_f \quad . \tag{4.10}$$

**5 этап.** Разложение дисперсии.

Из уравнения можно выявить следующее:

$$Y = \hat{Y} + (Y - \hat{Y}) \text{ или } Y = (b_0 + b_1 X) + (Y - b_0 - b_1 X) \quad (4.11)$$

Наблюдаемое значение $Y$ — Объясненное линейной зависимостью — Остаток или отклонение от линейной зависимости

В идеале, когда все точки лежат на прямой регрессии, все остатки равны нулю и значения $Y$ полностью вычисляются или объясняются линейной функцией от $X$.

Отнимая $\hat{Y}$ от обеих частей предыдущего равенства, имеется следующее:

$$Y - \bar{Y} = (Y - \hat{Y}) + (\hat{Y} - \bar{Y}) \; . \quad (4.12)$$

Несложными алгебраическими преобразованиями можно показать, что суммы квадратов складываются:

$$\sum (Y - \bar{Y})^2 = \sum (Y - \hat{Y})^2 + \sum (\hat{Y} - \bar{Y})^2 \quad (4.13)$$

или

$$SST = SSR + SSE \; , \quad (4.14)$$

где $\text{SST} = \sum (Y - \bar{Y})^2$, $\text{SSR} = \sum (\hat{Y} - \bar{Y})^2$, $\text{SSE} = \sum (Y - \hat{Y})^2$ .

Здесь SS обозначает "сумма квадратов" (Sum of Squares), а *T, R, E* – соответственно "общая" (Total), "регрессионная" (Regression) и "ошибки" (Error). С этими суммами квадратов связаны следующие величины степеней свободы:

- $df$ (SST) = n-1;
- $df$ (SSR) = n;
- $df$ (SSE) = n-2.

Так же, как и суммы квадратов, степени свободы связаны следующим соотношением.

$$n - 1 = 1 + (n\text{-}2) \; . \quad (4.15)$$

Если линейной связи нет, $Y$ не зависит от $X$ и дисперсия $Y$ оценивается значением выборочной дисперсии:

$$S_y^2 = \frac{1}{n-1} \sum (Y - \bar{Y})^2 \; . \quad (4.16)$$

Если, с другой стороны, связь между *X* и *Y* имеется, она может влиять на некоторые разности значений *Y.*

Регрессионная сумма квадратов, SSR, измеряет часть дисперсии *Y*, объясняемую линейной зависимостью. Сумма квадратов ошибок, SSE, – это оставшаяся часть дисперсии *Y*, или дисперсия *Y*, не объясненная линейной зависимостью.

**Разложение дисперсии**

| SST | = | SSR | + | SSE |
|---|---|---|---|---|
| Общая изменчивость *Y* | | Изменчивость, объясненная линейной зависимостью | | Остаток, или необъясненная изменчивость |

Суммы квадратов, связанные с разложением изменчивости *Y*, и их соответствующие величины степеней свободы могут быть размещены так, как показано в табл. 4.1, известной как *таблица анализа дисперсии* или *таблица ANOVA* (ANalisis Of VArianse).

Таблица 4.1

Таблица *ANOVA* для прямолинейной регрессии

| Источник | Сума квадратов | Степени свободы | Среднеквадратическое отклонение |
|---|---|---|---|
| Регрессия | SST | 1 | MSR = SSR / 1 |
| Ошибки | SSE | n – 2 | MSE = SSE / (n-2) |
| Общая | SSR | n – 1 | |

Последний столбец таблицы *ANOVA* – это *среднеквадратичные значения.* Среднеквадратичное регрессии, MSR – это регрессионная сумма квадратов, разделенная на их величину степеней свободы. Аналогично среднеквадратичное ошибок, *MSE* – это сумма квадратов ошибок, разделенная на их величину степеней свободы.

Из уравнения 4.8 имеется следующее:

$$MSE = \frac{SSE}{n-2} = \frac{\sum (Y - \hat{Y})^2}{n-2} = S^2{}_{y*x} \quad , \qquad (4.17)$$

т.е. равенство *MSE* квадрату стандартной ошибки оценки. Отношение среднеквадратичных значений будет использовано для другой цели в этой главе дальше.

**6 этап.** Определение коэффициента детерминации.

*Коэффициент детерминации* измеряет долю изменчивости *Y,* которую можно объяснить с помощью информации об изменчивости (разнице значений) независимой переменной *X.*

Тождество $Y - \overline{Y} = (Y - \hat{Y}) + (\hat{Y} - \overline{Y})$ (формула 4.14) приводит к разбиению дисперсии, данному в уравнении 4.15. Для регрессионной прямой данных проводимого прогноза гипотетических точек данных разбиение графически представлено на рис. 4.1.

Если величина *Y* не зависит от *X*, специалисту следует ожидать значения *Y*, близкие к $\overline{Y}$, а разности $Y - \overline{Y}$ просто отражают случайные отклонения. Однако в действительности величина Y зависит от *X*, что демонстрируется функцией регрессии. На рисунке взято значение *X,* большее $\overline{X}$ , и известно, что *X* и *Y* имеют значительную отрицательную корреляцию (r = -0,86). Общее расстояние по вертикали равно $Y - \overline{Y}$, величина $\hat{Y} - \overline{Y}$, следовательно "объясняется" изменением *X*, тогда как оставшееся по вертикали расстояние $Y - \hat{Y}$ "не объясняется" изменением *X.*

Показатель SST измеряет общую вариацию относительно $\overline{Y}$, а ее часть, объясненная изменением *X*, соответствует SSR. Оставшаяся, или необъясненная вариация соответствует SSE. Отношение объясненной вариации к общей называется выборочным *коэффициентом детерминации* и обозначается $r^2$.

$$r^2 = \frac{\text{объясненная\_вариация}}{\text{общая\_вариация}} = \frac{SSR}{SST} = \frac{\sum(\hat{Y} - \overline{Y})^2}{\sum(Y - \overline{Y})^2} = 1 - \frac{\text{объясненная\_вариация}}{\text{общая\_вариация}} = 1 - \frac{SSR}{SST} = 1 - \frac{\sum(\hat{Y} - \overline{Y})^2}{\sum(Y - \overline{Y})^2} \quad (4.18)$$

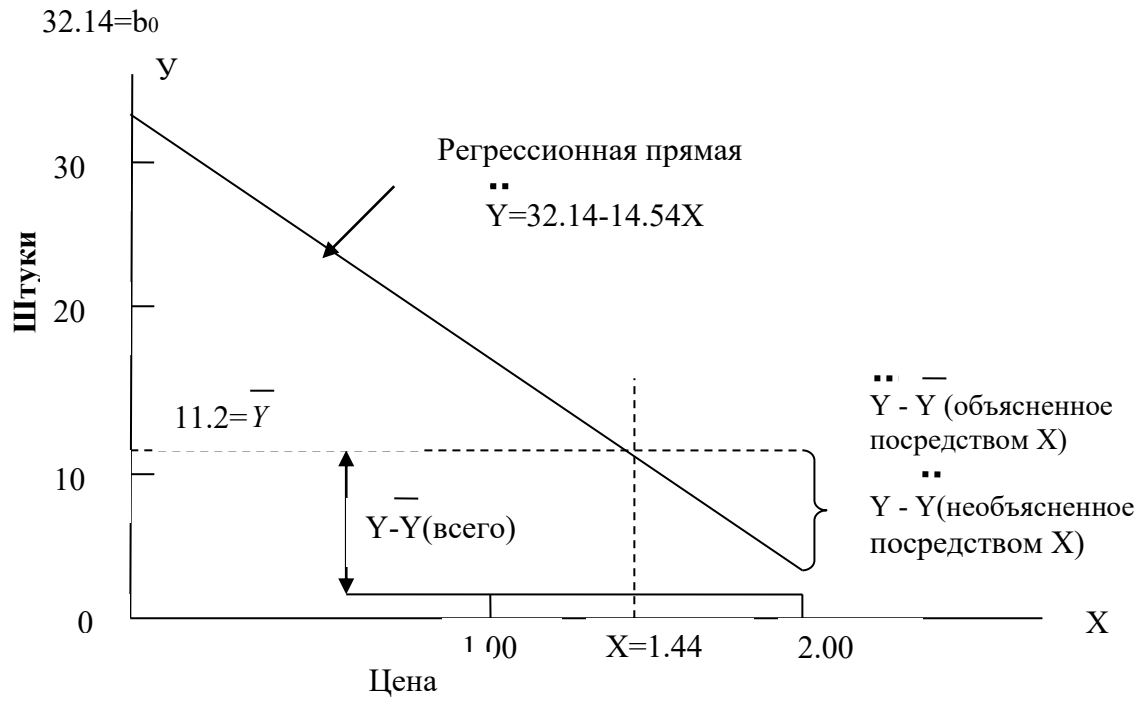



Рис. 4.1. Объясненная и необъясненная дисперсии для данных прогноза

3. Практический пример построения прогноза на основе регрессионного анализа.

Менеджер отдела закупок сети супермаркетов анализирует зависимость спроса на фрукты (мандарины) от их розничной цены. Для исследования были случайным образом отобраны данные за 10 дней торговли в одном из магазинов сети. Необходимо построить регрессионную модель, оценить тесноту связи между ценой и объемом продаж, а также спрогнозировать возможный объем реализации при установлении цены на уровне 1,4 усл. ед.

Таблица 4.2

Данные о продаже мандаринов

| День торговли | Объем продаж (кг) – Y | Цена за 1 кг (усл. ед.) – X |
|---|---|---|
| 1 | 110 | 1,2 |
| 2 | 65 | 1,9 |
| 3 | 52 | 1,6 |
| 4 | 125 | 1,4 |
| 5 | 104 | 1,5 |
| 6 | 158 | 1,1 |
| 7 | 48 | 1,5 |
| 8 | 122 | 1,3 |
| 9 | 175 | 0,9 |
| 10 | 208 | 1,0 |

**Решение.**

**Этап 1.** Для наглядного изображения исходных данных и дальнейшего анализа и прогнозирования составляется диаграмма рассеивания для исходных данных, представленная на рис. 4.2.

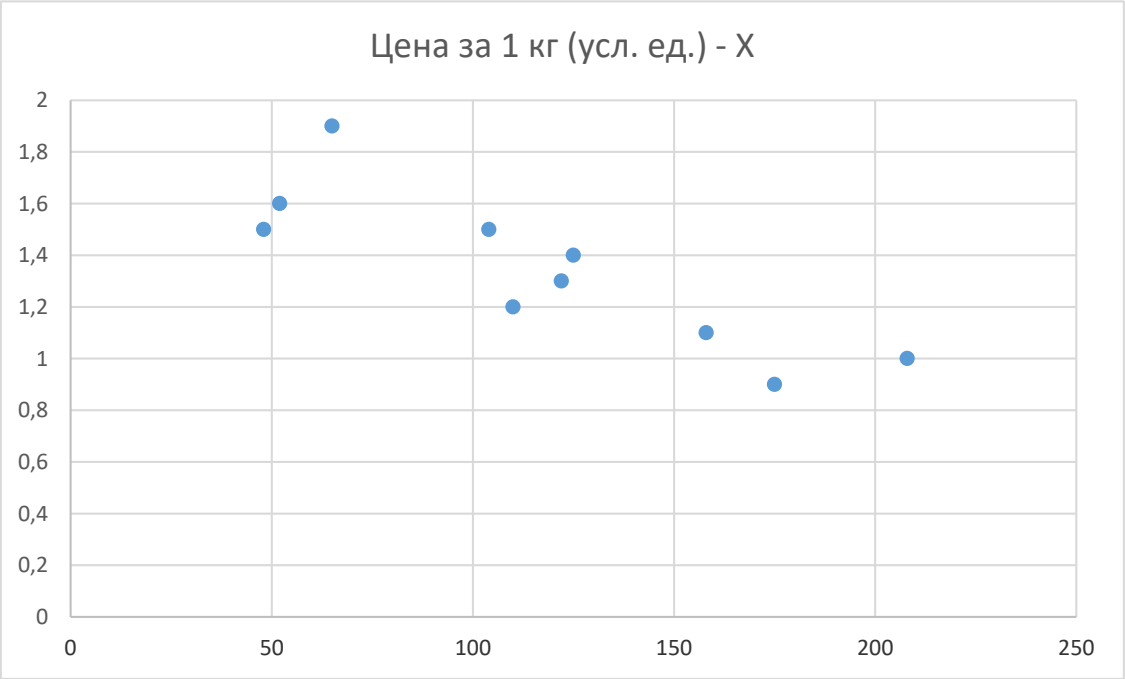



Рис. 4.2. Диаграмма рассеивания

Диаграмма показывает, что имеет место обратная линейная зависимость между переменной Y (объемом продаж) и переменной $X$ (цена за 1 кг). Можно сделать вывод, что при возрастании цены объем продаж уменьшается.

Таким образом, далее целесообразно оценить количественную меру обнаруженной зависимости. Для этого вычисляется выборочный коэффициент корреляции на основе формулы 4.19.

$$r = \frac{n * \sum XY - (\sum X)(\sum Y)}{\sqrt{n\sum X^2 - (\sum X)^2} * \sqrt{n\sum Y^2 - (\sum Y)^2}} \quad . \qquad (4.19)$$

Вспомогательные расчеты представляются в табл. 4.3.

Таблица 4.3

Расчеты коэффициента корреляции

| n = 10 | Y | X | XY | $Y^2$ | $X^2$ |
|---|---|---|---|---|---|
| 1 | 110 | 1,2 | 132,0 | 12100 | 1,44 |
| 2 | 65 | 1,9 | 123,5 | 4225 | 3,61 |
| 3 | 52 | 1,6 | 83,2 | 2704 | 2,56 |
| 4 | 125 | 1,4 | 175,0 | 15625 | 1,96 |
| 5 | 104 | 1,5 | 156,0 | 10816 | 2,25 |
| 6 | 158 | 1,1 | 173,8 | 24964 | 1,21 |
| 7 | 48 | 1,5 | 72,0 | 2304 | 2,25 |
| 8 | 122 | 1,3 | 158,6 | 14884 | 1,69 |
| 9 | 175 | 0,9 | 157,5 | 30625 | 0,81 |
| 10 | 208 | 1,0 | 208,0 | 43264 | 1,00 |
| **Сумма** | **1167** | **13,4** | **1439,6** | **163511** | **18,78** |

$$r = \frac{10*1439{,}6 - 13{,}4*1167}{\sqrt{10*18{,}78 - 18{,}78^2} * \sqrt{10*163511 - 163511^2}} = -0{,}828.$$

Расчет коэффициента корреляции можно легко выполнить в Excel с помощью надстройки "Анализ данных" → "Корреляция". По результатам этого анализа полученное значение выборочного коэффициента корреляции, равное -0,828, свидетельствует о сильной обратной связи между ценой товара (X) и объемом продаж (Y). Это означает, что при увеличении цены на мандарины количество проданных килограммов закономерно уменьшается, что полностью соответствует классическому закону спроса.

Тем не менее, коэффициент корреляции лишь показывает направление и степень связи, но не дает ответа на важный практический вопрос: на сколько именно килограммов снизятся продажи при увеличении цены на одну условную единицу?

Для получения этого ответа необходимо построить регрессионную модель. На диаграмме рассеяния можно провести линию тренда, которая максимально точно пройдет среди точек наблюдений. Наклон этой линии (коэффициент регрессии, обозначаемый как b) покажет, на сколько единиц (в килограммах) в среднем изменится объем продаж (Y) при увеличении цены (X) на одну условную единицу.

Ожидаемый знак коэффициента b будет отрицательным, подтверждая обратную зависимость, а его абсолютное значение позволит

количественно оценить чувствительность спроса к ценовым изменениям.

**Этап 2.** Необходимо провести требуемую прямую, ориентируя ее визуально так, чтобы она максимально приближалась к отмеченным на диаграмме точкам. Делать это можно различными способами. Однако важно выбрать такой метод определения прямой наилучшего приближения, который обеспечит одинаковый результат у любого человека при анализе одних и тех же данных. Для однозначного определения этой прямой чаще всего используют критерий наименьших квадратов.

При помощи метода наименьших квадратов рассчитываются оценки коэффициентов регрессии. Эти вычисления осуществляются на базе уравнений 4.3 и 4.4, а также числовых значений, приведенных в табл. 4.3. В результате определяется следующее:

$$b_1 = \frac{10*1439,6 - 13,4*1167}{10*18,78 - 13,4^2} = -150,7,$$

$$b_0 = 116,7 - (-150,7)*1,34) = 318,7.$$

Тогда уравнение прямой регрессии, определенное по методу наименьших квадратов, будет иметь следующий вид:

$$\hat{Y} = 318,7 - 150,7X. \qquad (4.21)$$

Свободный член b0 в регрессионной модели по сути представляет собой значение зависимой переменной YYY, если бы цена товара XXX была равна нулю. Согласно уравнению, если бы мандарины отдавались бесплатно (то есть цена – 0), средний объем продаж составлял бы 318,7 кг. Однако такая интерпретация не имеет практического смысла и противоречит здравому смыслу по нескольким причинам. Во-первых, в анализируемом диапазоне цен (от 0,9 до 1,9 условных единиц) данных о ценах, близких к нулю, нет. Экстраполяция модели за пределы наблюдаемых значений может привести к некорректным результатам. Во-вторых, в реальности цена не бывает нулевой, и поведение спроса при бесплатной раздаче может значительно отличаться от закономерностей, выявленных в рамках платного сегмента. Также следует учитывать, что модель отражает зависимость только внутри диапазона цен от 0,9 до 1,9 условных единиц – за его пределами прогнозы требуют осторожности, поскольку характер связи может измениться. Поэтому значение свободного члена – это ско-

рее расчетный коэффициент, который обеспечивает правильное расположение линии регрессии относительно данных, и не подлежит прямой экономической интерпретации.

Коэффициент регрессии b1= −150,7 имеет четкое экономическое значение. Он показывает, насколько в среднем изменится объем продаж при увеличении цены на одну условную единицу. В данном случае, увеличение цены на один рубль или условную единицу ведет к снижению продаж примерно на 151 килограмм. Отрицательный знак коэффициента подтверждает обратную зависимость между ценой и спросом, что полностью соответствует классической теории – повышение цены вызывает сокращение объема продаж. Эта количественная мера чувствительности спроса позволяет понять, что каждое увеличение цены на одну условную единицу в среднем приводит к потере примерно 151 килограмма продаж.

Практическая ценность полученной модели заключается в возможности оценивать влияние изменения цены на объем продаж. Зная коэффициент, можно прогнозировать снижение продаж при планируемом повышении цены. Так, например, при повышении цены на 0,5 условных единицы ожидается снижение объема продаж примерно на 75 килограммов. Также модель помогает в оптимизации ценовой политики – сопоставляя потери в объеме и возможный рост дохода, можно определить такую цену, которая максимизирует прибыль. Кроме того, она дает возможность более точно планировать закупки, исходя из ожидаемых изменений спроса при различных ценовых стратегиях.

Следует учитывать, что модель ограничена по применению: она отражает зависимость только в диапазоне цен от 0,9 до 1,9 условных единиц. За его пределами характер связи может измениться. Также она учитывает только влияние цены и не включает влияние других факторов, таких как доходы населения, сезонность, цены на заменяющие товары или рекламные акции, которые могут существенно влиять на спрос. Кроме того, полученные коэффициенты являются оценками, основанными на выборке, и могут варьироваться при изменениях данных или состава выборки.

Зависимость между переменными X и Y можно визуально продемонстрировать на диаграмме рассеивания, на которой изображена линия, представляющая собой наилучшее приближение этой связи (рис. 4.3).

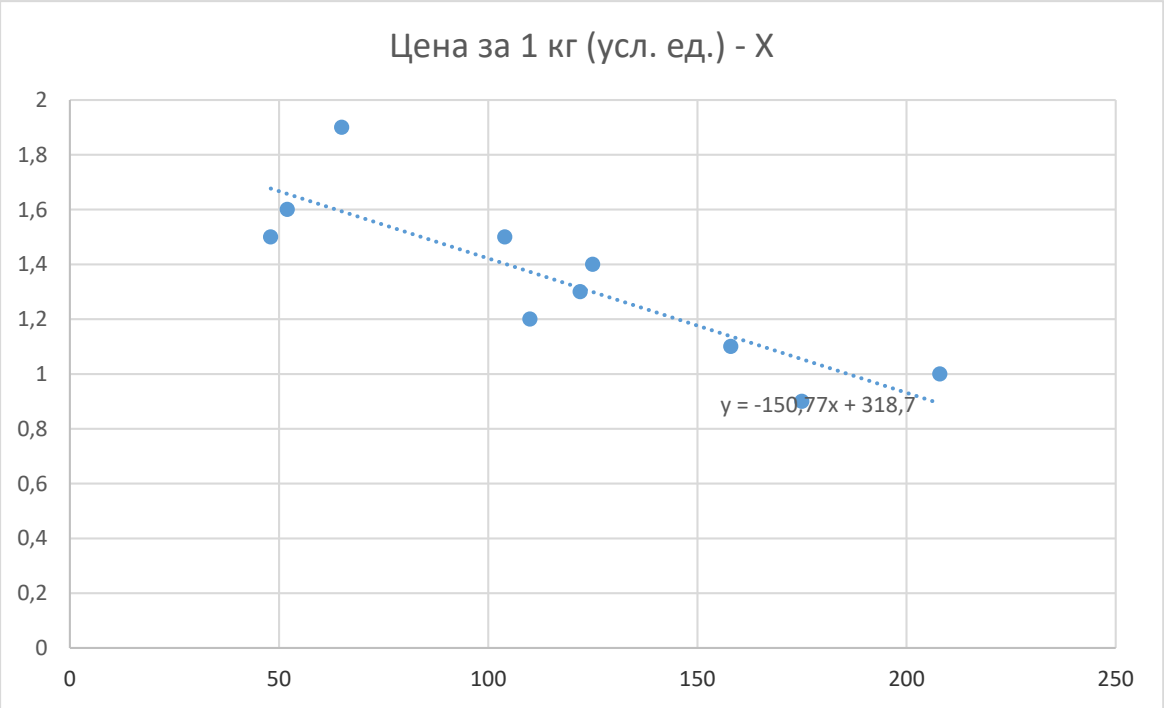



Рис. 4.3. Данные прогноза

Обратите внимание на то, что вертикальные отрезки от точек данных до прямой проведены пунктиром. Сумма квадратов длин отрезков, проведенных к этой прямой, должна быть меньше аналогичной суммы квадратов длин, проведенных к любой другой прямой. (Для данных специалиста ПЭО сумма квадратов длин равна SSE = 59,14). Из метода наименьших квадратов следует, что данная прямая является наилучшим приближением для заданных 10 точек исходных данных.

**Этап 3.** Определение стандартной ошибки.

Для данных специалиста ПЭО стандартная ошибка оценки равна следующему:

$$S_{y*x} = \sqrt{\frac{163511 - 318{,}7*1167 - (-150{,}7)*1439{,}6}{10-2}} = \sqrt{1066{,}9} = 32{,}66.$$

Полученное значение стандартной ошибки регрессии Syx= 32,66 кг свидетельствует о том, что, хотя цена является важным фактором, определяющим спрос на мандарины, она не единственная. Существенная доля вариации объема продаж объясняется иными причинами. Вопрос о статистической значимости модели и возможности ее улучшения будет исследован на следующих этапах анализа (проверка

значимости коэффициентов, анализ остатков, расчет доверительных интервалов).

**Этап 4.** Прогнозирование величины *Y*.

Предположим, что менеджер торговой сети хочет получить прогноз количества мандаринов (в кг), которое будет продано при установлении цены 1,63 усл. ед. за килограмм**.**

Используя полученное ранее уравнение регрессии:

Y=318,7−150,7·X*Y*=318,7−150,7·*X*

подставляем значение X=1,63:

Y^=318,7−150,7×1,63=318,7−245,64=73,06 кг*Y*^=318,7−150,7×1,63=318,7−245,64=73,06 кг

Таким образом, точечный прогноз объема продаж составляет 73,06 кг**.**

Данный прогноз – это значение величины Y*Y*, соответствующее координате X=1,63 на регрессионной прямой.

Графически 95%-ный интервал прогноза значений *Y* для данных специалиста представлен на рис. 4.4.

Однако важно понимать, что реальные значения величины Y*Y*, соответствующие рассматриваемым значениям X*X*, не лежат в точности на регрессионной прямой. Фактически они разбросаны относительно прямой в соответствии с величиной стандартной ошибки регрессии Syx=32,66 кг.

Более того, построенная нами выборочная регрессионная прямая является лишь **оценкой** истинной регрессионной прямой генеральной совокупности, основанной на выборке всего из 10 пар данных. Другая случайная выборка из 10 пар данных дала бы иную выборочную прямую регрессии. Это аналогично ситуации, когда различные выборки из одной и той же генеральной совокупности дают различные значения выборочного среднего.

Для учета этой неопределенности строится **интервальный прогноз** – доверительный интервал, который с заданной вероятностью (обычно 95%) накрывает истинное значение Y*Y* для заданного X*X* (рис. 4.4).

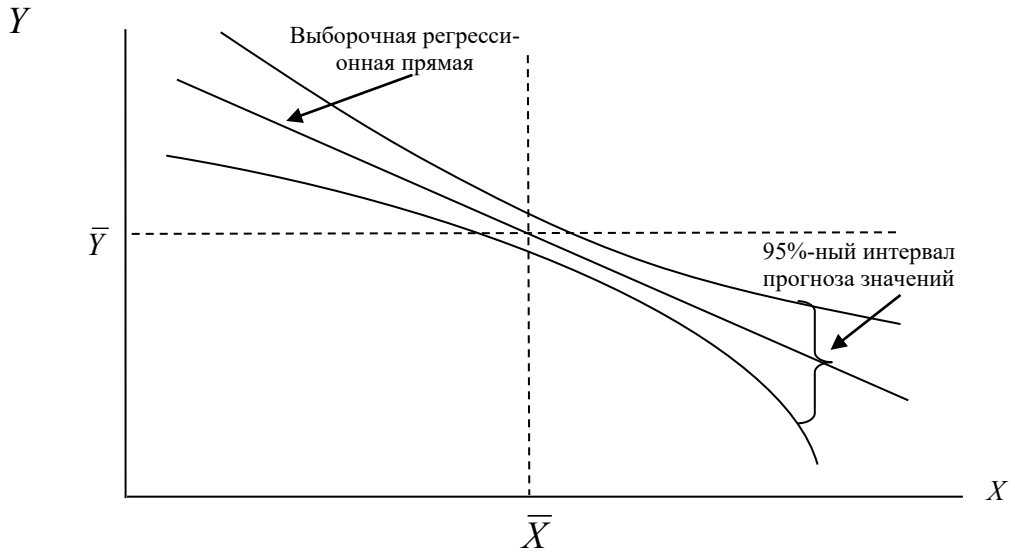



Рис. 4.4. 95%-ный интервал прогноза значений $Y$

Используя результаты из табл. 4.3 и уравнения 4.11, определяется стандартная ошибка прогноза в точке $X$=1,63 .

Таблица 4.4

Расчет стандартной ошибки прогноза

| $X$ | $(X-\overline{X})$ |
|---|---|
| 1,2 | 0,0196 |
| 1,9 | 0,3136 |
| 1,6 | 0,0676 |
| 1,4 | 0,0036 |
| 1,5 | 0,0256 |
| 1,1 | 0,0576 |
| 1,5 | 0,0256 |
| 1,3 | 0,0016 |
| 0,9 | 0,1936 |
| 1,0 | 0,1156 |
| | $\sum(X-\overline{X})^2 = 0{,}8240$ |

$$S_f = 32{,}66\sqrt{1+\frac{1}{10}+\frac{(1{,}63-1{,}34)^2}{0{,}824}} = 35{,}79\,.$$

Доверительный интервал прогноза (95%)

Табличное значение t0,025;8=2,306*t*0,025;8=2,306.

Y^±t·Spred*Y*^±*t*·*Spred*73,06±2,306×35,7973,06±2,306×35,792,306×35,79≈82,522,306×35,79≈82,52

Границы интервала:

Нижняя: 73,06−82,52=−9,4673,06−82,52=−9,46 кг → 0 кг (объем не может быть отрицательным)

Верхняя: 73,06+82,52=155,5873,06+82,52=155,58 кг

Таким образом, 95% доверительный интервал прогноза:

(0; 155,58) кг(0;155,58) кг.

**Этап 5.** Разложение дисперсии.

Товаровед начал свой анализ данных с информации об объемах продаж только за 10 недель (переменная Y). Если другой информации не поступит, он может использовать выборочное среднее Y=11,2 как прогноз количества продаваемых мандаринов для каждой недели. Ошибки или отклонения, связанные с этим прогнозом, равны $Y - \hat{Y}$, и сумма квадратов ошибок даст $\sum(Y-\hat{Y})^2$ . Последнее значение, $\sum(Y-\hat{Y})^2$, в точности равно SST, обшей сумме квадратов, введенной в уравнение 5.10. Таким образом, SST измеряет отклонение значения Y от прогноза, использующего лишь значения Y в его вычислении. (Если анализ остановить на этом этапе, отклонения Y следует измерять выборочной дисперсией $S^2{}_y = \sum(Y-\bar{Y})^2/(n-1)$ вместо SST=$\sum(Y-\bar{Y})^2$. Выборочная дисперсия является обычной мерой изменчивости наблюдений одной переменной.) Прогноз величины $\bar{Y}$, значения отклонения Y – $\bar{Y}$ суммы квадратов SST=$\sum(Y-\bar{Y})^2$ приведены в табл. 4.5. (Сумма отклонений Y – $\bar{Y}$ всегда равна нулю, поскольку среднее $\bar{Y}$ является математическим центром значений Y).

Таблица 4.5

Отклонения для данных прогноза и значения прогноза $\bar{Y}$

| № наблюдения | Фактический объем продаж (кг) | Средний объем продаж (кг) | Отклонения (Y- $\bar{Y}$ ) | (Y- $\bar{Y}$ )$^2$ |
|---|---|---|---|---|
| 1 | 110 | 116,7 | -6,7 | 44,89 |
| 2 | 65 | 116,7 | -51,7 | 2 672,89 |
| 3 | 52 | 116,7 | -64,7 | 4 186,09 |
| 4 | 125 | 116,7 | 8,3 | 68,89 |
| 5 | 104 | 116,7 | -12,7 | 161,29 |
| 6 | 158 | 116,7 | 41,3 | 1 705,69 |
| 7 | 48 | 116,7 | -68,7 | 4 719,69 |
| 8 | 122 | 116,7 | 5,3 | 28,09 |
| 9 | 175 | 116,7 | 58,3 | 3 398,89 |
| 10 | 208 | 116,7 | 91,3 | 8 335,69 |

Товаровед также имеет информацию о значениях переменной X$X$ (цене за 1 кг мандаринов), соответствующих величинам Y$Y$ (объему продаж). Коэффициент корреляции между ценой и объемом продаж составляет r=−0,828, что указывает на тесную обратную связь. Можно ожидать, что с помощью этой дополнительной переменной мы сможем объяснить часть изменчивости (разностей) значений Y$Y$, не объясненной простым средним прогнозом Y¯$\bar{Y}$.

По расчетам линейный прогноз пар значений X−Y$X-Y$ задается уравнением:

Y^=318,7−150,7·X$\hat{Y}=318,7-150,7\cdot X$

Таблица, подобная табл. 4.5, может быть построена при Y^$\hat{Y}$ в качестве прогноза значений Y$Y$. Результат приводится в табл. 4.6. (Если свободный член включен в уравнение регрессии, сумма отклонений ∑(Yi−Y^i)$\sum(Yi-\hat{Y}i)$ всегда равна нулю).

.

Таблица 4.6

Отклонения для данных при значении прогноза $\hat{Y}$

| № | Цена (усл. ед.) | Фактический объем (кг) | Прогноз | Отклонение | (Yi−Y^i)2(*Yi* −*Y^i*)2 |
|---|---|---|---|---|---|
| 1 | 1,2 | 110 | 318,7 – 150,7 × 1,2 = 318,7 – 180,84 = 137,86 | -27,86 | 776,18 |
| 2 | 1,9 | 65 | 318,7 – 150,7 × 1,9 = 318,7 – 286,33 = 32,37 | 32,63 | 1 064,72 |
| 3 | 1,6 | 52 | 318,7 – 150,7 × 1,6 = 318,7 – 241,12 = 77,58 | -25,58 | 654,34 |
| 4 | 1,4 | 125 | 318,7 – 150,7 × 1,4 = 318,7 – 210,98 = 107,72 | 17,28 | 298,60 |
| 5 | 1,5 | 104 | 318,7 – 150,7 × 1,5 = 318,7 – 226,05 = 92,65 | 11,35 | 128,82 |
| 6 | 1,1 | 158 | 318,7 – 150,7 × 1,1 = 318,7 – 165,77 = 152,93 | 5,07 | 25,70 |
| 7 | 1,5 | 48 | 318,7 – 150,7 × 1,5 = 318,7 – 226,05 = 92,65 | -44,65 | 1 993,62 |
| 8 | 1,3 | 122 | 318,7 – 150,7 × 1,3 = 318,7 – 195,91 = 122,79 | -0,79 | 0,62 |
| 9 | 0,9 | 175 | 318,7 – 150,7 × 0,9 = 318,7 – 135,63 = 183,07 | -8,07 | 65,12 |
| 10 | 1,0 | 208 | 318,7 – 150,7 × 1,0 = 318,7 – 150,7 = 168,00 | 40,00 | 1 600,00 |
| **Сумма** | **13,4** | **1 167** | **1 167,00** | **0,00** | **6 607,72** |

Сравнение табл. 4.5 и 4.6 показывает, что использование $\hat{Y}$ в качестве прогноза значения Y приводит, вообще говоря, к меньшим отклонениям (по абсолютной величине) и существенно меньшим суммам квадратов остатков (ошибок), чем применение для прогноза значения $\bar{Y}$. Использование соответствующих значений *X* уменьшает ошибку прогноза (предсказания). Таким образом, знание значений *X* помогает лучше объяснить разности *Y*. Но в какой мере может помочь знание значений *X*? Ответ на этот вопрос можно получить посредством разбиения изменчивости.

Используя данные из табл. 4.5, 4.6 и уравнение 4.14, имеется

$\text{SST}=\sum(Y-\bar{Y})^2=25322{,}10;$

$\text{SSE}=\sum(Y-\hat{Y})^2=6607{,}72$и, следовательно,

$\text{SSR}=\sum(\hat{Y}-\bar{Y})^2=\text{SST}-\text{SSE}=25322{,}10-6607{,}72=18714{,}38.$

Разбиение изменчивости является следующим:

| SST | = | SSR | + | SSE |
|---|---|---|---|---|
| Общая вариация | | Объясненная вариация | | Необъясненная вариация |

Для изменчивости, оставшейся после предсказания *Y* через значение $\bar{Y}$, специалист получил следующее значение:

$\frac{SSR}{SST}=0{,}739.$

Это та часть, которая объясняется взаимосвязью значений *Y* и *X*. Доля вариации *Y* относительно $\bar{Y}$, равная 1 – 0,739 = 0,261, осталась необъясненной. С этой точки зрения знание значений соответствующей переменной *X* приводит к лучшему прогнозу значений *Y*, чем прогноз, полученный из значения $\bar{Y}$, не зависящего от X.

Разбиение изменчивости для данных прогноза может быть представлено в таблице анализа дисперсии *ANOVA*, общий вид которой представлен в табл. 4.1., 4.7.

Таблица 4.7

Таблица *ANOVA* по данным прогноза

| Источник вариации | Сумма квадратов (SS) | Число степеней свободы (df) | Средний квадрат (MS) |
|---|---|---|---|
| Регрессия (объясненная) | SSR = 18 714,38 | 1 | MSR = 18 714,38 |
| Остатки (ошибки) | SSE = 6 607,72 | 8 | MSE = 825,97 |
| Общая вариация | SST = 25 322,10 | 9 | – |

Разбиение изменчивости ясно показано в столбце с суммами квадратов. Необходимо обратить внимание на то, что с учетом погрешности округления MSE=6607,72/8=825,97

**Этап 6.** Расчет коэффициента детерминации $r^2$ .

Для данных прогнозиста коэффициент был вычислен ранее. Значение коэффициента детерминации также можно легко получить из таблицы *ANOVA*, представленной табл. 4.7.

$$SST=\sum(Y-\bar{Y})^2=25322{,}10;\ SSR=\sum(\hat{Y}-\bar{Y})^2=18714{,}38;$$

$$SSE=\sum(Y-\hat{Y})^2=6607{,}72$$

и $$r^2=\frac{18714{,}38}{25322{,}10}=0{,}739.$$

Кроме того, $r^2$ можно вычислить через остаточную вариацию:

$$r^2=1-\frac{6607{,}72}{25322{,}10}=1-0{,}261=0{,}739.$$

Значение коэффициента детерминации R2 =0,739 показывает, что примерно 73,9% вариации объема продаж мандаринов (Y) можно объяснить изменениями в цене товара (X). Это означает, что большая часть изменений в продажах связана с колебаниями цены. Однако оставшиеся около 26,1% изменений не поддаются объяснению именно ценой. Эти остаточные вариации могут быть вызваны другими факторами, которые не были учтены в проведенном регрессионном анализе. Среди возможных причин – сезонные колебания (например, рост спроса перед праздниками), влияние рекламных акций и выкладки товара в торговых точках, качество конкретной партии мандаринов, наличие товаров-заменителей (апельсины, грейпфруты), уровень доходов покупателей, а также такие факторы, как день недели или погодные условия.

В случае прямолинейной регрессии коэффициент детерминации $r^2$ равен квадрату коэффициента корреляции *r:*

коэффициент детерминации = (коэффициент корреляции)$^2$,

$r^2 = (r)^2$.

Значит для данных специалиста, с учетом погрешности округления,

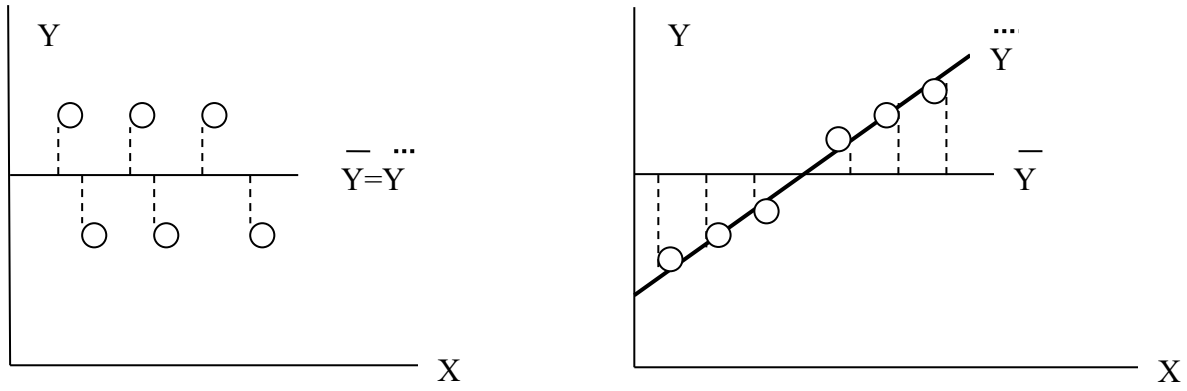
$(r)2=(-0,828)2=0,686$.

Обнаруживается небольшое расхождение: коэффициент детерминации R2=0,739 отличается от квадрата корреляционного коэффициента (r)2=0,686. Такое различие связано с погрешностями округления при вычислении коэффициентов регрессии и сумм квадратов. Теоретически эти значения должны совпадать, так как они отражают одну и ту же степень связи между переменными. Для дальнейшего анализа мы возьмем за основу значение R2=0,739, полученное из таблицы ANOVA, поскольку оно считается более точным.

Коэффициент корреляции *r* показывает не только степень связи между двумя переменными, но и её направление. В исследованных данных, собранных товароведом, обнаружена отрицательная связь: r=−0,828. Это логично с экономической точки зрения: при росте цены спрос на товар снижается. В других случаях значение r может указывать на положительную взаимосвязь. При работе с большим числом переменных иногда важно учитывать, в каком направлении связаны пары переменных. Следует помнить, что при возведении коэффициента корреляции в квадрат (r2) получается всегда положительное число, что лишает информации о направлении связи.

Коэффициент детерминации *R*2 показывает степень объяснения вариации зависимой переменной *Y* её связью с независимой *X*, в отличие от коэффициента корреляции *r*. Он отражает, какая часть изменчивости Y объясняется различиями в *X*. Это интерпретацию можно расширить и на случай множественной регрессии, когда *Y* зависит от нескольких переменных *X*.

На рис. 4.5 иллюстрируется два крайних случая для значения коэффициента $r^2$: $r^2 = 0$ и $r^2=1$. В случае *(а)* изменчивость *Y никак* не объясняется изменениями *X:* диаграмма рассеивания не показывает никакой линейной взаимосвязи между значениями величин *X* и *Y.* В случае (б), когда коэффициент $r^2 = 1$, изменчивость *Y* полностью объясняется, если известны значения *X:* все точки данных в нашей выборке лежат на прямой регрессии.




$r^2 = 1 - \sum(Y - \hat{Y})^2 / \sum(Y - \bar{Y})^2 = 1 - 1 = 0$ $r^2 = 1 - \sum(Y - \hat{Y})^2 / \sum(Y - \bar{Y})^2 = 1 - 0 = 1$

а) линейная корреляция отсутствует б) четко выраженная линейная корреляция

Рис. 4.5. Интерпретация крайних значений коэффициента детерминации $r^2$

*Примечание*. Выполненные расчеты для исходных и прогнозируемых данных по всем пунктам можно проверить с помощью компьютерных программ, например, используя функцию регрессионного анализа в Excel.

# Глава 5. МЕТОД РАЗНОСТИ РАЗНОСТЕЙ (DIFFERENCE-IN-DIFFERENCES)

Метод разности разностей (Difference-in-Differences, DiD) является одним из наиболее распространенных квази-экспериментальных методов в современной эконометрике, используемых для оценки причинно-следственных связей (causal effects) различных политических вмешательств и институциональных изменений. Согласно исследованиям, почти 25% всех эмпирических рабочих документов NBER и 17% эмпирических статей в пяти ведущих экономических журналах упоминают DiD.

История метода восходит к 1840-м годам, а первые известные применения связаны с работами Игнаца Земмельвайса и Джона Сноу по эпидемиологии. В экономических исследованиях метод получил широкое распространение во второй половине XX века и сегодня является стандартным инструментом в арсенале прикладного экономиста.

Разность разностей – это квази-экспериментальный метод, который измеряет причинный эффект некоторого неслучайного вмешательства (treatment) путем сравнения изменений результативного показателя во времени между группой, подвергшейся воздействию (treatment group), и группой, не подвергшейся воздействию (control group).

Базовый принцип DiD заключается в ответе на контрфактический вопрос: *что произошло бы с результативным показателем, если бы данное вмешательство не имело места?*. Если мы можем ответить на этот вопрос, то можем сравнить этот ответ с фактической ситуацией, где вмешательство было реализовано. Истинное воздействие лечения – это разница между фактическими значениями и ответом на контрфактический вопрос.

**Базовый дизайн: две группы и два периода**

**Структура данных**

Простейший дизайн DiD требует наличия:

**Двух временных периодов**: до вмешательства (pre-treatment) и после вмешательства (post-treatment)

**Двух групп**: экспериментальной (treatment group), которая подвергается воздействию во втором периоде, и контрольной (control

group), которая не подвергается воздействию на протяжении всего периода наблюдения [1; 2; 5].

Наблюдаемые средние значения результативного показателя могут быть представлены в виде табл. 5.1.

Таблица 5.1

Средние значения результативного показателя в дизайне 2×2

| Группа | До вмешательства (T=0) | После вмешательства (T=1) | Разность во времени | | | | |
|---|---|---|---|---|---|---|---|
| Экспериментальная (D=1) | E[$Y_0$ | D=1] | E[$Y_1$ | D=1] | $\Delta_1$ = E[$Y_1$ | D=1] – E[$Y_0$ | D=1] |
| Контрольная (D=0) | E[$Y_0$ | D=0] | E[$Y_1$ | D=0] | $\Delta_0$ = E[$Y_1$ | D=0] – E[$Y_0$ | D=0] |

**Расчет оценки DiD**

Оценка DiD вычисляется как двойная разность:

$DiD=(Ytreatment,post–Ytreatment,pre)–(Ycontrol,post–Ycontrol,pre)$

Эта двойная разность позволяет:

1. **Первая разность** (для каждой группы) – элиминирует постоянные во времени индивидуальные эффекты (time-invariant heterogeneity)

2. **Вторая разность** (между группами) – элиминирует общие временные тренды и макроэкономические шоки [1; 5; 8]

**Регрессионная реализация DiD**

**Базовая регрессионная модель**

На практике оценка DiD обычно реализуется через регрессионную модель. Для случая двух периодов и двух групп модель имеет вид [1; 5; 10]:

$Yit=\alpha+\beta\cdot Treati+\gamma\cdot Postt+\delta\cdot(Treati\times Postt)+\varepsilon it$ $Yit=\alpha+\beta\cdot Treati+\gamma\cdot Postt+\delta\cdot(Treati\times Postt)+\varepsilon it$

где

Y{it} – результативный показатель для единицы i в момент времени t;

Treat_i – дамми-переменная, равная 1 для единиц в экспериментальной группе;

Post_t – дамми-переменная, равная 1 для периода после вмешательства;

(Treat_i \times Post_t) – взаимодействие, равное 1 только для экспериментальной группы в пост-периоде;

delta – коэффициент DiD, представляющий собой оценку эффекта вмешательства;

varepsilon_{it} – случайная ошибка.

Коэффициент delta в точности соответствует оценке двойной разности, вычисленной по средним значениям.

**Модель с фиксированными эффектами**

Более общая спецификация, которая может быть распространена на случай многих периодов, использует фиксированные эффекты [1; 5]:

$$Yit=\alpha i+\lambda t+\delta \cdot Dit+\varepsilon it$$

где

- alpha_i – индивидуальные фиксированные эффекты (учитывают постоянные различия между единицами);

- ambda_t – временные фиксированные эффекты (учитывают общие макроэкономические шоки);

- D_{it} – индикатор воздействия (равен 1 для единиц i в периоды t, когда они подвергались лечению).

**Реализация DiD в R**

1. **Подготовка данных**

Данные в формате панели (id, time, outcome, treatment indicator)

Для многоэтапного DiD: переменная с годом первого лечения (0 для never-treated)

Проверка на пропущенные значения и сбалансированность панели

2. **Визуализация и предварительный анализ**

```r
# График средних по группам
# Проверка баланса ковариат
```

3. **Проверка параллельных трендов**

```r
# Event study с помощью did::aggte(type = "dynamic")
```

```
# Placebo-тесты
```

4. **Выбор метода оценки**

Два периода, две группы → lm() или feols()

Много периодов, одинаковое время лечения → feols() с взаимодействиями

Много периодов, staggered treatment → did::att_gt() или fixest::sunab()

5. **Робастность**

Различные спецификации (с контролем/без контроля)

Различные контрольные группы (never-treated vs not-yet-treated)

Кластеризация стандартных ошибок на уровне вмешательства

**Пакеты R для DiD-анализа**

| Пакет | Функции | Назначение |
|---|---|---|
| **did** | att_gt(), aggte(), ggdid() | Многоэтапный DiD (Callaway & Sant'Anna) |
| **fixest** | feols(), sunab() | Фиксированные эффекты, Sun & Abraham |
| **plm** | plm() | Панельные данные, базовые модели |
| **lfe** | felm() | Фиксированные эффекты, инструментальные переменные |
| **MatchIt** | matchit() | Propensity score matching для PSM-DiD |
| **weights** | weighted.did() | Взвешенный DiD |
| **spdep**, **splm** | spml() | Пространственный DiD |
| **DRDID** | drdid() | Doubly robust DiD |
| **did2s** | did2s() | Двухшаговый DiD (Gardner, 2022) |

## Установка и загрузка пакетов

```
# Установка основных пакетов
install.packages(c("did", "fixest", "plm", "lfe",
"MatchIt",
          "weights", "spdep", "splm", "DRDID",
"did2s"))

# Загрузка в сессии
library(did)
library(fixest)
library(plm)
library(lfe)
```

```
library(MatchIt)
library(weights)
library(spdep)
library(splm)
library(DRDID)
library(did2s)
library(tidyverse) # для работы с данными и визуализа-
ции
```

## Лабораторная работа № 1
## Анализ эффективности рекламных затрат

**Цель работы:** Изучить взаимосвязь между затратами на digital-рекламу и количеством привлеченных клиентов, построить прогнозную модель.

**Контекст:** Маркетинговый отдел интернет-магазина собрал данные о еженедельных расходах на контекстную рекламу (X, тыс. руб.) и количестве новых зарегистрированных пользователей (Y, чел.).

**Исходные данные:**

| Неделя | Затраты на рекламу (X) | Новые клиенты (Y) |
|---|---|---|
| 1 | 4.1 | 125 |
| 2 | 5.4 | 138 |
| 3 | 6.3 | 143 |
| 4 | 5.4 | 143 |
| 5 | 4.8 | 145 |
| 6 | 4.6 | 130 |
| 7 | 6.2 | 140 |
| 8 | 6.1 | 151 |
| 9 | 6.4 | 158 |
| 10 | 7.1 | 165 |

**Задания:**

1. Используя программные средства (Excel, R), постройте точечную диаграмму рассеивания. Существует ли визуальная линейная взаимосвязь между затратами на рекламу и количеством новых клиентов?

2. Рассчитайте коэффициент корреляции Пирсона. Интерпретируйте его значение: насколько сильна связь?

3. Постройте модель парной линейной регрессии (уравнение прогноза) Y = b0 + b1*X методом наименьших квадратов. Запишите полученное уравнение.

4. Проверьте значимость коэффициента регрессии b1 (углового коэффициента) на 5%-ном уровне значимости. Сделайте вывод о наличии линейной зависимости в генеральной совокупности.

5. Рассчитайте коэффициент детерминации $R^2$. Какой процент вариации количества новых клиентов объясняется вариацией затрат на рекламу?

6. Составьте прогноз количества новых клиентов при уровне затрат на рекламу 8.0 тыс. руб.

7. **Для реализации в Excel:**

Выполните задания 2-5 с помощью пакета анализа «Регрессия» (надстройка «Анализ данных»).

Постройте на графике линию тренда, отобразите уравнение регрессии и величину достоверности аппроксимации ($R^2$).

8. **Для реализации в R:**

Напишите скрипт, выполняющий задания 2-6. Скрипт должен выводить в консоль: уравнение регрессии, коэффициент детерминации, p-value для коэффициента, а также прогнозное значение.

Постройте диаграмму рассеивания с добавленной линией регрессии (используя ggplot2 или базовые графики plot() и abline()).

9. Напишите краткий аналитический отчет (вывод) по результатам работы. Стоит ли маркетологам увеличивать бюджет на рекламу?

## Лабораторная работа № 2
## Производительность труда и время обслуживания

**Цель работы:** Оценить зависимость объема выпуска продукции от времени, затраченного на обслуживание оборудования, и построить интервальный прогноз.

**Контекст:** На производственном участке зафиксированы показатели времени планово-предупредительного обслуживания станков (X, часы) и объем произведенной продукции за смену (Y, усл. ед.).

**Исходные данные:**

| Наблюдение | Время обслуживания (X) | Объем изделий (Y) |
|---|---|---|
| 1 | 3.6 | 30.6 |
| 2 | 4.1 | 30.5 |
| 3 | 0.8 | 2.4 |
| 4 | 5.7 | 42.2 |
| 5 | 3.4 | 21.8 |
| 6 | 1.8 | 6.2 |
| 7 | 4.3 | 40.1 |
| 8 | 0.2 | 2.0 |
| 9 | 2.6 | 15.5 |
| 10 | 1.3 | 6.5 |

**Задания:**

1. Постройте диаграмму рассеивания. Опишите характер зависимости.

2. Рассчитайте выборочный коэффициент корреляции. Проверьте его значимость на 5% уровне.

3. Определите уравнение регрессии Y по X методом наименьших квадратов.

4. **Для реализации в Excel:**

Используя функцию ЛИНЕЙН, получите коэффициенты регрессии и их стандартные ошибки.

Постройте доверительный интервал для среднего значения Y при X = 3.0.

5. **Для реализации в R:**

Постройте модель линейной регрессии с помощью lm().

Используя функцию predict(), получите точечный прогноз и 99%-ный доверительный интервал для индивидуального значения Y при X = 3.0. Интерпретируйте полученный интервал.

6. Проверьте значимость углового коэффициента на 5%-ном уровне.

7. Составьте отчет, включив в него сравнение точечной и интервальной оценок. Почему интервальная оценка полезнее для планирования?

## Лабораторная работа № 3
## Анализ бюджета автопарка

**Цель работы:** Выявить зависимость расходов на содержание автомобиля от его возраста и использовать модель для бюджетного планирования.

**Контекст:** Транспортный отдел логистической компании анализирует ежегодные затраты на ремонт и обслуживание автомобилей (Y, долл.) в зависимости от их возраста (X, полных лет с начала эксплуатации).

**Исходные данные:**

| Автомобиль | Расходы на содержание (Y) | Возраст (X) |
|---|---|---|
| A | 859 | 8 |
| B | 682 | 5 |
| C | 471 | 3 |
| D | 708 | 9 |
| E | 1094 | 11 |
| F | 224 | 2 |
| G | 320 | 1 |
| H | 651 | 8 |
| I | 1049 | 12 |

**Задания:**

1. Постройте диаграмму рассеивания. Подтверждается ли гипотеза о росте расходов с увеличением возраста?
2. Найдите уравнение парной регрессии, описывающее зависимость расходов от возраста.
3. Проверьте значимость модели в целом и ее коэффициентов.
4. **Для реализации в Excel:**

Используя надстройку «Анализ данных» (Регрессия), получите таблицу с дисперсионным анализом (ANOVA). Интерпретируйте F-статистику.

На основе полученного уравнения спрогнозируйте расходы для автомобиля возрастом 5 лет.

5. **Для реализации в R:**

После построения модели, извлеките остатки (residuals) и постройте график «Остатки от предсказанных значений». Проверьте, выполняется ли условие гомоскедастичности.

Сделайте прогноз для автомобиля возрастом 5 лет.

6. Напишите вывод: существует ли положительная взаимосвязь и как отдел может использовать эту модель при формировании годового бюджета?

## Лабораторная работа № 4
## Прогнозирование продаж в ритейле

**Цель работы:** Построить модель для прогнозирования недельных продаж книжной продукции на основе торговых площадей.

**Контекст:** Сеть книжных магазинов исследует влияние длины стеллажей (полочного пространства) на выручку. Для анализа собраны данные по 11 торговым точкам за неделю (X – суммарная длина стеллажей, ед. изм.; Y – количество проданных книг, шт.).

**Исходные данные:**

| Магазин | Длина стеллажей (X) | Продажи книг (Y) |
|---|---|---|
| 1 | 6.8 | 275 |
| 2 | 3.3 | 142 |
| 3 | 4.1 | 168 |
| 4 | 4.2 | 197 |
| 5 | 4.8 | 215 |
| 6 | 3.9 | 188 |
| 7 | 4.9 | 241 |
| 8 | 7.7 | 295 |
| 9 | 3.1 | 125 |
| 10 | 5.9 | 266 |
| 11 | 5.0 | 200 |

**Задания:**

1. Вычислите и интерпретируйте коэффициент корреляции.

2. Определите уравнение регрессии. Проверьте значимость коэффициента наклона (b1) при $\alpha=0.05$.

3. Рассчитайте коэффициент детерминации. Хорошо ли модель описывает данные?

4. **Для реализации в Excel:**

Используя инструмент «Регрессия», получите стандартную ошибку оценки (стандартное отклонение остатков). Что она означает?

Спрогнозируйте продажи для магазина с длиной стеллажей, равной 4.0 ед. изм.

5. **Для реализации в R:**

Постройте модель. Создайте новый data.frame с точкой прогноза (X=4.0). Используйте predict, чтобы получить предсказанное значение

и 95% доверительный интервал для среднего значения Y. Визуализируйте модель, добавив на график точку прогноза.

6. Составьте отчет. Ответьте на вопрос: достаточно ли одного фактора (длина стеллажей) для точного прогнозирования продаж?

## Лабораторная работа № 5
## Эффективность почтовой рассылки

**Цель работы:** Оценить эффективность маркетинговой кампании (директ-мейл) и построить интервальный прогноз отклика.

**Контекст:** Компания, торгующая по каталогам, хочет спрогнозировать количество заказов в зависимости от масштаба рассылки. Данные по 12 городам: X – количество разосланных каталогов (тыс. шт.), Y – количество полученных заказов (тыс. шт.).

**Исходные данные:**

| Город | Заказы (Y) | Каталоги (X) |
|---|---|---|
| A | 24 | 6 |
| B | 16 | 2 |
| C | 23 | 5 |
| D | 15 | 1 |
| E | 32 | 10 |
| F | 25 | 7 |
| G | 18 | 15 |
| H | 18 | 3 |
| I | 35 | 11 |
| J | 34 | 13 |
| K | 15 | 2 |
| L | 32 | 12 |

**Задания:**

1. Постройте диаграмму рассеивания. Обратите внимание на город G. Можно ли его назвать выбросом (нетипичным наблюдением)?

2. Определите уравнение регрессии Y по X.

3. Проверьте значимость линейной взаимосвязи между переменными на уровне значимости 0.05 (используйте F-тест или t-тест для коэффициента).

4. **Для реализации в Excel:**

Вычислите стандартную ошибку оценки (Se). Постройте 90%-ный доверительный интервал для среднего значения Y при X = 10 (тыс. каталогов).

5. **Для реализации в R:**

Постройте модель. Выведите сводку (summary). Постройте график остатков. Есть ли на графике остатков подтверждение, что город G – выброс?

Постройте 90%-ный интервал прогноза (prediction interval) для количества заказов при X = 10 (тыс. каталогов). В чем отличие доверительного интервала от интервала прогноза? Интерпретируйте результат.

6. Напишите отчет для отдела маркетинга: насколько эффективна рассылка (сколько заказов приносит 1000 дополнительных каталогов)?

## Лабораторная работа № 6
## Макроэкономический анализ

**Цель работы:** Исследовать зависимость между размером банковских вкладов и ключевой процентной ставкой.

**Контекст:** Аналитик банка изучает поведение вкладчиков. Собраны годовые данные за 10 лет: X – средняя процентная ставка по депозитам (%), Y – общий объем привлеченных вкладов (тыс. у.е.).

**Исходные данные:**

| Год | Ставка по вкладам (X) | Объем вкладов (Y) |
|---|---|---|
| 1 | 4.8 | 1060 |
| 2 | 5.1 | 940 |
| 3 | 5.9 | 920 |
| 4 | 5.1 | 1110 |
| 5 | 4.8 | 1590 |
| 6 | 3.8 | 2050 |
| 7 | 3.7 | 2070 |
| 8 | 4.5 | 2030 |
| 9 | 4.9 | 1780 |
| 10 | 6.2 | 1420 |

**Задания:**

1. Постройте диаграмму рассеивания. Какой характер зависимости наблюдается (прямая или обратная)?

2. Рассчитайте коэффициент корреляции. Сделайте вывод о тесноте связи.

3. Постройте уравнение линейной регрессии, где Y (вклады) – зависимая переменная.

4. **Для реализации в Excel:**

Получите уравнение регрессии. Рассчитайте остатки (разницу между фактическими и предсказанными значениями).

Спрогнозируйте объем вкладов, если процентная ставка упадет до 4.0%.

5. **Для реализации в R:**

Постройте модель. Проверьте, значима ли модель? Является ли ставка значимым предиктором?

Рассчитайте общую вариацию (TSS) и необъясненную вариацию (RSS). Убедитесь, что TSS = RSS + ESS.

6. Составьте отчет. Можно ли использовать данную модель для прогноза при ставке 4%? Насколько точен этот прогноз (оцените по $R^2$)?

## Лабораторная работа № 7
## Рынок недвижимости и макроэкономика

**Цель работы:** Оценить влияние банковской учетной ставки на деловую активность в строительной отрасли.

**Контекст:** Аналитик строительного холдинга исследует зависимость между количеством выданных разрешений на строительство (Y, шт.) и размером ключевой ставки Центробанка (X, %).

**Исходные данные:**

| Месяц | Разрешения (Y) | Ключевая ставка (X) |
|---|---|---|
| 1 | 786 | 10.2 |
| 2 | 494 | 12.6 |
| 3 | 289 | 13.5 |
| 4 | 892 | 9.7 |
| 5 | 343 | 10.8 |
| 6 | 888 | 9.5 |
| 7 | 509 | 10.9 |
| 8 | 987 | 9.2 |
| 9 | 187 | 14.2 |

**Задания:**

1. Постройте диаграмму рассеивания. Подтверждается ли предположение о том, что высокая ставка «охлаждает» рынок строительства?

2. Рассчитайте уравнение регрессии. Дайте интерпретацию коэффициенту b1: на сколько единиц в среднем изменяется количество разрешений при изменении ставки на 1%?

3. Вычислите коэффициент детерминации $R^2$. Какая доля вариации количества разрешений объясняется изменением ставки?

4. **Для реализации в Excel:**

Используя инструмент «Регрессия», проверьте значимость коэффициента b1 на уровне 5%. Постройте диаграмму рассеивания с линией тренда.

5. **Для реализации в R:**

Постройте модель. Создайте вектор новых значений ставки (например, от 8 до 15) и постройте предсказанные линию регрессии с доверительной областью (с помощью `geom_smooth(method = "lm")` в ggplot2 или функции `predict`).

6. Напишите вывод: как изменение денежно-кредитной политики влияет на планы строителей?

## Лабораторная работа № 8
## Управление качеством на производстве

**Цель работы:** Проанализировать зависимость количества дефектов от объема производственной партии для оптимизации контроля качества.

**Контекст:** На заводе по производству комплектующих фиксируется количество бракованных изделий (Y, шт.) в зависимости от размера произведенной партии (X, шт.).

**Исходные данные:**

| Партия | Размер партии (X) | Брак (Y) |
|---|---|---|
| 1 | 25 | 4 |
| 2 | 50 | 8 |
| 3 | 75 | 6 |
| 4 | 100 | 16 |
| 5 | 125 | 22 |
| 6 | 150 | 27 |
| 7 | 175 | 36 |
| 8 | 200 | 49 |
| 9 | 225 | 53 |
| 10 | 250 | 70 |
| 11 | 275 | 82 |
| 12 | 300 | 95 |
| 13 | 325 | 109 |

**Задания:**

1. Постройте диаграмму рассеивания. С увеличением размера партии растет ли количество брака? Выглядит ли зависимость линейной?

2. Рассчитайте выборочный коэффициент корреляции. Проверьте его значимость.

3. Постройте уравнение регрессии.

4. **Для реализации в Excel:**

Спрогнозируйте количество бракованных изделий для партии размером 300 шт. Сравните прогноз с фактическим значением из таблицы (партия №12).

Постройте график остатков. Есть ли в остатках какой-либо паттерн (систематическое отклонение)?

5. **Для реализации в R:**

Постройте модель. Используйте функцию predict, чтобы получить прогноз и 95% доверительный интервал для партии размером 300 шт.

Рассчитайте общую (TSS) и необъясненную (RSS) вариации.

6. Составьте аналитическую записку технологу: как можно использовать эту модель для планирования объема выборки при контроле качества?

## Лабораторная работа № 9

## Оценка стоимости недвижимости (Кадастровая и Рыночная)

**Цель работы:** Построить модель для оценки рыночной стоимости объектов недвижимости на основе кадастровой (инвентаризационной) оценки и оценить ее прогностическую способность.

**Контекст:** Риелторское агентство "Городской стандарт" провело анализ данных по 30 сделкам купли-продажи домов в одном районе города. Цель исследования – понять, насколько кадастровая стоимость (X, тыс. у.е.) соответствует реальной рыночной цене (Y, тыс. у.е.), и возможно ли использовать кадастровую оценку для быстрого определения справедливой цены продажи.

**Исходные данные:**

| Дом | Кадастровая стоимость (X) | Рыночная стоимость (Y) | Дом | Кадастровая стоимость (X) | Рыночная стоимость (Y) |
|---|---|---|---|---|---|
| 1 | 68,2 | 87,4 | 16 | 74,0 | 88,4 |
| 2 | 74,6 | 88,0 | 17 | 72,8 | 93,6 |
| 3 | 64,6 | 87,2 | 18 | 80,4 | 92,8 |
| 4 | 80,2 | 94,0 | 19 | 74,2 | 90,6 |
| 5 | 76,0 | 94,2 | 20 | 80,0 | 91,6 |
| 6 | 78,0 | 93,6 | 21 | 81,6 | 92,8 |
| 7 | 76,0 | 88,4 | 22 | 75,6 | 89,0 |
| 8 | 77,0 | 92,2 | 23 | 79,4 | 91,8 |
| 9 | 75,2 | 90,4 | 24 | 82,2 | 98,4 |
| 10 | 72,4 | 90,4 | 25 | 67,0 | 89,8 |
| 11 | 80,0 | 93,6 | 26 | 72,0 | 97,2 |
| 12 | 76,4 | 91,4 | 27 | 73,6 | 95,2 |
| 13 | 70,2 | 89,6 | 28 | 71,4 | 88,8 |
| 14 | 75,8 | 91,8 | 29 | 81,0 | 97,4 |
| 15 | 79,2 | 94,8 | 30 | 80,6 | 95,4 |

**Задания:**

**Часть 1. Предварительный анализ данных**

1. Постройте диаграмму рассеивания (точечный график) для переменных X и Y. Визуально оцените:

Наличие и направление связи между кадастровой и рыночной стоимостью.

Насколько тесной представляется эта связь?

Имеются ли явные выбросы (нетипичные наблюдения)?

2. Рассчитайте коэффициент корреляции Пирсона. Интерпретируйте полученное значение. Проверьте гипотезу о том, что коэффициент корреляции в генеральной совокупности значимо отличается от нуля (при уровне значимости $\alpha = 0,05$). Какой вывод можно сделать о наличии линейной взаимосвязи?

**Часть 2. Построение регрессионной модели**

3. Используя метод наименьших квадратов, постройте уравнение парной линейной регрессии, где рыночная стоимость (Y) выступает в роли зависимой переменной, а кадастровая стоимость (X) – независимой. Запишите полученное уравнение в виде: $Y = b_0 + b_1 \cdot X$.

4. Дайте экономическую интерпретацию коэффициенту регрессии $b_1$. Что означает его величина?

5. Рассчитайте коэффициент детерминации $R^2$. Какая доля вариации рыночной стоимости домов объясняется вариацией их кадастровой стоимости? Достаточно ли точно модель описывает данные?

**Часть 3. Реализация в Excel**

6. Используя надстройку «Пакет анализа» → инструмент «Регрессия», выполните расчеты. В полученной таблице найдите и выпишите:

Коэффициенты уравнения регрессии.

Коэффициент детерминации $R^2$.

Стандартную ошибку регрессии (Se) – это стандартное отклонение остатков.

Результаты дисперсионного анализа (ANOVA), обратите внимание на F-статистику и ее значимость.

7. На основе построенной модели спрогнозируйте рыночную стоимость дома, кадастровая стоимость которого составляет **90,5 тыс. у.е.** (предположим, что такой дом выставлен на продажу).

8. Проанализируйте остатки регрессии (разности между фактическими и предсказанными значениями). Можно ли заметить в них какую-либо закономерность?

**Часть 4. Реализация в R**

9. Напишите скрипт на языке R, который выполняет следующие задачи:

Создает векторы или датафрейм с исходными данными.

Строит диаграмму рассеивания с помощью базовой функции `plot()` или библиотеки `ggplot2`.

Строит модель линейной регрессии с помощью функции `lm()`. Записывает результат в объект, например, `model`.

Выводит в консоль подробную сводку по модели с помощью `summary(model)`.

Извлекает из сводки значение F-статистики и p-value для проверки значимости регрессии в целом.

10. Используя функцию `predict()`, выполните следующие действия:

Создайте новый датафрейм с точкой прогноза: `new_data <- data.frame(X = 90.5)`.

Получите точечный прогноз рыночной стоимости для X = 90,5.

Получите **95% доверительный интервал** для *среднего* значения рыночной стоимости всех домов с кадастровой стоимостью 90,5 тыс. у.е. Интерпретируйте полученный интервал.

11. **(Важный аналитический вопрос)** Оцените надежность прогноза для X = 90,5.

Сравните значение X = 90,5 с диапазоном значений X в исходной выборке (найдите минимум и максимум кадастровой стоимости).

Постройте график, демонстрирующий опасность экстраполяции:

Нанесите исходные точки (X, Y).

Добавьте линию регрессии.

Добавьте доверительные интервалы для линии регрессии (можно использовать geom_smooth(method = "lm", se = TRUE) в ggplot2 или рассчитать их вручную).

Отметьте на графике точку прогноза X = 90,5.

Визуально оцените, попадает ли точка прогноза в область данных, на основе которых строилась модель. Сделайте вывод о том, допустимо ли использовать построенную модель для прогноза при данном значении X.

**Часть 5. Выводы и заключение**

12. На основе всех проведенных расчетов и построенных графиков напишите развернутый вывод, который должен содержать ответы на следующие вопросы:

Существует ли статистически значимая линейная зависимость между кадастровой и рыночной стоимостью на рассматриваемом рынке недвижимости?

Насколько сильна эта зависимость? Можно ли считать кадастровую стоимость хорошим предиктором (предсказывающим фактором) для рыночной цены?

Если дом имеет кадастровую стоимость 90,5 тыс. у.е., какой можно дать прогноз по его рыночной цене? Будет ли этот прогноз надежным с точки зрения статистики? Почему?

Какие рекомендации вы могли бы дать риелторскому агентству по использованию кадастровой стоимости в своей работе?

## Лабораторная работа № 10
## Оценка эффективности государственной программы поддержки малого бизнеса методом разности разностей (Difference-in-Differences)

**Цель работы**

Освоить применение метода разности разностей (Difference-in-Differences, DiD) для оценки причинно-следственных эффектов экономических политик и программ. Научиться проверять предположения метода, интерпретировать результаты и формулировать практические рекомендации.

**Задания**

В 2022 году правительство региона запустило программу поддержки малого бизнеса в отдельных городах. Программа включала:

Налоговые льготы для вновь зарегистрированных малых предприятий

Субсидирование части затрат на аренду помещений

Бесплатные консультационные услуги по бизнес-планированию

Программа стартовала в 2022 году в 6 городах (экспериментальная группа). Другие 6 городов не получали поддержки (контрольная группа). Вам предоставлены данные по количеству вновь зарегистрированных малых предприятий (на 1000 жителей) за 4 года: 2 года до программы (2020-2021) и 2 года после (2022-2023).

**Исходные данные:**

| Город | Группа | 2020 | 2021 | 2022 | 2023 |
|---|---|---|---|---|---|
| A | Контроль | 12.5 | 13.0 | 13.2 | 13.8 |
| B | Контроль | 8.3 | 8.7 | 9.0 | 9.4 |
| C | Контроль | 15.2 | 15.8 | 16.1 | 16.5 |
| D | Контроль | 10.1 | 10.5 | 10.8 | 11.2 |
| E | Контроль | 14.3 | 14.9 | 15.2 | 15.7 |
| F | Контроль | 9.7 | 10.2 | 10.5 | 11.0 |
| G | Эксперимент | 11.8 | 12.3 | 14.5 | 16.8 |
| H | Эксперимент | 13.2 | 13.9 | 15.8 | 17.2 |
| I | Эксперимент | 10.5 | 11.1 | 13.2 | 14.9 |
| J | Эксперимент | 16.3 | 17.0 | 18.9 | 21.3 |
| K | Эксперимент | 12.7 | 13.4 | 15.1 | 16.5 |
| L | Эксперимент | 14.8 | 15.4 | 17.3 | 19.6 |

**Метод разности разностей (Difference-in-Differences, DiD)** – это квази-экспериментальный метод оценки причинно-следственных эффектов, который сравнивает изменения результативного показателя во времени между группой, подвергшейся воздействию (treatment group), и группой, не подвергшейся воздействию (control group).

Базовая формула:

$\delta DiD=(\bar{Y}treatment,post-\bar{Y}treatment,pre)-(\bar{Y}control,post-\bar{Y}control,pre)$

**Ключевое предположение**: параллельные тренды (parallel trends) – в отсутствие программы экспериментальная и контрольная группы следовали бы параллельным траекториям.

**Часть 1. Подготовка данных и описательный анализ**

**Задание 1.1.** Преобразуйте данные из "широкого" формата (wide format) в "длинный" формат (long format), необходимый для панельного анализа. В длинном формате должны быть колонки: city, year, registrations, treatment (1 для экспериментальной группы), post (1 для 2022-2023 годов).

*Подсказка:* используйте функции pivot_longer() в R или Power Query в Excel.

**Задание 1.2.** Рассчитайте средние значения регистраций для каждой группы (контрольная/экспериментальная) в каждый год. Заполните таблицу:

| Год | Контрольная группа | Экспериментальная группа |
|---|---|---|
| 2020 | | |
| 2021 | | |
| 2022 | | |
| 2023 | | |

**Задание 1.3.** Постройте график динамики средних значений для обеих групп на одном рисунке. Визуально оцените:

- Выполняется ли предположение о параллельных трендах в пред-программный период (2020-2021)?
- Наблюдается ли расхождение трендов после 2022 года?

**Часть 2. Расчет DiD вручную**

**Задание 2.1.** Используя средние значения, рассчитайте разность разностей "вручную" (как в формуле выше). Заполните таблицу:

| Показатель | Контрольная группа | Экспериментальная группа |
|---|---|---|
| Среднее до программы (2020-2021) | | |
| Среднее после программы (2022-2023) | | |
| Изменение (после – до) | $\Delta_0$ = | $\Delta_1$ = |

**Задание 2.2.** Вычислите DiD-оценку эффекта программы: $\delta\hat{}=\Delta 1-\Delta 0=$_____ $\delta\hat{}=\Delta 1-\Delta 0=$_____

**Задание 2.3.** Интерпретируйте полученное значение. Что оно означает для бизнеса? Сколько дополнительных предприятий на 1000 жителей создала программа?

**Часть 3. Реализация в Excel**

**Задание 3.1.** Создайте в Excel переменные для регрессионной модели:

Treatment (1 для экспериментальных городов)

Post (1 для 2022-2023)

Treatment × Post (взаимодействие)

**Задание 3.2.** Постройте регрессионную модель:

$\text{Registrations} = \alpha + \beta_1 \cdot \text{Treatment} + \beta_2 \cdot \text{Post} + \beta_3 \cdot (\text{Treatment} \times \text{Post}) + \varepsilon$

Используйте надстройку "Анализ данных" → "Регрессия".

**Задание 3.3.** Заполните таблицу результатов:

| Коэффициент | Значение | p-value | Стандартная ошибка |
|---|---|---|---|
| Intercept ($\alpha$) | | | |
| Treatment ($\beta_1$) | | | |
| Post ($\beta_2$) | | | |
| Treatment×Post ($\beta_3$) | | | |

**Задание 3.4.** На уровне значимости 0.05 проверьте гипотезу об эффективности программы ($H_0$: $\beta_3 = 0$ против $H_1$: $\beta_3 \neq 0$). Является ли эффект статистически значимым?

**Задание 3.5.** Сравните значение $\beta_3$ из регрессии с вашим ручным расчетом из Задания 2.2. Объясните, почему они совпадают или различаются.

**Часть 4. Реализация в R**

**Задание 4.1.** Загрузите данные в R и выполните необходимые преобразования. Создайте длинный формат данных и добавьте переменные treatment, post и did (взаимодействие).

**Задание 4.2.** Постройте график средних значений по группам с доверительными интервалами. Добавьте вертикальную линию в год начала программы (2022).

**Вопрос:** Подтверждает ли график предположение о параллельных трендах?

**Задание 4.3.** Оцените регрессионную модель DiD тремя способами и сравните результаты

**Задание 5.1.** Проанализируйте остатки модели. Постройте график остатков и проверьте их на нормальность.

**Задание 5.3.** Проверьте чувствительность результатов к исключению отдельных городов. По очереди исключите каждый город из экспериментальной группы и переоцените модель.

**Задание 6.1.** Используя построенную модель, спрогнозируйте количество регистраций в экспериментальных городах в 2024 году при условии, что программа продолжается. Сравните с прогнозом при отсутствии программы.

**Задание 6.2.** Рассчитайте экономическую эффективность программы, если известно, что:

Каждое новое малое предприятие создает в среднем 2.5 рабочих места

Средняя зарплата в регионе – 45 000 руб./мес.

Бюджет программы составил 15 млн руб. на город в 2022-2023 годах

*Примечание для студентов: При выполнении работы в R убедитесь, что десятичные дроби записываются через точку (например, 68.2), а не через запятую. Для построения графиков с доверительными интервалами в* ggplot2 *используйте функцию* geom_smooth(method = "lm").

# Глава 6. МНОГОМЕРНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

В простой линейной регрессии рассматривалась взаимосвязь между независимой и зависимой переменными. Связь между двумя переменными часто позволяет точно предсказать значение зависимой переменной, если известно значение независимой переменной. Однако для точного прогнозирования зависимой переменной обычно требуется знать значения более чем одной независимой переменной. Регрессионные модели с несколькими независимыми переменными называются *моделями многомерной регрессии.*

Выбор уравнения многомерной регрессии с наиболее подходящими для прогноза переменными проводится следующим образом:

1. Определение набора возможных независимых переменных.

2. Исключение переменных, не имеющих существенного отношения к решению поставленной задачи (если переменная характеризуется значительными ошибками измерения, дублирует другие независимые переменные (мультиколлинеарность), точные данные по ней недоступны);

*3.* Выбор окончательного вида уравнения с «наилучшими» независимыми переменными, при этом решается задача обеспечения наилучшего прогноза с наименьшими затратами.

Области применения многомерного регрессионного анализа различны:

- отражение взаимосвязи уровня зарплаты работников с географическим расположением компаний, уровнем безработицы в регионе, темпами роста промышленности, членством в союзах, отраслью промышленности или уровнем зарплаты в конкурирующих фирмах;

- анализ изменения цены на акции исходя из получаемых дивидендов, доходов от каждой акции, дробления акций, ожидаемой процентной ставки, объемов сбережений и уровня инфляции;

- исследование влияния на изменение мнения покупателей размеров рекламного бюджета, выбора средств информации, повторения информации, частоты рекламных акций или выбора рекламирующей персоны;

- анализ зависимости объема продаж от расходов на рекламу, уровня цен, маркетинговых расходов конкурентов и разовых заработков покупателей, а также от большого числа других переменных.

Таким образом, целью лабораторной работы является приобретение практических навыков построения уравнения многомерной регрессии предлагаемой социально-экономической ситуации с помощью инструмента анализа данных Excel и R.

1. Методические положения построения модели многомерной регрессии на основе практического примера.

В табл. 6.1 представлены исходные данные для проведения расчетов, где, $Y$ – выработка продукции, $x_1$ - коэффициент обновления основных фондов, $x_2$ – доля рабочих высокой квалификации.

Необходимо ответить на следующие вопросы:

1. Оценить показатели вариации каждого признака и сделать вывод о возможностях применения МНК для их изучения.

2. Проанализировать линейные коэффициенты парной и частной корреляции.

3. Написать уравнение множественной регрессии, оценить значимость его параметров, пояснить их экономический смысл.

Таблица 6.1

Исходные данные для многомерной регрессии

| Номер предприятия | y | x1 | x2 | № предприятия | y | x1 | x2 |
|---|---|---|---|---|---|---|---|
| 1 | 7 | 3,9 | 10 | 11 | 9 | 6 | 21 |
| 2 | 7 | 3,9 | 14 | 12 | 11 | 6,4 | 22 |
| 3 | 7 | 3,7 | 15 | 13 | 9 | 6,8 | 22 |
| 4 | 7 | 4 | 16 | 14 | 11 | 7,2 | 25 |
| 5 | 7 | 3,8 | 17 | 15 | 12 | 8 | 28 |
| 6 | 7 | 4,8 | 19 | 16 | 12 | 8,2 | 29 |
| 7 | 8 | 5,4 | 19 | 17 | 12 | 8,1 | 30 |
| 8 | 8 | 4,4 | 20 | 18 | 12 | 8,5 | 31 |
| 9 | 8 | 5,3 | 20 | 19 | 14 | 9,6 | 32 |
| 10 | 10 | 6,8 | 20 | 20 | 14 | 9 | 36 |

4. С помощью F-критерия Фишера оценить статистическую надежность уравнения регрессии и $R^2_{yx_1x_2}$. Сравнить значения скорректированного и нескорректированного коэффициентов множественной детерминации.

5. С помощью частных F-критериев Фишера оценить целесообразность включения в уравнение множественной регрессии фактора $x_1$ после $x_2$ и фактора $x_2$ после $x_1$.

6. Рассчитать средние частные коэффициенты эластичности и дать на их основе сравнительную оценку силы влияния факторов на результат.

***Решение с помощью Excel***

**1.** Для оценки показателя вариации каждого признака необходимо составить сводную таблицу основных статистических характеристик для одного или нескольких массивов данных, которую можно получить с помощью инструмента анализа данных, **Описательная статистика.** Для этого следует выполнить следующие шаги:

1) введите исходные данные или откройте существующий файл, содержащий анализируемые данные;

в главном меню выберите последовательно пункты **Сервис/Анализ данных/Описательная статистика**, после чего щелкните по кнопке **ОК**;

2) заполните диалоговое окно ввода данных и параметров вывода (рис. 6.1).

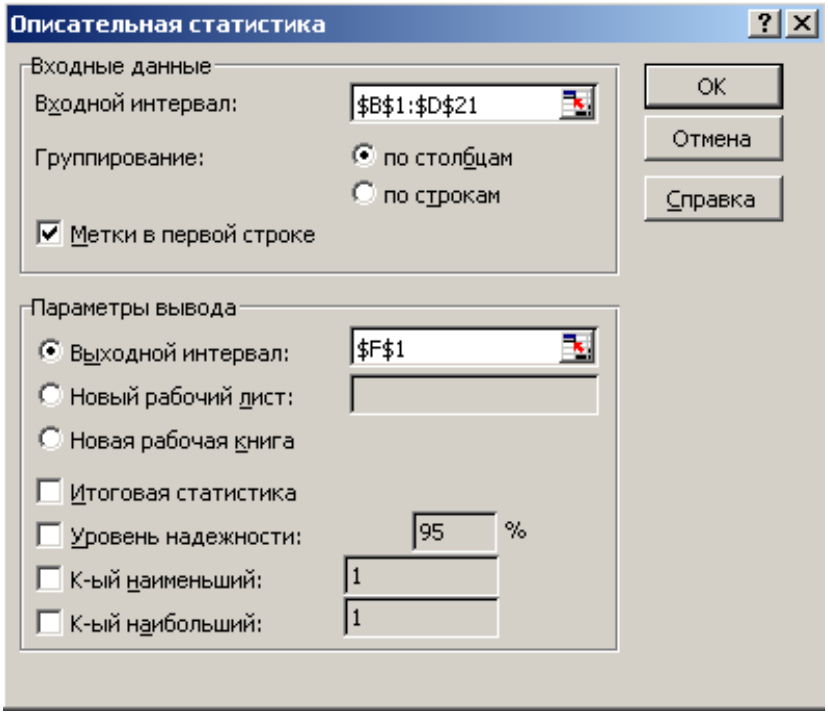



Рис. 6.1. Диалоговое окно ввода параметров инструмента *Описательная статистика*

*Входной интервал* – диапазон, содержащий анализируемые данные, это может быть одна или несколько строк (столбцов).

*Группирование* – по столбцам или строкам – необходимо указать дополнительно.

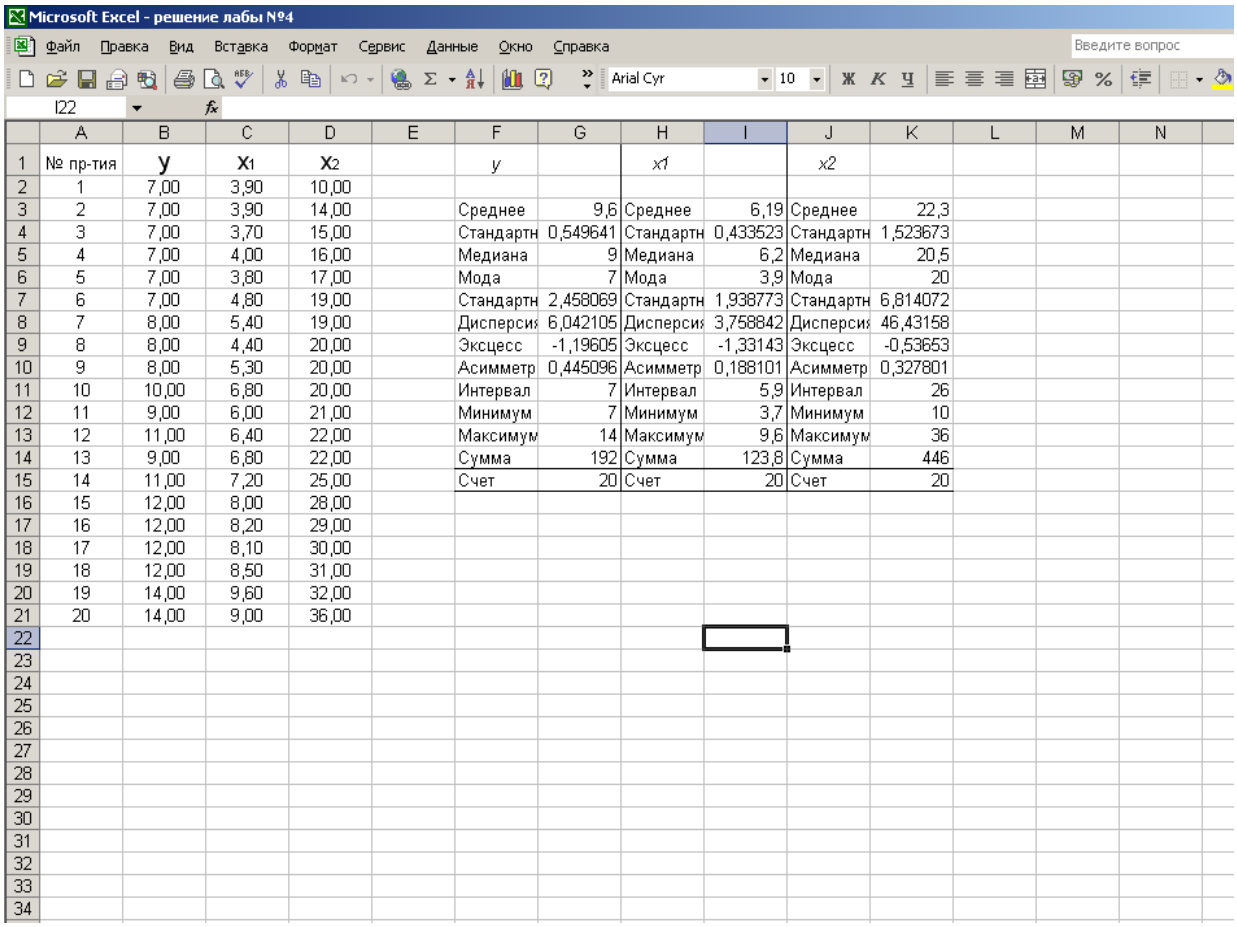
*Метки* – флажок, который указывает, содержит ли первая строка названия столбцов или нет.

*Выходной интервал* – достаточно указать верхнюю левую ячейку будущего диапазона.

*Новый рабочий лист* – можно задать произвольное имя нового листа.

Если необходимо получить дополнительную информацию по *итоговой статистике, уровню надежности, k-го наибольшего и наименьшего значений,* установите соответствующие флажки в диалоговом окне. Щелкните по кнопке **OK**.

Результаты вычисления соответствующих показателей для каждого признака представлены на рис. 6.2.




| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | № пр-тия | y | x1 | x2 | | y | | x1 | | x2 | | | | |
| 2 | 1 | 7,00 | 3,90 | 10,00 | | | | | | | | | | |
| 3 | 2 | 7,00 | 3,90 | 14,00 | | Среднее | 9,6 | Среднее | 6,19 | Среднее | 22,3 | | | |
| 4 | 3 | 7,00 | 3,70 | 15,00 | | Стандартн | 0,549641 | Стандартн | 0,433523 | Стандартн | 1,523673 | | | |
| 5 | 4 | 7,00 | 4,00 | 16,00 | | Медиана | 9 | Медиана | 6,2 | Медиана | 20,5 | | | |
| 6 | 5 | 7,00 | 3,80 | 17,00 | | Мода | 7 | Мода | 3,9 | Мода | 20 | | | |
| 7 | 6 | 7,00 | 4,80 | 19,00 | | Стандартн | 2,458069 | Стандартн | 1,938773 | Стандартн | 6,814072 | | | |
| 8 | 7 | 8,00 | 5,40 | 19,00 | | Дисперсия | 6,042105 | Дисперсия | 3,758842 | Дисперсия | 46,43158 | | | |
| 9 | 8 | 8,00 | 4,40 | 20,00 | | Эксцесс | -1,19605 | Эксцесс | -1,33143 | Эксцесс | -0,53653 | | | |
| 10 | 9 | 8,00 | 5,30 | 20,00 | | Асимметр | 0,445096 | Асимметр | 0,188101 | Асимметр | 0,327801 | | | |
| 11 | 10 | 10,00 | 6,80 | 20,00 | | Интервал | 7 | Интервал | 5,9 | Интервал | 26 | | | |
| 12 | 11 | 9,00 | 6,00 | 21,00 | | Минимум | 7 | Минимум | 3,7 | Минимум | 10 | | | |
| 13 | 12 | 11,00 | 6,40 | 22,00 | | Максимум | 14 | Максимум | 9,6 | Максимум | 36 | | | |
| 14 | 13 | 9,00 | 6,80 | 22,00 | | Сумма | 192 | Сумма | 123,8 | Сумма | 446 | | | |
| 15 | 14 | 11,00 | 7,20 | 25,00 | | Счет | 20 | Счет | 20 | Счет | 20 | | | |
| 16 | 15 | 12,00 | 8,00 | 28,00 | | | | | | | | | | |
| 17 | 16 | 12,00 | 8,20 | 29,00 | | | | | | | | | | |
| 18 | 17 | 12,00 | 8,10 | 30,00 | | | | | | | | | | |
| 19 | 18 | 12,00 | 8,50 | 31,00 | | | | | | | | | | |
| 20 | 19 | 14,00 | 9,60 | 32,00 | | | | | | | | | | |
| 21 | 20 | 14,00 | 9,00 | 36,00 | | | | | | | | | | |
| 22 | | | | | | | | | | | | | | |
| 23 | | | | | | | | | | | | | | |
| 24 | | | | | | | | | | | | | | |
| 25 | | | | | | | | | | | | | | |
| 26 | | | | | | | | | | | | | | |
| 27 | | | | | | | | | | | | | | |
| 28 | | | | | | | | | | | | | | |
| 29 | | | | | | | | | | | | | | |
| 30 | | | | | | | | | | | | | | |
| 31 | | | | | | | | | | | | | | |
| 32 | | | | | | | | | | | | | | |
| 33 | | | | | | | | | | | | | | |
| 34 | | | | | | | | | | | | | | |

Рис. 6.2. Результат применения инструмента *Описательная статистика*

Сравнивая значения средних квадратических $\sigma_y$, $\sigma_{x1}$, $\sigma_{x21}$ отклонений и средних величин $\bar{y}$, $\bar{x}_1$ $\bar{x}_2$ и определяя коэффициенты вариации, приходим к выводу о повышенном уровне варьирования признаков, хотя и в допустимых пределах, не превышающих 35%.

$$\nu_y = \frac{\sigma_y}{\bar{y}} * 100\% = \frac{2{,}45807}{9{,}6} * 100\% = 25{,}6\%;$$

$$\nu_{x_1} = \frac{\sigma_{x_1}}{\bar{x}_1} * 100\% = \frac{1{,}93877}{6{,}19} * 100\% = 31{,}3\%;$$

$$\nu_{x_{21}} = \frac{\sigma_{x_{21}}}{\bar{x}_2} * 100\% = \frac{6{,}81407}{22{,}3} * 100\% = 30{,}6\%.$$

Следовательно, совокупность предприятий однородна, и для ее изучения могут использоваться метод наименьших квадратов и вероятностные методы оценки статистических гипотез.

***2.*** Значения линейных коэффициентов парной корреляции определяют тесноту попарно связанных переменных, использованных в данном уравнении множественной регрессии. Линейные коэффициенты частной корреляции оценивают тесноту связи значений двух переменных, исключая влияние всех других переменных, представленных в уравнении множественной регрессии.

К сожалению, в ППП Excel нет специального инструмента для расчета линейных коэффициентов частной корреляции. Матрицу парных коэффициентов корреляции переменных можно рассчитать, используя инструмент анализа данных **Корреляция.** Для этого:

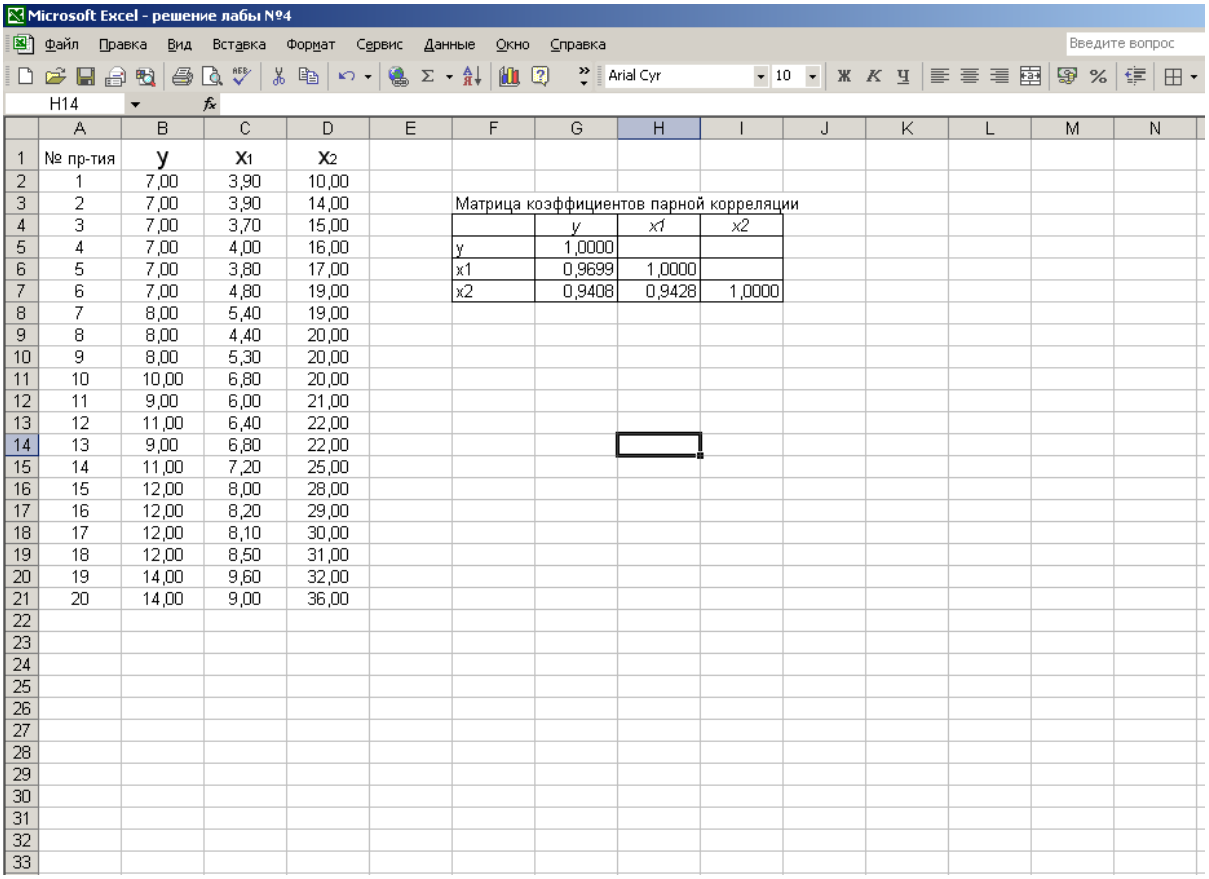
1) в главном меню последовательно выберите пункты **Сервис/ Анализ данных/ Корреляция.** Щелкните по кнопке **ОК**;

2) заполните диалоговое окно ввода данных и параметров вывода;

3) результаты вычислений – матрица коэффициентов парной корреляции – представлены на рис. 6.3.

Значения коэффициентов парной корреляции указывают на весьма тесную связь выработки ***y*** как с коэффициентом обновления основных фондов – $x_1$, так и с долей рабочих высокой квалификации

– $x_2$ ($r_{yx_1} = 0{,}9699$ и $r_{yx_2} = 0{,}9408$). Но в то же время, межфакторная связь $r_{x_1x_2} = 0{,}9428$ весьма тесная и превышает тесноту связи $x_2$ с у. В связи с этим для улучшения данной модели можно исключить из нее фактор $x_2$ как малоинформативный, недостаточно статистически надежный.




| | A | B | C | D |
|---|---|---|---|---|
| 1 | № пр-тия | y | x1 | x2 |
| 2 | 1 | 7,00 | 3,90 | 10,00 |
| 3 | 2 | 7,00 | 3,90 | 14,00 |
| 4 | 3 | 7,00 | 3,70 | 15,00 |
| 5 | 4 | 7,00 | 4,00 | 16,00 |
| 6 | 5 | 7,00 | 3,80 | 17,00 |
| 7 | 6 | 7,00 | 4,80 | 19,00 |
| 8 | 7 | 8,00 | 5,40 | 19,00 |
| 9 | 8 | 8,00 | 4,40 | 20,00 |
| 10 | 9 | 8,00 | 5,30 | 20,00 |
| 11 | 10 | 10,00 | 6,80 | 20,00 |
| 12 | 11 | 9,00 | 6,00 | 21,00 |
| 13 | 12 | 11,00 | 6,40 | 22,00 |
| 14 | 13 | 9,00 | 6,80 | 22,00 |
| 15 | 14 | 11,00 | 7,20 | 25,00 |
| 16 | 15 | 12,00 | 8,00 | 28,00 |
| 17 | 16 | 12,00 | 8,20 | 29,00 |
| 18 | 17 | 12,00 | 8,10 | 30,00 |
| 19 | 18 | 12,00 | 8,50 | 31,00 |
| 20 | 19 | 14,00 | 9,60 | 32,00 |
| 21 | 20 | 14,00 | 9,00 | 36,00 |

Матрица коэффициентов парной корреляции

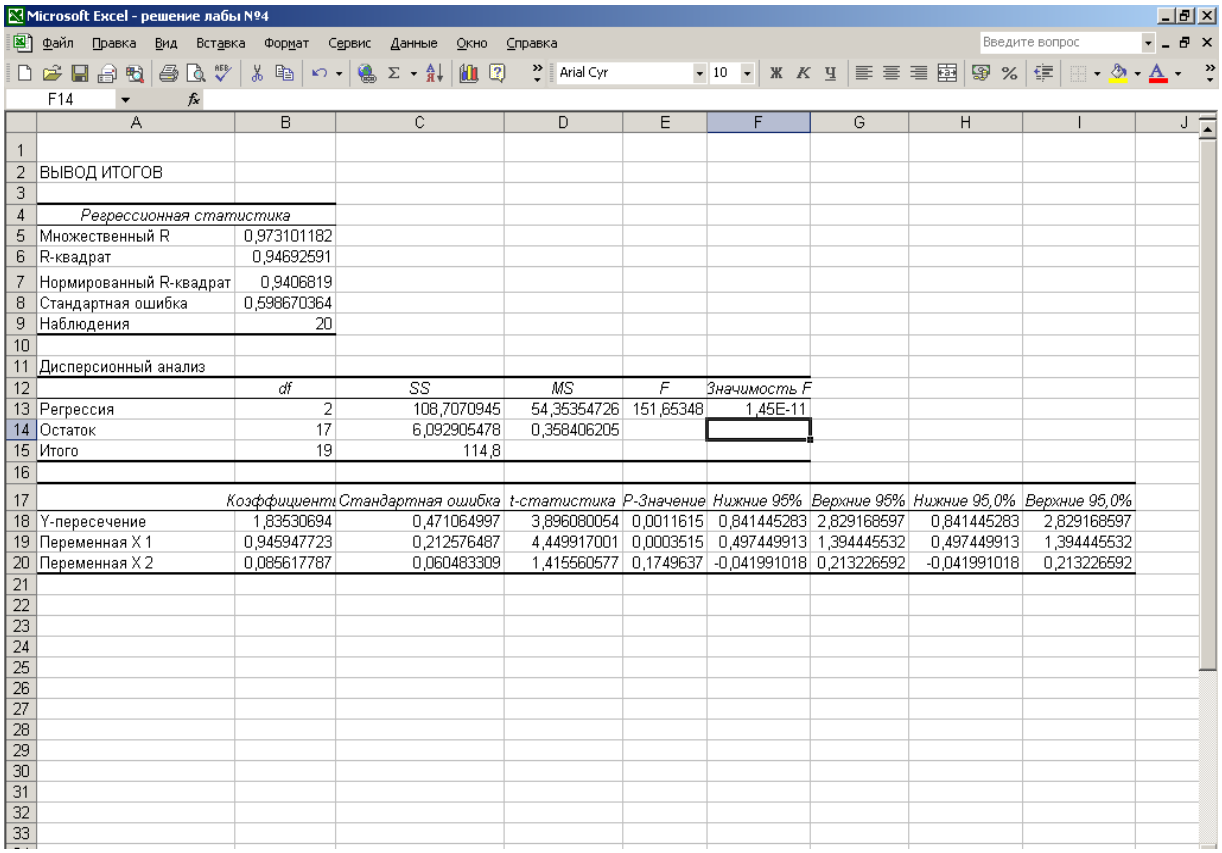
| | y | x1 | x2 |
|---|---|---|---|
| y | 1,0000 | | |
| x1 | 0,9699 | 1,0000 | |
| x2 | 0,9408 | 0,9428 | 1,0000 |

Рис. 6.3. Матрица коэффициентов парной корреляции

Коэффициенты частной корреляции дают более точную характеристику тесноты связи двух признаков, чем коэффициенты парной корреляции**.** Если сравнивать коэффициенты парной и частной корреляции, можно сказать, что из-за высокой межфакторной зависимости коэффициенты парной корреляции дают завышенные оценки тесноты связи, именно по этой причине рекомендуется при наличии сильной коллинеарности (взаимосвязи) факторов исключать из исследования тот фактор, у которого теснота парной зависимости меньше, чем теснота межфакторной связи.

***3.*** Вычисление параметров линейного уравнения множественной регрессии.

Эта операция проводится с помощью инструмента анализа данных **Регрессия.** Она аналогична расчету параметров парной линейной регрессии, а отличие от парной регрессии состоит только в том, что в диалоговом окне при заполнении параметров *входной интервал* X следует указывать не один столбец, а все столбцы, содержащие значения факторных признаков. Результаты анализа представлены на рис. 6.4.




| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | |
| 2 | ВЫВОД ИТОГОВ | | | | | | | | | |
| 3 | | | | | | | | | | |
| 4 | *Регрессионная статистика* | | | | | | | | | |
| 5 | Множественный R | 0,973101182 | | | | | | | | |
| 6 | R-квадрат | 0,94692591 | | | | | | | | |
| 7 | Нормированный R-квадрат | 0,9406819 | | | | | | | | |
| 8 | Стандартная ошибка | 0,598670364 | | | | | | | | |
| 9 | Наблюдения | 20 | | | | | | | | |
| 10 | | | | | | | | | | |
| 11 | Дисперсионный анализ | | | | | | | | | |
| 12 | | *df* | *SS* | *MS* | *F* | *Значимость F* | | | | |
| 13 | Регрессия | 2 | 108,7070945 | 54,35354726 | 151,65348 | 1,45E-11 | | | | |
| 14 | Остаток | 17 | 6,092905478 | 0,358406205 | | | | | | |
| 15 | Итого | 19 | 114,8 | | | | | | | |
| 16 | | | | | | | | | | |
| 17 | | *Коэффициенты* | *Стандартная ошибка* | *t-статистика* | *P-Значение* | *Нижние 95%* | *Верхние 95%* | *Нижние 95,0%* | *Верхние 95,0%* | |
| 18 | Y-пересечение | 1,83530694 | 0,471064997 | 3,896080054 | 0,0011615 | 0,841445283 | 2,829168597 | 0,841445283 | 2,829168597 | |
| 19 | Переменная X 1 | 0,945947723 | 0,212576487 | 4,449917001 | 0,0003515 | 0,497449913 | 1,394445532 | 0,497449913 | 1,394445532 | |
| 20 | Переменная X 2 | 0,085617787 | 0,060483309 | 1,415560577 | 0,1749637 | -0,041991018 | 0,213226592 | -0,041991018 | 0,213226592 | |
| 21 | | | | | | | | | | |
| 22 | | | | | | | | | | |
| 23 | | | | | | | | | | |
| 24 | | | | | | | | | | |
| 25 | | | | | | | | | | |
| 26 | | | | | | | | | | |
| 27 | | | | | | | | | | |
| 28 | | | | | | | | | | |
| 29 | | | | | | | | | | |
| 30 | | | | | | | | | | |
| 31 | | | | | | | | | | |
| 32 | | | | | | | | | | |
| 33 | | | | | | | | | | |

Рис. 6.4. Результат применения инструмента *Регрессия*

По результатам вычислений составим уравнение множественной регрессии вида

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2;$$

$$\hat{y} = 1{,}8353 + 0{,}9459 * x_1 + 0{,}0856 * x_2.$$

Величина $b_0$ оценивает агрегированное влияние прочих (кроме учтенных в модели факторов $x_2$ и $x_1$) факторов на результат y. Вели-

чины $b_1$ и $b_2$ указывают, что с увеличением $x_2$ и $x_1$ на единицу результат увеличивается соответственно на 0,9459 и 0,0856 млн. руб. Сравнивать эти значения не следует, т.к. они зависят от единиц измерения каждого признака и потому несопоставимы между собой.

Значения случайных ошибок параметров $b_0$, $b_1$ и $b_2$ с учетом округления составят: $m_{b_0} = 0{,}4711$, $m_{b_1} = 0{,}2126$, $m_{b_2} = 0{,}0605$. Они показывают, какое значение данной характеристики сформировалось под влиянием случайных факторов. Эти значения используются для расчета t-критерия Стьюдента $t_{b_0} = 3{,}90$; $t_{b_1} = 4{,}45$; $t_{b_2} = 1{,}42$.

Если значения t – критерия больше 2 – 3, можно сделать вывод о существенности данного параметра, который формируется под воздействием неслучайных причин. Здесь статистически значимыми являются $b_0$ и $b_1$, а величина $b_2$ сформировалась под воздействием случайных причин, поэтому фактор $x_2$, силу влияния которого оценивает $b_2$, можно исключить как несущественно влияющий, неинформативный.

На это же указывает показатель вероятности случайных значений параметров регрессии: если $\alpha$ меньше принятого нами уровня (обычно 0,1; 0,05 или 0,01; это соответствует 10%, 5% или 1% вероятности), делают вывод о несущественной природе данного значения параметра, т.е. о том, что он статистически значим и надежен. В противном случае принимается гипотеза о случайной природе значения коэффициентов уровня. Здесь $\alpha_{x_2} = 17{,}5\% > 5\%$, что позволяет рассматривать $x_2$ как неинформативный фактор и удалить его для улучшения данного уравнения.

***4.*** Оценку надежности уравнения регрессии в целом и показателя тесноты связи $R_{yx_1x_2}$ дает F-критерий Фишера:

$$F_{факт} = \frac{\Sigma\left(\hat{y}_{x_1x_2} - \hat{y}\right)^2}{m} / \frac{\Sigma\left(y - \hat{y}_{x_1x_2}\right)^2}{n-m-1} = 151{,}65.$$

По данным таблицы дисперсионного анализа, представленной на рис. 6.4, $F_{факт}$=151,65. Вероятность случайно получить такое значение F-критерия составляет 0 (см. значимость *F*), что не превышает допустимый уровень значимости 5%; об этом свидетельствует величина *P – значение* из этой же таблицы. Следовательно, полученное значение неслучайно, оно сформировалось под влиянием существенных факторов, т.е. подтверждается статистическая значимость всего уравнения и показателя тесноты связи $R^2_{yx_1x_2}$.

Значения скорректированного и нескорректированного линейных коэффициентов множественной детерминации приведены на рис. 6.4 в рамках регрессионной статистики. Нескорректированный коэффициент множественной детерминации $R^2_{yx_1x_2}$ =0,9469 оценивает долю вариации результата за счет представленных в уравнении факторов в общей вариации результата. Здесь эта доля составляет 94,7% и указывает на весьма высокую степень обусловленности вариации результата вариацией факторов, иными словами – на весьма тесную связь факторов с результатом.

Скорректированный коэффициент множественной детерминации $R^2_{yx_1x_2}$ =0,9407 определяет тесноту связи с учетом степеней свободы общей и остаточной дисперсий. Он дает такую оценку тесноты связи, которая не зависит от числа факторов в модели и потому может сравниваться по разным моделям с разным числом факторов. Оба коэффициента указывают на весьма высокую (более 90%) детерминированность результата *y* в модели факторами $x_1$ и $x_2$.

***5.*** Информация для оценки с помощью частных F- критериев Фишера целесообразности включения в модель фактора $x_1$ после фактора $x_2$ и фактора $x_2$ после фактора $x_1$ может быть получена в **ППП Statgraphics**. Частный F- критерий показывает статистическую значимость включения фактора $x_2$ после того, как в нее включен фактор $x_1$.

Но по данным, вычисленным с помощью ППП Excel, можно сделать общий вывод, который состоит в том, что множественная мо-

дель с факторами $x_1$ и $x_2$ с $R^2_{yx_1x_2}$ =0,9469 содержит неинформативный фактор $x_2$. Если исключить фактор $x_2$, то можно ограничиться уравнением парной регрессии более простым, хорошо детерминированным, пригодным для анализа и для прогноза.

$$\hat{y}_x = \alpha_0 + \alpha_1 x = 1{,}99 + 1{,}23 * x \ ;$$

$$r^2_{yx} = 0{,}9407.$$

***6.*** Средние частные коэффициенты эластичности $\overline{Э}_{yxj}$ показывают, на сколько процентов от значения своей средней $\bar{y}$ изменяется результат при изменении фактора $x_j$ на 1% от своей средней $\bar{x}_j$ и при фиксированном воздействии на у всех прочих факторов, включенных в уравнение регрессии. Для линейной зависимости

$$\overline{Э}_{yxj} = b_j \frac{\bar{x}_j}{\bar{y}}, \qquad (5.1)$$

где $b_j$ - коэффициент регрессии при $x_j$ в уравнении множественной регрессии.

Здесь $\overline{Э}_{yx_1} = \frac{0{,}9459 * 6{,}19}{9{,}6} = 0{,}6099\%$,

$\overline{Э}_{yx_2} = \frac{0{,}0856 * 22{,}3}{9{,}6} = 0{,}1989\%$.

По значениям частных коэффициентов эластичности можно сделать вывод о более сильном влиянии на результат у признака фактора $x_1$, чем признака фактора $x_2$: 0,6% против 0,2%.

**Решение с помощью R**

**1. Подготовка данных и загрузка необходимых библиотек**

Первым шагом необходимо создать скрипт в R, загрузить данные и подключить библиотеки, которые потребуются для анализа.

```
r
# ================================================
# Множественный регрессионный анализ в R
# ================================================
```

```r
# Очистка рабочей области (опционально)
rm(list = ls())

# Подключение необходимых библиотек
# install.packages(c("psych", "car", "lmtest",
"ggplot2", "corrplot"))
library(psych)   # для описательной статистики
library(car)    # для VIF и дополнительных тестов
library(lmtest)   # для тестов гетероскедастичности
library(ggplot2)  # для визуализации
library(corrplot)  # для визуализации корреляций
library(stats)   # базовые статистические функции
library(ggpubr)   # для объединения графиков

# Ввод исходных данных
# Создаем векторы с данными из таблицы
y <- c(7, 7, 7, 7, 7, 7, 8, 8, 8, 10, 9, 11, 9, 11,
12, 12, 12, 12, 14, 14)
x1 <- c(3.9, 3.9, 3.7, 4.0, 3.8, 4.8, 5.4, 4.4, 5.3,
6.8, 6.0, 6.4, 6.8, 7.2, 8.0, 8.2, 8.1, 8.5, 9.6, 9.0)
x2 <- c(10, 14, 15, 16, 17, 19, 19, 20, 20, 20, 21,
22, 22, 25, 28, 29, 30, 31, 32, 36)

# Создаем датафрейм для удобства работы
data <- data.frame(
 предприятие = 1:20,
 y = y,
 x1 = x1,
 x2 = x2
)

# Просмотр первых строк данных
print("Первые 6 строк данных:")
head(data)
```

**2. Оценка показателей вариации (аналог "Описательная статистика" в Excel)**

В R аналогом инструмента "Описательная статистика" является функция summary() для базовой статистики и функция describe() из пакета psych для более подробного анализа.

```
# ================================================
# 1. Оценка показателей вариации
# ================================================

# Базовая описательная статистика
print("Базовая описательная статистика:")
summary(data[, c("y", "x1", "x2")])

# Расширенная описательная статистика с использованием пакета psych
print("Расширенная описательная статистика (psych):")
describe(data[, c("y", "x1", "x2")])

# Ручной расчет коэффициентов вариации
# Коэффициент вариации = (Стандартное отклонение / Среднее) * 100%

stats <- data.frame(
 Признак = c("y (выработка)", "x1 (обновление фон-дов)", "x2 (доля квалифицированных)"),
 Среднее = c(mean(data$y), mean(data$x1), mean(data$x2)),
 Ст_отклонение = c(sd(data$y), sd(data$x1), sd(data$x2))
)
stats$Коэф_вариации <- round((stats$Ст_отклонение / stats$Среднее) * 100, 2)

print("Расчет коэффициентов вариации:")
print(stats)
```

```
# Вывод об однородности совокупности
cat("\nВЫВОД:",
  "\nКоэффициенты вариации:", stats$Коэф_вариации,
  "\nВсе коэффициенты вариации менее 35%, следователь-
но, совокупность предприятий однородна.",
  "\nЭто означает, что для изучения зависимости можно
применять метод наименьших квадратов",
  "\nи вероятностные методы оценки статистических ги-
потез.\n")
```

**Результат выполнения кода должен показать**, что коэффициенты вариации находятся в допустимых пределах (менее 35%), что свидетельствует об однородности совокупности.

### 3. Анализ коэффициентов парной и частной корреляции

#### 3.1. Матрица парных коэффициентов корреляции (аналог инструмента "Корреляция" в Excel)

```
# ================================================
# 2. Анализ коэффициентов корреляции
# 2.1 Матрица парных коэффициентов корреляции
# ================================================

# Расчет матрицы парных корреляций
cor_matrix <- cor(data[, c("y", "x1", "x2")])
print("Матрица парных коэффициентов корреляции:")
print(round(cor_matrix, 4))

# Визуализация матрицы корреляций
par(mfrow = c(1, 1)) # возвращаем стандартный режим
отображения
corrplot(cor_matrix,
     method = "number",
     type = "upper",
     tl.col = "black",
     tl.cex = 0.8,
     number.cex = 0.8,
     title = "Матрица парных коэффициентов
корреляции",
```

```
      mar = c(0, 0, 2, 0))

# Анализ корреляций
cat("\nАНАЛИЗ ПАРНЫХ КОЭФФИЦИЕНТОВ КОРРЕЛЯЦИИ:",
  "\n- Связь между y и x1 (коэффициент обновления): r
=", round(cor_matrix[1,2], 4),
  "\n- Связь между y и x2 (доля квалифицированных): r
=", round(cor_matrix[1,3], 4),
  "\n- Межфакторная связь (x1 и x2): r =",
round(cor_matrix[2,3], 4))

# Проверка на мультиколлинеарность
if (abs(cor_matrix[2,3]) > 0.7) {
 cat("\n\nПРОБЛЕМА: Межфакторная корреляция превышает
0.7, что указывает на наличие",
   "\nмультиколлинеарности. Это может привести к завы-
шенным оценкам тесноты связи",
   "\nи снижению надежности модели.")
} else {
 cat("\n\nМежфакторная корреляция не превышает крити-
ческий уровень 0.7,",
   "\nпроблема мультиколлинеарности отсутствует или
незначительна.")
}
```

**3.2. Расчет коэффициентов частной корреляции**

В отличие от Excel, в R можно легко рассчитать коэффициенты частной корреляции, которые очищены от влияния других переменных.

```
# ================================================
# 2.2 Коэффициенты частной корреляции
# ================================================

# Функция для расчета частной корреляции между y и x1
при фиксированном x2
# Используем формулу: r_yx1.x2 = (r_yx1 - r_yx2 *
r_x1x2) / sqrt((1 - r_yx2^2) * (1 - r_x1x2^2))
```

```
r_yx1 <- cor_matrix[1, 2]
r_yx2 <- cor_matrix[1, 3]
r_x1x2 <- cor_matrix[2, 3]

# Частная корреляция y с x1 при фиксированном x2
r_yx1_x2 <- (r_yx1 - r_yx2 * r_x1x2) / sqrt((1 -
r_yx2^2) * (1 - r_x1x2^2))

# Частная корреляция y с x2 при фиксированном x1
r_yx2_x1 <- (r_yx2 - r_yx1 * r_x1x2) / sqrt((1 -
r_yx1^2) * (1 - r_x1x2^2))

# Частная корреляция x1 с x2 при фиксированном y (для
полноты анализа)
r_x1x2_y <- (r_x1x2 - r_yx1 * r_yx2) / sqrt((1 -
r_yx1^2) * (1 - r_yx2^2))

print("Коэффициенты частной корреляции:")
cat("\n1. Частная корреляция между y и x1 (при исклю-
чении влияния x2): r_yx1.x2 =", round(r_yx1_x2, 4))
cat("\n2. Частная корреляция между y и x2 (при исклю-
чении влияния x1): r_yx2.x1 =", round(r_yx2_x1, 4))
cat("\n3. Частная корреляция между x1 и x2 (при исклю-
чении влияния y): r_x1x2.y =", round(r_x1x2_y, 4))

# Сравнение парных и частных коэффициентов
cat("\n\nСРАВНЕНИЕ ПАРНЫХ И ЧАСТНЫХ КОЭФФИЦИЕНТОВ:",
  "\n- Парная корреляция y-x1:", round(r_yx1, 4), "|
Частная корреляция y-x1.x2:", round(r_yx1_x2, 4),
  "\n- Парная корреляция y-x2:", round(r_yx2, 4), "|
Частная корреляция y-x2.x1:", round(r_yx2_x1, 4))

cat("\n\nВЫВОД: Из-за высокой межфакторной корреляции
парные коэффициенты дают завышенные оценки",
  "\nтесноты связи по сравнению с частными. Это под-
тверждает наличие эффекта мультиколлинеарности.")
```

**4. Построение уравнения множественной регрессии**

```
# ================================================
# 3. Построение уравнения множественной регрессии
# ================================================

# Построение модели множественной регрессии
model <- lm(y ~ x1 + x2, data = data)

# Вывод результатов модели
print("РЕЗУЛЬТАТЫ МНОЖЕСТВЕННОЙ РЕГРЕССИИ:")
summary(model)

# Извлечение коэффициентов
coef_model <- coef(model)
b0 <- coef_model[1]
b1 <- coef_model[2]
b2 <- coef_model[3]

cat("\nУРАВНЕНИЕ МНОЖЕСТВЕННОЙ РЕГРЕССИИ:")
cat("\nŷ =", round(b0, 4), "+", round(b1, 4), "* x1
+", round(b2, 4), "* x2")
cat("\n\nЭКОНОМИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ:")
cat("\n- Свободный член b0 =", round(b0, 4), "оценива-
ет агрегированное влияние прочих",
  "\n (не учтенных в модели) факторов на выработку
продукции.")
cat("\n- Коэффициент b1 =", round(b1, 4), "показывает,
что при увеличении коэффициента обновления",
  "\n основных фондов на 1% (при фиксированной доле
квалифицированных рабочих)",
  "\n выработка продукции увеличивается в среднем на",
round(b1, 4), "млн руб.")
cat("\n- Коэффициент b2 =", round(b2, 4), "показывает,
что при увеличении доли рабочих высокой",
  "\n квалификации на 1% (при фиксированном коэффици-
енте обновления фондов)",
```

```
  "\n выработка продукции увеличивается в среднем на",
round(b2, 4), "млн руб.")

# Стандартные ошибки коэффициентов
std_errors <- summary(model)$coefficients[, "Std. Er-
ror"]
cat("\n\nСТАНДАРТНЫЕ ОШИБКИ КОЭФФИЦИЕНТОВ (случайные
ошибки):")
cat("\n- SE(b0) =", round(std_errors[1], 4))
cat("\n- SE(b1) =", round(std_errors[2], 4))
cat("\n- SE(b2) =", round(std_errors[3], 4))

# t-статистики и p-значения
t_values <- summary(model)$coefficients[, "t value"]
p_values <- summary(model)$coefficients[, "Pr(>|t|)"]

cat("\n\nt-КРИТЕРИИ СТЬЮДЕНТА (для проверки значимости
коэффициентов):")
for (i in 1:length(t_values)) {
 sig_status <- ifelse(p_values[i] < 0.05, "значим",
"не значим")
 cat("\n-", names(t_values)[i], ": t =",
round(t_values[i], 3),
   ", p-value =", format.pval(p_values[i], digits =
4),
   "→ коэффициент", sig_status)
}

# Вывод о значимости факторов
if (p_values[3] > 0.05) {
 cat("\n\nВЫВОД: Коэффициент при x2 (доля квалифициро-
ванных рабочих) статистически не значим",
   "\n(p-value > 0.05). Это означает, что фактор x2
может быть неинформативным, и его",
   "\nможно исключить из модели для улучшения уравне-
ния.")
}
```

**5. Оценка надежности уравнения регрессии (F-критерий Фишера)**

```
# ================================================
# 4. Оценка надежности уравнения регрессии
# ================================================

# Извлечение данных для F-критерия из сводки модели
model_summary <- summary(model)

# R² и скорректированный R²
r_squared <- model_summary$r.squared
adj_r_squared <- model_summary$adj.r.squared

cat("КОЭФФИЦИЕНТЫ ДЕТЕРМИНАЦИИ:")
cat("\n- Нескорректированный R² =", round(r_squared,
4))
cat("\n- Скорректированный R² =", round(adj_r_squared,
4))
cat("\n\nИнтерпретация: R² =", round(r_squared, 4),
"означает, что",
  round(r_squared * 100, 2), "% вариации выработки
продукции объясняется",
  "вариацией включенных в модель факторов (коэффициент
обновления фондов",
  "и доля квалифицированных рабочих).")

# F-статистика
f_statistic <- model_summary$fstatistic
f_value <- f_statistic[1]
df1 <- f_statistic[2] # число степеней свободы для ре-
грессии
df2 <- f_statistic[3] # число степеней свободы для
остатков

# Расчет p-value для F-статистики
f_p_value <- pf(f_value, df1, df2, lower.tail = FALSE)
```

```
cat("\n\nF-КРИТЕРИЙ ФИШЕРА:")
cat("\n- F-статистика =", round(f_value, 4))
cat("\n- Степени свободы: df1 =", df1, ", df2 =", df2)
cat("\n- p-value =", format.pval(f_p_value, digits =
6))

if (f_p_value < 0.05) {
 cat("\n\nВЫВОД: p-value < 0.05, следовательно, урав-
нение регрессии статистически значимо.",
   "\nПолученное значение F-критерия неслучайно, оно
сформировалось под влиянием",
   "\nсущественных факторов. Подтверждается статисти-
ческая значимость всего уравнения",
   "\nи показателя тесноты связи R².")
} else {
 cat("\n\nВЫВОД: p-value > 0.05, уравнение регрессии
статистически не значимо.")
}

# Таблица дисперсионного анализа (ANOVA)
cat("\n\nДИСПЕРСИОННЫЙ АНАЛИЗ (ANOVA):")
anova_table <- anova(model)
print(anova_table)
```

**6. Частные F-критерии Фишера (целесообразность включения факторов)**

```
# ================================================
# 5. Частные F-критерии Фишера
# ================================================

# Для оценки целесообразности включения фактора x2 по-
сле x1
# Строим модель только с x1
model_x1 <- lm(y ~ x1, data = data)

# Строим полную модель (уже есть – model)
# Частный F-критерий для включения x2 после x1
```

```r
# Расчет сумм квадратов
rss_x1 <- sum(residuals(model_x1)^2) # остаточная сумма квадратов для модели с x1
rss_full <- sum(residuals(model)^2)  # остаточная сумма квадратов для полной модели
df_diff <- 1 # разность числа факторов (включили один новый фактор)
df_resid_full <- df.residual(model)  # остаточные степени свободы полной модели

# Частный F-критерий для x2 после x1
f_partial_x2 <- ((rss_x1 - rss_full) / df_diff) / (rss_full / df_resid_full)
p_value_partial_x2 <- pf(f_partial_x2, df_diff, df_resid_full, lower.tail = FALSE)

# Аналогично для включения x1 после x2
model_x2 <- lm(y ~ x2, data = data)
rss_x2 <- sum(residuals(model_x2)^2)

f_partial_x1 <- ((rss_x2 - rss_full) / df_diff) / (rss_full / df_resid_full)
p_value_partial_x1 <- pf(f_partial_x1, df_diff, df_resid_full, lower.tail = FALSE)

cat("ЧАСТНЫЕ F-КРИТЕРИИ ФИШЕРА:")
cat("\n\n1. Целесообразность включения фактора x2 (доля квалифицированных) после x1:")
cat("\n  F-статистика =", round(f_partial_x2, 4))
cat("\n  p-value =", format.pval(p_value_partial_x2, digits = 4))

if (p_value_partial_x2 < 0.05) {
```

```
 cat("\n  ВЫВОД: p-value < 0.05 → включение фактора x2
после x1 статистически обосновано.")
} else {
 cat("\n  ВЫВОД: p-value > 0.05 → включение фактора x2
после x1 НЕ обосновано,",
   "\n  фактор x2 можно исключить из модели как неин-
формативный.")
}

cat("\n\n2. Целесообразность включения фактора x1 (об-
новление фондов) после x2:")
cat("\n  F-статистика =", round(f_partial_x1, 4))
cat("\n  p-value =", format.pval(p_value_partial_x1,
digits = 4))

if (p_value_partial_x1 < 0.05) {
 cat("\n  ВЫВОД: p-value < 0.05 → включение фактора x1
после x2 статистически обосновано.")
} else {
 cat("\n  ВЫВОД: p-value > 0.05 → включение фактора x1
после x2 НЕ обосновано.")
}

# Дополнительно: VIF для проверки мультиколлинеарности
cat("\n\nПРОВЕРКА МУЛЬТИКОЛЛИНЕАРНОСТИ (VIF):")
vif_values <- vif(model)
print(vif_values)

for (i in 1:length(vif_values)) {
 if (vif_values[i] > 10) {
  cat("\n- VIF для", names(vif_values)[i], "=",
round(vif_values[i], 2),
    "> 10 → критическая мультиколлинеарность")
 } else if (vif_values[i] > 5) {
```

```
  cat("\n- VIF для", names(vif_values)[i], "=",
round(vif_values[i], 2),
    "> 5 → умеренная мультиколлинеарность")
 } else {
  cat("\n- VIF для", names(vif_values)[i], "=",
round(vif_values[i], 2),
    "→ мультиколлинеарность незначительна")
 }
}
```

### 7. Коэффициенты эластичности

```
# ================================================
# 6. Средние частные коэффициенты эластичности
# ================================================

# Расчет средних значений
mean_y <- mean(data$y)
mean_x1 <- mean(data$x1)
mean_x2 <- mean(data$x2)

# Коэффициенты эластичности по формуле: Э_j = b_j *
(mean_xj / mean_y)
E1 <- b1 * (mean_x1 / mean_y)
E2 <- b2 * (mean_x2 / mean_y)

cat("СРЕДНИЕ ЧАСТНЫЕ КОЭФФИЦИЕНТЫ ЭЛАСТИЧНОСТИ:")
cat("\n\nКоэффициент эластичности для x1 (обновление
фондов):")
cat("\nЭ1 =", round(b1, 4), "* (", round(mean_x1, 2),
"/", round(mean_y, 2), ") =", round(E1 * 100, 2), "%")
cat("\nИнтерпретация: При увеличении коэффициента об-
новления основных фондов на 1%",
  "\not своего среднего уровня (при фиксированной доле
квалифицированных рабочих)",
```

```
  "\nвыработка продукции увеличивается на", round(E1 *
100, 2), "% от своего среднего уровня.")

cat("\n\nКоэффициент эластичности для x2 (доля квали-
фицированных):")
cat("\nЭ2 =", round(b2, 4), "* (", round(mean_x2, 2),
"/", round(mean_y, 2), ") =", round(E2 * 100, 2), "%")
cat("\nИнтерпретация: При увеличении доли рабочих вы-
сокой квалификации на 1%",
  "\nот своего среднего уровня (при фиксированном ко-
эффициенте обновления фондов)",
  "\nвыработка продукции увеличивается на", round(E2 *
100, 2), "% от своего среднего уровня.")

cat("\n\nСРАВНИТЕЛЬНАЯ ОЦЕНКА СИЛЫ ВЛИЯНИЯ ФАКТОРОВ:")
if (abs(E1) > abs(E2)) {
 cat("\nФактор x1 (обновление фондов) оказывает более
сильное влияние на результат (",
   round(E1 * 100, 2), "%), чем фактор x2 (", round(E2
* 100, 2), "%).")
} else {
 cat("\nФактор x2 (доля квалифицированных) оказывает
более сильное влияние на результат (",
   round(E2 * 100, 2), "%), чем фактор x1 (", round(E1
* 100, 2), "%).")
}
```

### 8. Диагностика модели и визуализация

```
# ================================================
# 7. Диагностика модели и визуализация
# ================================================

# Добавим предсказанные значения и остатки в датафрейм
data$fitted <- fitted(model)
```

```r
data$residuals <- residuals(model)
data$std_residuals <- rstandard(model)

# 7.1 График соответствия фактических и предсказанных значений
p1 <- ggplot(data, aes(x = fitted, y = y)) +
 geom_point(size = 3, color = "blue") +
 geom_abline(intercept = 0, slope = 1, color = "red",
linetype = "dashed", size = 1) +
 labs(title = "Фактические vs Предсказанные значения",
    x = "Предсказанные значения y", y = "Фактические
значения y") +
 theme_minimal() +
 annotate("text", x = min(data$fitted), y =
max(data$y),
      label = paste("R² =", round(r_squared, 4)),
hjust = 0)

# 7.2 График остатков от предсказанных значений
p2 <- ggplot(data, aes(x = fitted, y = residuals)) +
 geom_point(size = 3, color = "darkgreen") +
 geom_hline(yintercept = 0, color = "red", linetype =
"dashed") +
 geom_smooth(method = "loess", se = TRUE, color =
"blue", fill = "lightblue") +
 labs(title = "Остатки vs Предсказанные значения",
    x = "Предсказанные значения", y = "Остатки") +
 theme_minimal()

# 7.3 Q-Q plot для проверки нормальности остатков
p3 <- ggplot(data, aes(sample = residuals)) +
 stat_qq(size = 3) +
 stat_qq_line(color = "red") +
 labs(title = "Q-Q график (нормальность остатков)",
```

```
    x = "Теоретические квантили", y = "Выборочные
квантили") +
 theme_minimal()

# 7.4 Масштабно-локационный график (проверка го-
москедастичности)
p4 <- ggplot(data, aes(x = fitted, y =
sqrt(abs(std_residuals)))) +
 geom_point(size = 3, color = "purple") +
 geom_smooth(method = "loess", se = TRUE, color =
"red", fill = "lightpink") +
 labs(title = "Масштабно-локационный график",
    x = "Предсказанные значения", y =
"√|Стандартизованные остатки|") +
 theme_minimal()
# Объединение графиков
grid_plots <- ggarrange(p1, p2, p3, p4, ncol = 2, nrow
= 2)
print(grid_plots)

# 7.5 Тесты для проверки предпосылок МНК
cat("\n\nДИАГНОСТИКА МОДЕЛИ:")
cat("\n\n1. Тест на гетероскедастичность (Breusch-
Pagan):")
bp_test <- bptest(model)
print(bp_test)
if (bp_test$p.value > 0.05) {
 cat("  ВЫВОД: p-value > 0.05, гетероскедастичность
отсутствует (условие гомоскедастичности выполняет-
ся).")
} else {
 cat("  ВЫВОД: p-value < 0.05, присутствует гетеро-
скедастичность.")
}
```

```
cat("\n\n2. Тест на нормальность остатков (Shapiro-
Wilk):")
sw_test <- shapiro.test(data$residuals)
print(sw_test)
if (sw_test$p.value > 0.05) {
 cat("  ВЫВОД: p-value > 0.05, остатки распределены
нормально.")
} else {
 cat("  ВЫВОД: p-value < 0.05, распределение остатков
отличается от нормального.")
}

# 7.6 Дополнительный график: трехмерная визуализация
# (требует установки библиотеки scatterplot3d)
# install.packages("scatterplot3d")
library(scatterplot3d)

# Создание 3D графика
s3d <- scatterplot3d(data$x1, data$x2, data$y,
           pch = 16, color = "blue",
           main = "3D визуализация: y ~ x1 + x2",
           xlab = "x1 (обновление фондов)",
           ylab = "x2 (доля квалифицированных)",
           zlab = "y (выработка)",
           type = "p", angle = 45)

# Добавление плоскости регрессии
s3d$plane3d(model, lty.box = "solid", col = "red")
```

### 9. Прогнозирование по модели

```
# ================================================
# 8. Прогнозирование по модели
# ================================================
```

```r
# Создание нового наблюдения для прогноза
# Например, предприятие с x1 = 7.5 и x2 = 25
new_data <- data.frame(x1 = 7.5, x2 = 25)

# Точечный прогноз
point_prediction <- predict(model, newdata = new_data)

# Доверительный интервал для среднего значения
conf_interval <- predict(model, newdata = new_data,
interval = "confidence", level = 0.95)

# Интервал прогноза для индивидуального значения
pred_interval <- predict(model, newdata = new_data,
interval = "prediction", level = 0.95)

cat("ПРОГНОЗИРОВАНИЕ ПО МОДЕЛИ:")
cat("\n\nИсходные данные для прогноза:")
cat("\n- x1 (коэффициент обновления фондов) = 7.5%")
cat("\n- x2 (доля квалифицированных рабочих) = 25%")

cat("\n\nТочечный прогноз:")
cat("\nŷ =", round(point_prediction, 4), "млн руб.")

cat("\n\n95% доверительный интервал для среднего зна-
чения:")
cat("\nНижняя граница:", round(conf_interval[2], 4),
  "| Верхняя граница:", round(conf_interval[3], 4))
cat("\nИнтерпретация: С вероятностью 95% средняя выра-
ботка для всех предприятий",
```

```
  "\nс характеристиками x1=7.5, x2=25 находится в ука-
занных пределах.")
cat("\n\n95% интервал прогноза для индивидуального
значения:")
cat("\nНижняя граница:", round(pred_interval[2], 4),
  "| Верхняя граница:", round(pred_interval[3], 4))
cat("\nИнтерпретация: С вероятностью 95% выработка
конкретного предприятия",
  "\nс характеристиками x1=7.5, x2=25 будет находиться
в указанных пределах.")
```

### 10. Итоговый отчет

```
# ================================================
# 9. ИТОГОВЫЙ ОТЧЕТ ПО РЕЗУЛЬТАТАМ АНАЛИЗА
# ================================================

cat("\n", paste(rep("=", 80), collapse = ""))
cat("\n\t\tИТОГОВЫЙ ОТЧЕТ ПО МНОЖЕСТВЕННОМУ РЕГРЕССИ-
ОННОМУ АНАЛИЗУ")
cat("\n", paste(rep("=", 80), collapse = ""))

cat("\n\n1. ХАРАКТЕРИСТИКА ИСХОДНЫХ ДАННЫХ:")
cat("\n  - Объем выборки: n = 20 предприятий")
cat("\n  - Зависимая переменная: Y - выработка продук-
ции (млн руб.)")
cat("\n  - Фактор X1: коэффициент обновления основных
фондов (%)")
cat("\n  - Фактор X2: доля рабочих высокой квалифика-
ции (%)")
```

```
cat("\n  - Коэффициенты вариации: Y =",
round(stats$Коэф_вариации[1], 2),
  "%, X1 =", round(stats$Коэф_вариации[2], 2),
  "%, X2 =", round(stats$Коэф_вариации[3], 2), "%")
cat("\n  - Вывод: совокупность однородна, МНК приме-
ним.")

cat("\n\n2. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ:")
cat("\n  - Парная корреляция y-x1:",
round(cor_matrix[1,2], 4))
cat("\n  - Пар
```

3. Задание лабораторной работы.

В предложенных вариантах задач требуется разработать прогноз социально-экономического развития исследуемого показателя с применением метода многомерной регрессии, используя последовательность аналитических расчетов, приведенных в данном практическом примере.

## Лабораторная работа № 1
## Прогнозирование эффективности работы менеджеров по продажам

**Цель работы:** Построить модель множественной регрессии для прогнозирования объема продаж менеджеров на основе их личностных характеристик и профессиональных качеств.

**Контекст:** Крупная торговая компания "Регион-Сбыт" проводит отбор кандидатов на должность менеджера по продажам. Отделу персонала необходимо разработать статистическую модель, которая позволит прогнозировать будущую результативность сотрудников (ежемесячный объем продаж) на основе данных, собираемых при приеме на работу. В выборке представлены данные по 17 действующим сотрудникам.

**Исходные данные:**

| Сотрудник | Объем продаж за месяц (Y), ед. | Результат теста способностей (X1) | Возраст (X2), лет | Результат теста тревожности (X3) | Опыт работы (X4), лет | Средний балл аттестата (X5) |
|---|---|---|---|---|---|---|
| 1 | 44 | 10 | 22,1 | 4,9 | 0 | 2,4 |
| 2 | 47 | 19 | 22,5 | 3,0 | 1 | 2,6 |
| 3 | 60 | 27 | 23,1 | 1,5 | 0 | 2,8 |
| 4 | 71 | 31 | 24,0 | 0,6 | 3 | 2,7 |
| 5 | 61 | 64 | 22,6 | 1,8 | 2 | 2,0 |
| 6 | 60 | 81 | 21,7 | 3,3 | 1 | 2,5 |
| 7 | 58 | 42 | 22,0 | 3,2 | 0 | 2,5 |
| 8 | 56 | 67 | 22,4 | 2,1 | 0 | 2,3 |
| 9 | 66 | 48 | 22,6 | 6,0 | 1 | 2,8 |
| 10 | 61 | 64 | 21,1 | 1,8 | 1 | 3,4 |
| 11 | 51 | 57 | 22,5 | 3,8 | 0 | 3,0 |
| 12 | 47 | 10 | 22,2 | 4,5 | 1 | 2,7 |
| 13 | 53 | 48 | 24,8 | 4,5 | 0 | 2,8 |
| 14 | 74 | 96 | 24,8 | 0,1 | 3 | 3,8 |
| 15 | 65 | 75 | 22,6 | 0,9 | 0 | 3,7 |
| 16 | 33 | 12 | 20,5 | 4,8 | 0 | 2,1 |
| 17 | 54 | 47 | 21,9 | 2,3 | 1 | 1,8 |

**Задания:**

**Часть 1. Подготовка и предварительный анализ**

1. Определите зависимую переменную (Y) и независимые переменные (факторы X1-X5). Сформулируйте цель исследования.

2. Рассчитайте описательные статистики (среднее, минимум, максимум, стандартное отклонение) для всех переменных. Интерпретируйте полученные значения.

**Часть 2. Корреляционный анализ**

3. Постройте корреляционную матрицу для всех переменных. Выявите факторы, наиболее тесно связанные с объемом продаж (Y).

4. Проанализируйте корреляцию между независимыми переменными. Существует ли проблема мультиколлинеарности (наличие сильной связи между факторами)? Какие пары переменных имеют высокий коэффициент корреляции?

**Часть 3. Построение регрессионной модели**

5. Используя метод наименьших квадратов, постройте модель множественной линейной регрессии вида: $Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + b_3 \cdot X_3 + b_4 \cdot X_4 + b_5 \cdot X_5$

6. Запишите полученное уравнение регрессии. Дайте экономическую интерпретацию каждому коэффициенту регрессии.

**Часть 4. Реализация в Excel**

7. Используя надстройку «Пакет анализа» → инструмент «Регрессия», выполните расчеты. В полученной таблице найдите и выпишите:

Коэффициенты уравнения регрессии.

Множественный коэффициент детерминации $R^2$. Интерпретируйте его значение.

Стандартную ошибку регрессии.

Результаты дисперсионного анализа (ANOVA): F-статистику и ее значимость. Является ли модель в целом значимой?

p-значения для каждого коэффициента. Какие факторы являются статистически значимыми на уровне $\alpha = 0,05$?

8. На основе анализа значимости факторов предложите, какие переменные можно исключить из модели для ее упрощения.

**Часть 5. Реализация в R**

9. Напишите скрипт на языке R, который выполняет следующие задачи:

Создает датафрейм с исходными данными.

Строит матрицу корреляций и визуализирует ее с помощью библиотек corrplot или ggcorrplot.

Строит модель множественной линейной регрессии с помощью функции lm(). Сохраняет результат в объект.

Выводит в консоль подробную сводку по модели с помощью summary().

10. Используя функцию step() или stepAIC() (из библиотеки MASS), выполните пошаговый отбор наиболее значимых факторов (методом обратного исключения или прямого включения). Какая модель была выбрана как оптимальная?

11. Для итоговой модели постройте графики диагностики остатков (график "фактические значения – предсказанные значения", график остатков от предсказанных значений, Q-Q plot). Выполняется ли условие гомоскедастичности? Нормально ли распределены остатки?

**Часть 6. Прогнозирование**

12. Предположим, на собеседование пришел кандидат со следующими характеристиками:

Результат теста способностей (X1): 85 баллов

Возраст (X2): 23,5 года

Результат теста тревожности (X3): 2,0 балла

Опыт работы (X4): 2 года

Средний балл аттестата (X5): 3,5 балла

Используя построенную в R итоговую модель (после отбора факторов), спрогнозируйте ожидаемый месячный объем продаж для этого кандидата.

13. Постройте 95% доверительный интервал для среднего значения прогноза и 95% интервал прогноза для индивидуального значения.

**Часть 7. Выводы и заключение**

14. Напишите развернутый аналитический отчет для отдела персонала, который должен содержать ответы на вопросы:

Какие факторы в наибольшей степени влияют на результативность менеджеров по продажам?

Можно ли использовать предлагаемую модель для отбора кандидатов?

Каков прогнозируемый объем продаж для нового кандидата и какова точность этого прогноза?

Какие рекомендации по процедуре отбора можно дать компании на основе проведенного анализа?

## Лабораторная работа № 2
## Анализ времени обслуживания покупателей в розничной сети

**Цель работы:** Построить модель множественной регрессии для прогнозирования времени обслуживания покупателей в зависимости от параметров их покупок.

**Контекст:** Сеть супермаркетов "Эконом-Сити" проводит исследование с целью оптимизации работы кассовой зоны. Необходимо понять, какие факторы влияют на время обслуживания одного покупателя, чтобы более точно планировать загрузку персонала. Собраны данные по 18 покупателям.

**Исходные данные:**

| Покупатель | Время обслуживания (Y), мин. | Стоимость покупок (X1), ден. ед. | Количество единиц товара (X2) |
|---|---|---|---|
| 1 | 3,0 | 36 | 9 |
| 2 | 1,3 | 13 | 5 |
| 3 | 0,5 | 3 | 2 |
| 4 | 7,4 | 81 | 14 |
| 5 | 5,9 | 78 | 13 |
| 6 | 8,4 | 103 | 16 |
| 7 | 5,0 | 64 | 12 |
| 8 | 8,1 | 67 | 11 |
| 9 | 1,9 | 25 | 7 |
| 10 | 6,2 | 55 | 11 |
| 11 | 0,7 | 13 | 3 |
| 12 | 1,4 | 21 | 8 |
| 13 | 9,1 | 121 | 21 |
| 14 | 0,9 | 10 | 6 |
| 15 | 5,4 | 60 | 13 |
| 16 | 3,3 | 32 | 11 |
| 17 | 4,5 | 51 | 15 |
| 18 | 2,4 | 28 | 10 |

**Задания:**

**Часть 1. Предварительный анализ данных**

1. Определите зависимую и независимые переменные. Сформулируйте гипотезы о характере влияния каждого фактора на время обслуживания.

2. Постройте диаграммы рассеивания для каждой пары (Y с X1) и (Y с X2). Визуально оцените характер и тесноту связи.

**Часть 2. Корреляционный анализ**

3. Рассчитайте коэффициенты корреляции между всеми парами переменных. Заполните корреляционную матрицу.

4. Проанализируйте:

Какой фактор (стоимость покупок или количество единиц товара) сильнее коррелирует со временем обслуживания?

Существует ли мультиколлинеарность между независимыми переменными X1 и X2? Если да, то какие проблемы это может создать при построении регрессии?

**Часть 3. Построение регрессионной модели**

5. Постройте модель множественной линейной регрессии: $Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2$.

6. Запишите полученное уравнение. Интерпретируйте коэффициенты $b_1$ и $b_2$. Что происходит со временем обслуживания при увеличении стоимости покупок на 1 ден. ед.? При увеличении количества товаров на 1 единицу?

**Часть 4. Реализация в Excel**

7. Используя инструмент «Регрессия» (надстройка «Анализ данных»), получите результаты регрессионного анализа.

8. На основе выходной таблицы определите:

Множественный R и $R^2$. Какую долю вариации времени обслуживания объясняют включенные в модель факторы?

- Значимость модели в целом (F-статистика и ее p-значение).

Значимость каждого коэффициента регрессии (p-значения для t-статистики). Оба ли фактора являются статистически значимыми?

Запишите уравнение регрессии со стандартными ошибками коэффициентов.

9. Используя полученное уравнение, спрогнозируйте время обслуживания для покупателя, приобретающего **14 единиц товара** общей стоимостью **70 ден. ед.** (т.е. X1 = 70, X2 = 14).

**Часть 5. Реализация в R**

10. Напишите скрипт на R, выполняющий:

Загрузку данных и построение модели `lm(Y ~ X1 + X2, data = ...)`.

Вывод сводки `summary()`.

Построение трехмерного графика облака точек и плоскости регрессии (используя библиотеки `scatterplot3d` или `plotly`).

11. Проанализируйте остатки модели. Постройте график "остатки от предсказанных значений" и Q-Q plot для проверки нормальности остатков. Есть ли основания сомневаться в адекватности модели?

12. Для нового покупателя (X1 = 70, X2 = 14) получите:

Точечный прогноз.

95% доверительный интервал для среднего времени обслуживания всех покупателей с такими характеристиками.

95% интервал прогноза для времени обслуживания конкретного покупателя с такими характеристиками.

Сравните ширину доверительного интервала и интервала прогноза. Какой из них шире и почему?

**Часть 6. Сравнение моделей**

13. Постройте две дополнительные модели простой линейной регрессии:

Y от X1 (только стоимость покупок)

Y от X2 (только количество товаров)

14. Сравните три модели (две простые и одну множественную) по коэффициенту детерминации $R^2$. Какая модель лучше объясняет вариацию времени обслуживания? Имеет ли смысл использовать оба фактора одновременно?

**Часть 7. Выводы**

15. Подготовьте отчет для руководства супермаркета, в котором должны быть отражены:

Вывод о том, какой фактор (стоимость или количество) вносит больший вклад в увеличение времени обслуживания.

Рекомендуемая модель для прогнозирования времени обслуживания.

Прогноз времени для покупателя с заданными характеристиками.

Предложения по использованию модели для составления графиков работы кассиров.

## Лабораторная работа № 3
## Прогнозирование объема продаж автозапчастей в регионах

**Цель работы:** Разработать модель множественной регрессии для прогнозирования годового объема продаж автомобильных запчастей на основе инфраструктурных и социально-экономических показателей региона.

**Контекст:** Федеральная сеть магазинов автозапчастей "АвтоДеталь" планирует расширение и открытие новых точек в регионах. Для принятия решения об инвестициях необходимо построить модель, позволяющую прогнозировать потенциальный годовой объем продаж в зависимости от характеристик региона. В распоряжении аналитиков имеются данные по 11 регионам.

**Исходные данные:**

| Регион | Годовой объем продаж (Y), млн ден. ед. | Количество пунктов обслуживания (X1) | Количество зарегистрированных автомобилей (X2), млн шт. | Общий доход населения (X3), млрд ден. ед. |
|---|---|---|---|---|
| 1 | 52,3 | 2011 | 24,6 | 98,5 |
| 2 | 26,0 | 2850 | 22,1 | 31,1 |
| 3 | 20,2 | 650 | 7,9 | 34,8 |
| 4 | 16,0 | 480 | 12,5 | 32,7 |
| 5 | 30,0 | 1694 | 9,0 | 68,8 |
| 6 | 46,2 | 2302 | 11,5 | 94,7 |
| 7 | 35,0 | 2214 | 20,5 | 67,6 |
| 8 | 3,5 | 125 | 4,1 | 19,7 |
| 9 | 33,1 | 1840 | 8,9 | 67,9 |
| 10 | 25,2 | 1233 | 6,1 | 61,4 |
| 11 | 38,2 | 1699 | 9,5 | 75,6 |

**Задания:**

**Часть 1. Исследование исходных данных**

1. Определите результативный признак (Y) и факторные признаки (X1, X2, X3). Какие из факторов, по вашему мнению, должны оказывать положительное влияние на объем продаж, а какие – отрицательное?

2. Рассчитайте основные описательные статистики для всех переменных. Есть ли существенный разброс значений по регионам?

**Часть 2. Корреляционный анализ**

3. Постройте и проанализируйте матрицу парных коэффициентов корреляции.

4. Определите:

С каким фактором объем продаж связан наиболее тесно?

Есть ли проблема мультиколлинеарности? Обратите особое внимание на взаимосвязи между факторами X1, X2 и X3.

Если мультиколлинеарность присутствует, какие факторы ее вызывают?

**Часть 3. Построение полной регрессионной модели**

5. Постройте модель множественной линейной регрессии, включив все три фактора: $Y = b_0 + b_1 \cdot X_1 + b_2 \cdot X_2 + b_3 \cdot X_3$.

6. Интерпретируйте полученные коэффициенты регрессии. Соответствуют ли знаки коэффициентов вашим ожиданиям из п.1?

**Часть 4. Реализация в Excel**

7. Используя надстройку «Анализ данных» → «Регрессия», выполните расчеты для полной модели (с тремя факторами).

8. В полученной таблице найдите и проанализируйте:

Коэффициент детерминации $R^2$ (множественный $R^2$). Какова доля объясненной вариации объема продаж?

Значимость модели (F-статистика). Можно ли использовать модель для прогнозирования?

p-значения для каждого коэффициента. Все ли факторы являются статистически значимыми на уровне 5%? Если нет, укажите незначимые факторы.

9. Запишите полученное уравнение регрессии.

**Часть 5. Реализация в R**

10. Напишите скрипт на R, который:

Загружает данные.

Строит полную модель с тремя предикторами с помощью `lm()`.

Выводит сводку `summary()`.

11. Используя функцию `vif()` из библиотеки `car`, рассчитайте коэффициент инфляции дисперсии (VIF) для каждого фактора. Сделайте вывод о наличии мультиколлинеарности. Для каких факторов VIF > 5 или VIF > 10?

12. Выполните пошаговый отбор факторов (пошаговая регрессия) с использованием функции `step()` для автоматического выбора

наилучшей модели на основе информационного критерия Акаике (AIC). Какая модель была выбрана? Какие факторы в нее вошли?

13. Для итоговой модели (после отбора) постройте диагностические графики (остатки vs предсказанные, Q-Q plot, масштабно-локационный график). Проверьте выполнение предпосылок МНК.

**Часть 6. Прогнозирование для нового региона**

14. Компания рассматривает возможность выхода в регион №12 со следующими характеристиками:

Количество пунктов обслуживания (X1): 2500

Количество зарегистрированных автомобилей (X2): 20,2 млн шт.

Общий доход населения (X3): 40,0 млрд ден. ед.

Используя итоговую модель, полученную в R после отбора факторов, спрогнозируйте ожидаемый годовой объем продаж для этого региона.

15. Постройте доверительный интервал для среднего значения прогноза (с уровнем доверия 95%). Интерпретируйте результат.

**Часть 7. Сравнение моделей и выбор оптимальной**

16. Сравните две модели: полную (три фактора) и итоговую (после отбора) по следующим критериям:

Скорректированный коэффициент детерминации (Adjusted $R^2$).

Информационный критерий Акаике (AIC) – чем меньше, тем лучше.

Стандартная ошибка регрессии.

17. Какая модель является предпочтительной для прогнозирования и почему?

**Часть 8. Выводы и рекомендации**

18. Подготовьте аналитическую записку для отдела стратегического планирования сети "АвтоДеталь", которая должна содержать:

Краткое описание методики анализа.

Итоговую регрессионную модель (уравнение) для прогнозирования продаж.

Интерпретацию влияния каждого фактора, оставшегося в модели.

Прогноз для региона №12 с указанием точности прогноза (доверительный интервал).

Рекомендацию: стоит ли компании выходить в регион №12, если для безубыточности необходим годовой объем продаж не менее 30 млн ден. ед.?

*Примечание: Во всех работах при работе с R рекомендуется сохранять скрипты с расширением .R и включать в отчет как полученные численные результаты, так и графики. При работе с Excel – сохранять файлы с листами расчетов и копировать выходные таблицы в отчет.*

# ЗАКЛЮЧЕНИЕ

В современном мире, где данные становятся стратегическим ресурсом, способность принимать обоснованные решения на основе количественного анализа выходит на первый план. Учебное пособие было направлено на формирование у читателя целостного представления об эконометрике как о ключевом инструменте анализа данных.

В ходе изучения материала обучающийся прошел путь от описания тенденций развития явлений во времени до выявления глубинных причинно-следственных связей между показателями.

В первой части пособия были подробно рассмотрены методы прогнозирования динамических рядов. Освоив метод экстраполяции, читатель научился выявлять тренды и переносить их на будущее, а применение методов экспоненциального сглаживания позволило освоить гибкий инструментарий, где больший вес придается наиболее актуальным данным. Это дало понимание того, как строить краткосрочные и среднесрочные прогнозы в условиях инерционности экономических процессов.

Вторая часть работы была посвящена оценке согласованности экспертных мнений. Знакомство с коэффициентом конкордации расширило аналитический арсенал читателя, позволив ему перейти от анализа объективных цифр к исследованию субъективных оценок, что особенно важно в маркетинговых исследованиях и стратегическом планировании.

Центральное место в пособии занял регрессионный анализ, являющийся основой современной эконометрики. На конкретных примерах читатель научился не просто строить уравнения регрессии, но и содержательно интерпретировать их коэффициенты. Критически важным навыком, освоенным в данном разделе, стала оценка качества модели через призму показателей: стандартной ошибки, коэффициента детерминации и таблицы ANOVA. Это позволило понять, что «модель – это упрощение реальности»: даже самая качественная регрессия объясняет лишь часть вариации изучаемого явления, оставляя место для неучтенных факторов. Освоение процедур проверки значимо-

сти коэффициентов и построения доверительных интервалов заложило основу для статистически грамотного прогнозирования.

Практическая ценность пособия заключается в его прикладном характере. Выполняя лабораторные работы, читатель на реальных данных учится решать типичные задачи бизнеса: оценивать эффективность рекламы (лабораторная работа № 1), планировать производство (№ 2) и оптимизировать затраты (№ 3). Полученные навыки работы в Excel и среде статистического программирования R дают готовый инструментарий, востребованный в любой аналитической деятельности.

Таким образом, после освоения материала пособия, у обучающихся должны быть сформированы не только теоретические знания об эконометрических методах, но и практические компетенции, необходимые для проведения самостоятельных прикладных исследований. Понимание того, что любая модель содержит долю неопределенности (стохастическую компоненту), и умение эту неопределенность оценивать – главный итог освоения данного курса. Полученные знания станут надежной основой для дальнейшего углубленного изучения методов машинного обучения и статистического моделирования.

## БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Агаларов, З. С. Эконометрика : учеб. для вузов / З. С. Агаларов, А. И. Орлов. – 3-е изд. – М. : Дашков и К, 2024. – 380 с. – ISBN 978-5-394-05570-6.

2. Айвазян, С. А. Эконометрика-2: продвинутый курс с приложениями в финансах : учебник / С. А. Айвазян, Д. Фантаццини. – М. : Магистр : Инфра-М, 2024. – 944 с. – ISBN 978-5-9776-0333-1.

3. Вербик, М. Путеводитель по современной эконометрике / М. Вербик. – М. : ИД ВШЭ, 2025. – 672 с. – (Переводные учебники ВШЭ). – ISBN 978-5-7598-2724-5.

4. Галочкин, В. Т. Эконометрика : учеб. и практикум для вузов / В. Т. Галочкин. – М. : Юрайт, 2025. – 293 с. – (Высшее образование). – ISBN 978-5-534-14974-6.

5. Кацко, И. А. Эконометрика (продвинутый уровень) : учеб. пособие для вузов / И. А. Кацко [и др.]. – 2-е изд., стер. – СПб. : Лань, 2024. – 176 с. – ISBN 978-5-507-48946-6.

6. Ковалев, В. В. Дисперсионный анализ : учеб. пособие / В. В. Ковалев. – М. : Юрайт, 2024. – 180 с. – (Высшее образование). – ISBN 978-5-534-18392-4.

7. Кремер, Н. Ш. Эконометрика : учеб. и практикум для вузов / Н. Ш. Кремер, Б. А. Путко. – 4-е изд., испр. и доп. – М. : Юрайт, 2024. – 308 с. – (Высшее образование). – ISBN 978-5-534-08710-9.

8. Сурина, Е. Е. Методы анализа экономической информации и данных : учеб. пособие / Е. Е. Сурина. – 4-е изд., стер. – М. : ФЛИНТА, 2025. – 130 с. – ISBN 978-5-9765-2499-6.

9. Тимофеев, В. С. Эконометрика : учеб. для академ. бакалавриата / В. С. Тимофеев, А. В. Фаддеенков, В. Ю. Копылов. – 3-е изд., перераб. и доп. – Новосибирск : НГТУ, 2023. – 348 с. – ISBN 978-5-7782-4678-9.

10. Шанченко, Н. И. Эконометрика: компьютерный практикум в среде R Studio : учеб. пособие / Н. И. Шанченко. – Ульяновск : УлГТУ, 2024. – 215 с. – ISBN 978-5-9795-2245-6

# ПРИЛОЖЕНИЕ

Таблица П1

Исходные данные для выполнения лабораторной работы №1

**Вариант 1. Динамика добычи нефти**

| Условный год | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Добыча нефти, тыс. т | 273,1 | 295,2 | 312,6 | 365,8 | 389,9 | 403,4 | 408,6 | 397,8 | 367,9 |

**Вариант 2. Динамика объема бурения**

| Условный год | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Объем бурения, млн. м | 6,7 | 8,3 | 14,3 | 18,0 | 22,7 | 26,0 | 28,8 | 28,2 | 25,2 |

**Вариант 3. Динамика скорости бурения**

| Условный год | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Скорость бурения, м/ст-мес. | 2951 | 3763 | 2746 | 3510 | 3977 | 4610 | 5205 | 5280 | 4884 |

**Вариант 4. Динамика средней глубины скважин**

| Условный год | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Средняя глубина скважин, тыс. м | 2,05 | 2,25 | 2,40 | 2,48 | 2,52 | 2,53 | 2,55 | 2,55 | 2,56 |

**Вариант 5. Динамика коэффициента эксплуатации нефтяных скважин**

| Условный год | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Коэффициент эксплуатации, доли | 0,940 | 0,945 | 0,952 | 0,928 | 0,946 | 0,944 | 0,945 | 0,944 | 0,942 |

**Вариант 6. Динамика действующего фонда газовых скважин**

| Условный год | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Действующий фонд газовых скважин, скв. | 2 | 8 | 15 | 38 | 30 | 18 | 10 | 5 | 8 |

**Вариант 7. Динамика ввода в эксплуатацию новых скважин**

| Условный год | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Ввод новых скважин, скв. | 1 | 2 | 8 | 14 | 22 | 18 | 10 | 4 | 2 |

**Вариант 8. Динамика добычи газа**

| Условный год | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Добыча газа, млрд. м³ | 20,3 | 22,0 | 23,6 | 25,2 | 28,8 | 28,9 | 30,8 | 31,9 | 32,5 |

**Вариант 9. Динамика проходки на одну буровую бригаду**

| Условный год | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Проходка на 1 бригаду, тыс. м | 34,4 | 53,0 | 43,2 | 48,9 | 54,3 | 57,2 | 62,7 | 62,4 | 58,8 |

**Вариант 10. Динамика среднесуточного дебита скважин**

| Условный год | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Среднесуточный дебит, т | 105,8 | 125,4 | 94,3 | 43,9 | 36,6 | 31,2 | 26,2 | 22,2 | 18,7 |

**Вариант 11. Динамика объема перекачки газа**

| Условный год | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Объем перекачки газа, млрд. м³ | 256 | 282 | 301 | 322 | 345 | 366 | 392 | 424 | 453 |

**Вариант 12. Динамика средней дальности транспортировки нефти**

| Условный год | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Средняя дальность, тыс. км | 1,24 | 1,34 | 1,46 | 1,60 | 1,70 | 1,85 | 1,97 | 2,07 | 2,16 |

**Вариант 13. Динамика протяженности магистральных газопроводов**

| Условный год | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Протяженность, тыс. км | 99 | 103 | 111 | 117 | 125 | 132 | 136 | 144 | 149 |

**Вариант 14. Динамика коммерческой скорости бурения**

| Условный год | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Коммерческая скорость, м/ст.-мес. | 2951 | 3763 | 2746 | 3510 | 3977 | 4610 | 5205 | 5280 | 4684 |

**Вариант 15. Динамика объема реализации нефтепродуктов**

| Условный год | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Объем реализации, тыс. т | 50 | 53 | 56 | 57 | 57 | 58 | 63 | 65 | 63 |

Таблица П2

Квантиль распределения Стьюдента

| k \ p | 0,900 | 0,950 | 0,975 | 0,990 | 0,995 |
|---|---|---|---|---|---|
| 1 | 3,078 | 6,314 | 12,706 | 31,821 | 63,657 |
| 2 | 1,886 | 2,920 | 4,303 | 6,965 | 9.925 |
| 3 | 1,638 | 2,353 | 3,182 | 4,541 | 5,841 |
| 4 | 1,533 | 2,132 | 2,776 | 3,747 | 4,604 |
| 5 | 1,476 | 2,015 | 2,571 | 3,365 | 4,032 |
| 6 | 1,440 | 1,943 | 2,447 | 3,143 | 3,707 |
| 7 | 1,415 | 1,895 | 2,365 | 2,998 | 3,499 |
| 8 | 1,397 | 1,860 | 2,306 | 2,896 | 3,355 |
| 9 | 1,383 | 1,833 | 2,262 | 2,821 | 3,250 |
| 10 | 1,372 | 1,812 | 2,228 | 2,764 | 3,169 |
| 11 | 1,363 | 1,796 | 2,201 | 2,718 | 3,106 |
| 12 | 1,356 | 1,782 | 2,179 | 2,681 | 3,055 |
| 13 | 1,350 | 1,771 | 2,160 | 2,650 | 3,012 |
| 14 | 1,345 | 1,761 | 2,145 | 2,624 | 2,977 |
| 15 | 1,341 | 1,753 | 2,131 | 2,602 | 2,947 |
| 16 | 1,337 | 1,746 | 2,120 | 2,583 | 2,921 |
| 17 | 1,333 | 1,740 | 2,110 | 2,567 | 2,898 |
| 18 | 1,330 | 1,734 | 2,101 | 2,552 | 2,878 |
| 19 | 1,328 | 1,729 | 2,093 | 2,539 | 2,861 |
| 20 | 1,325 | 1,725 | 2,086 | 2,528 | 2,845 |
| 25 | 1,316 | 1,708 | 2,060 | 2,485 | 2,787 |
| 30 | 1,310 | 1,697 | 2,042 | 2,457 | 2,750 |
| 35 | 1,306 | 1,690 | 2,030 | 2,438 | 2,724 |
| 40 | 1,303 | 1,684 | 2,021 | 2,423 | 2,704 |
| 45 | 1,301 | 1,679 | 2,014 | 2,412 | 2,690 |
| 50 | 1,299 | 1,676 | 2,009 | 2,403 | 2,678 |
| 55 | 1,297 | 1,673 | 2,004 | 2,396 | 2,670 |
| 60 | 1,296 | 1,671 | 2,000 | 2,390 | 2,660 |
| 70 | 1,294 | 1,667 | 1,994 | 2,381 | 2,648 |
| 80 | 1,292 | 1,664 | 1,990 | 2,374 | 2,639 |
| 90 | 1,291 | 1,662 | 1,987 | 2,368 | 2,632 |
| 100 | 1,290 | 1,660 | 1,984 | 2,364 | 2,626 |

*Учебное электронное издание*

ГУБЕРНАТОРОВ Алексей Михайлович

АНАЛИЗ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ
ЭКОНОМЕТРИЧЕСКИХ МЕТОДОВ

Учебное пособие

*Издается в авторской редакции*

***Системные требования:*** Intel от 1,3 ГГц; Windows XP/7/8/10; Adobe Reader; дисковод CD-ROM.

**Тираж 9 экз.**

Издательство Владимирского государственного университета
имени Александра Григорьевича и Николая Григорьевича Столетовых.
600000, Владимир, ул. Горького, 87.