1 Регрессионный анализ

Предварительные сведения по регрессионному анализу

Имеется n значений независимых переменных $X_1, X_2,\ldots,X_p$ и зависимой переменной Y (рисунок 1).

Y	X_1	X_2	…	X_p
y_1	x_{11}	x_{12}	…	x_{1p}
y_2	x_{21}	x_{22}	…	x_{2p}
…	…	…	…	…
y_n	x_{n1}	x_{n2}	…	x_{np}

Рисунок 1 – Исходных данных для проведения регрессионного анализа

Наиболее часто встречаются следующие регрессионные зависимости:

1) линейная $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$;

2) степенная $Y = \beta_0\, X_1^{\beta_1} X_2^{\beta_2} \cdot \ldots \cdot X_p^{\beta_p} \varepsilon$;

3) экспоненциальная $Y = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon}$;

4) параболическая $Y = \beta_0 + \beta_1 X_1^2 + \beta_2 X_2^2 + \ldots + \beta_p X_p^2 + \varepsilon$;

5) гиперболическая $Y = \beta_0 + \dfrac{\beta_1}{X_1} + \dfrac{\beta_2}{X_2} + \ldots + \dfrac{\beta_p}{X_p} + \varepsilon$.

Пример 1: предположим, что зависимость расходов на продукты питания по совокупности семей характеризуется следующим уравнением:

$$\hat{y} = 0.5 + 0.35 x_1 + 0.73 x_2,$$

где y – расходы семьи за месяц на продукты питания, тыс. руб.;

x_1 – месячный доход на одного члена семьи, тыс. руб.;

x_2 – размер семьи, человек.

Из уравнения можно сделать вывод о том, что при увеличении только месячного дохода на 1 тыс. руб., расходы на продукты питания в среднем увеличатся на 350 руб., а при увеличении размера семьи на 1 человека расходы возрастут в среднем на 730 рублей.

Пример 2: предположим, что при исследовании спроса на мясо получено уравнение

$$\hat{y} = 0.82 \cdot x_1^{-2.63} x_2^{1.11},$$

где y – количество спрашиваемого мяса; x_1 – цена; x_2 – доход. Здесь коэффициенты β_1 и β_2 – эластичности. Следовательно, рост цен на 1% при том же доходе вызывает снижение спроса в среднем на 2,63%. Увеличение дохода на 1% обусловливает при неизменных ценах рост спроса на 1,11%.

Основное значение имеют линейные модели (относительно параметров регрессии) в силу своей простоты. Нелинейные формы зависимости часто преобразуются к линейным путем линеаризации.

Наилучшие оценки параметров

$\beta_0, \beta_1, \beta_2, \ldots, \beta_p$: $b_0, b_1, b_2, \ldots, b_p$ определяются методом наименьших квадратов (МНК). Оценки коэффициенты регрессии находятся по критерию:

$$F = \sum_{i=1}^{n} \left[y_i - \varphi\left(x_{i1}, x_{i2}, \ldots, x_{ip}\right) \right]^2 \to \min, \qquad (1)$$

где y_i $(i = 1, 2, \ldots, n)$ – значения результативного фактора (зависимой переменной) Y в i-м наблюдении; $x_{i1}, x_{i2}, \ldots, x_{ip}$ – i-е наблюдения независимых переменных $X_1, X_2, \ldots, X_p$; n – количество наблюдений.

Реализация этого критерия приводит к системе уравнений

$$\frac{\partial F}{\partial b_j} = 0, \quad j = 0, 1, 2, \ldots, p, \qquad (2)$$

из которого определяют значения параметров $b_0, b_1, b_2, \ldots, b_p$.

Пример проведения регрессионного анализа с помощью R и RStudio

Построение линейной модели множественной регрессии и ее анализ проведем в **R и RStudio** для следующих исходных данных.

По условным данным за два года (см. таблица 1) изучается зависимость оборота розничной торговли (Y, млрд. руб.) от ряда факторов: X_1 – денежные доходы населения, млрд. руб.; X_2 – доля доходов, используемая на покупку товаров и оплату услуг, млрд. руб.; X_3 – уровень инфляции за последний год, %; X_4 – официальный курс рубля по отношению к доллару США (данные условные).

Таблица 1 – Исходные данные

Месяц	**Y**	**X_1**	**X_2**	**X_3**	**X_4**
1	76,4	117,7	81,6	10,3	28,5
2	77,6	123,8	73,2	11,4	28,7
3	88,2	126,9	75,3	12,2	29,1
4	87,3	134,1	71,3	11,5	29,2
5	82,5	123,1	77,3	11,2	29,1
6	79,4	126,7	76,1	10,5	29,2
7	80,3	130,4	76,6	9,4	29,3
8	80,1	129,3	84,7	9,5	29,2
9	105,2	145,4	92,4	9,3	29,1
10	102,5	163,8	80,3	9,2	28,7
11	108,7	164,8	82,6	9,4	28,4
12	104,5	165,3	70,9	9,7	27,8
13	103,7	164,1	89,9	8,2	27,7
14	117,8	183,7	81,3	8,4	27,6
15	115,8	195,8	83,7	8,2	27,5
16	117,8	219,4	76,1	8,1	27,5
17	118,4	209,8	80,4	7,8	27,4
18	120,4	223,3	78,1	7,2	27,5
19	123,8	223,6	79,8	8,2	27,6
20	134,9	236,6	82,1	7,5	27,7
21	130,5	236,6	83,2	7,4	27,8
22	140,7	248,6	80,8	7,3	28,7
23	150,4	253,4	81,8	7,4	28,3
24	172,7	254,3	87,5	7,5	28,1

Требуется:

1. Для заданного набора данных построить линейную модель множественной регрессии.
2. Оценить адекватность и значимость построенного уравнения регрессии.

3. Выделить значимые и незначимые факторы в модели.
4. Построить уравнение регрессии со статистически значимыми факторами. Дать экономическую интерпретацию параметров модели.
5. Определить расчетные значения зависимой переменной *Y* и построить графики фактических значений зависимой переменной и расчетных. По оси абсцисс откладываются номера наблюдений.
6. Определить оборот розничной торговли для двух вариантов значений независимых переменных:
   a. Для одного набора значений переменных X_1=130, X_2=90, X_3=10
   b. Для таблицы значений переменных

X_1	X_2	X_3
100	200	12
150	190	9
120	175	10

7. Оценить гетероскедастичность дисперсии остатков с помощью тестов Гольдфельда– Квандта и Бреуша-Пагана. Найти робастные стандартные ошибки параметров модели регрессии.
8. Определить наличие автокорреляции остатков с помощью теста Дарбина-Уотсона.

Решение: Выполним задания 1–6.

В R для построения модели линейной регрессии можно воспользоваться функцией lm() из библиотеки stats, которая по умолчанию подключена в среде RStudio (аргументы функции приведены в таблице 2).

```
model <- lm(formula, data, na.action)
```

Таблица 2 – Аргументы функции lm()

<table>
<tr>
<td rowspan="2">formula</td>
<td colspan="2">y~x_1+x_2+x_3
Здесь y – это зависимая переменная; x_1, x_2, x_3 – независимые переменные.
Если по умолчанию в модель линейной регрессии должны войти все предикторы, можно записать y~.
Знак “+” используется для разделения предикторов, знак “-“ используется для исключения предикторов из модели, если нужно исключить свободный член, то записывают “-1”
Кроме знака “+” между переменными можно поставить следующие символы:</td>
</tr>
<tr>
<td>:</td>
<td>Обозначает взаимодействие между независимыми переменными. Предсказание значений y по значениям x, z и взаимодействия между x и z будет закодировано как $y \sim x + z + x{:}z$</td>
</tr>
<tr>
<td rowspan="3"></td>
<td>*</td>
<td>Краткое обозначение для всех возможных взаимодействий. Код $y \sim x * z * w$ в полном виде означает
$y \sim x + z + w + x{:}z + x{:}w + z{:}w + x{:}z{:}w$</td>
</tr>
<tr>
<td>^</td>
<td>Обозначает взаимодействия до определенного порядка. Код $y \sim (x + z + w)$^2 в полном виде будет записан как
$y \sim x + z + w + x{:}z + x{:}w + z{:}w$</td>
</tr>
<tr>
<td>I()</td>
<td>Элемент в скобках интерпретируется как арифметическое выражение. Например,
$y \sim x + (z + w)$^2
означает $y \sim x + z + w + z{:}w$.
Для сравнения $y \sim x$ + I(($z + w$)^2) означает
$y \sim x + h$, где h – это новая переменная, полученная при возведении в квадрат суммы z и w</td>
</tr>
</table>

	function	В формулах можно использовать математические функции. Например, $\log(y) \sim x + z + w$ будет предсказывать значения $\log(y)$ по значениям x, z и w
data	Таблица с исходными данными	
na.action	Действие в случае наличия в данных NA. По умолчанию такие наблюдения игнорируются	

В таблице 3 перечислены функции, которые используются совместно с результатом функции lm()

Таблица 3 – Функции, которые можно использовать с аргументом model<-lm()

Функция	Действие
summary()	Показывает детальную информацию о подогнанной модели
coefficients()	Перечисляет параметры модели (свободный член и регрессионные коэффициенты)
confint()	Вычисляет доверительные интервалы для параметров модели (по умолчанию доверительная вероятность 95%)
fitted()	Выводит на экран предсказанные значения, согласно подогнанной модели
residuals()	Показывает остатки для подогнанной модели
anova()	Создает таблицу ANOVA (дисперсионного анализа) для подогнанной модели или таблицу ANOVA, сравнивающую две или более моделей
vcov()	Выводит ковариационную матрицу для параметров модели. По главной диагонали этой матрицы расположены выборочные дисперсии оцененных параметров модели, корни квадратные из которых дают значения стандартных

	ошибок для параметров уравнения регрессии.
AIC()	Вычисляет информационный критерий Акаике (Akaike’s Information Criterion)
plot()	Создает диагностические диаграммы для оценки адекватности модели
predict()	Использует подогнанную модель для предсказания зависимой переменной для нового набора данных

Для выполнения **1-го пункта** задания, необходимо сохранить файл с исходной таблицей как .csv файл. Для этого в Excel нужно сохранить файл как .csv (разделитель – запятые). Пусть, например, этот файл сохранен под именем «**К ПЗ множ регр .csv**» в папке R/examples. Затем в R-studio можно сделать эту папку текущей. Для этого нужно выбрать в меню Session/Set Working Directory команду Choose Directory… и указать путь к папке, в которой будут храниться файлы. Теперь этот файл нужно выбрать. Можно воспользоваться вкладкой Environment, выбрать Import Dataset/From Text (base)….(см. рисунок 1). Это равносильно выполнению команды

```
df <- read.csv2("К ПЗ множ регр.csv")
View(df)
```

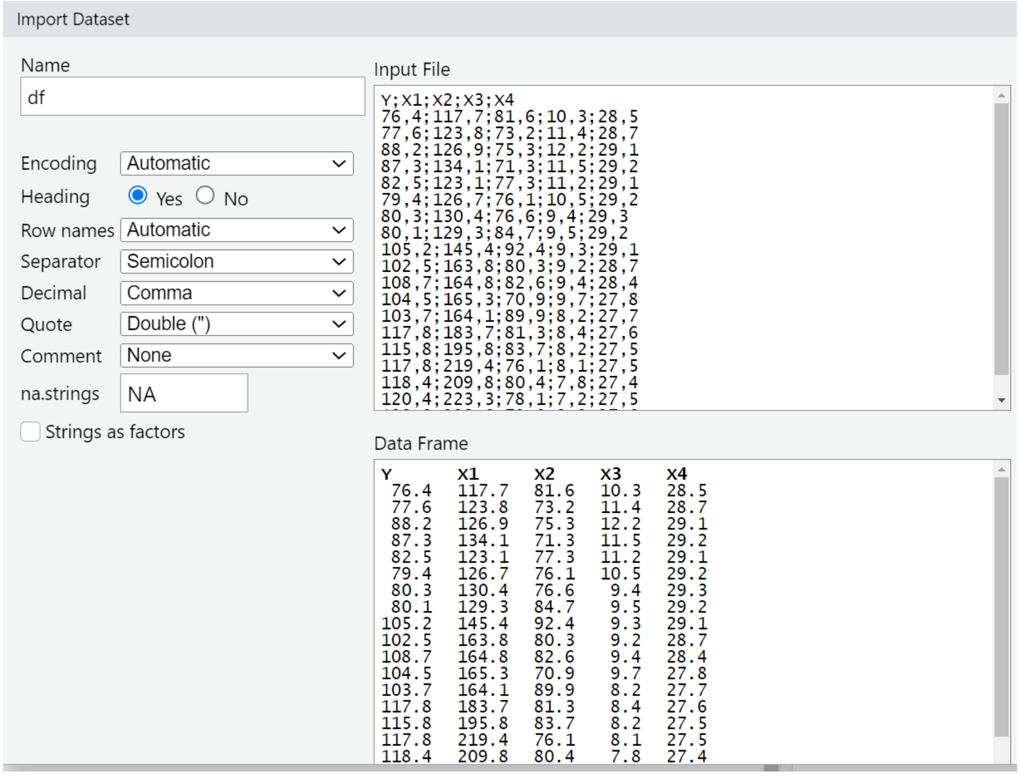
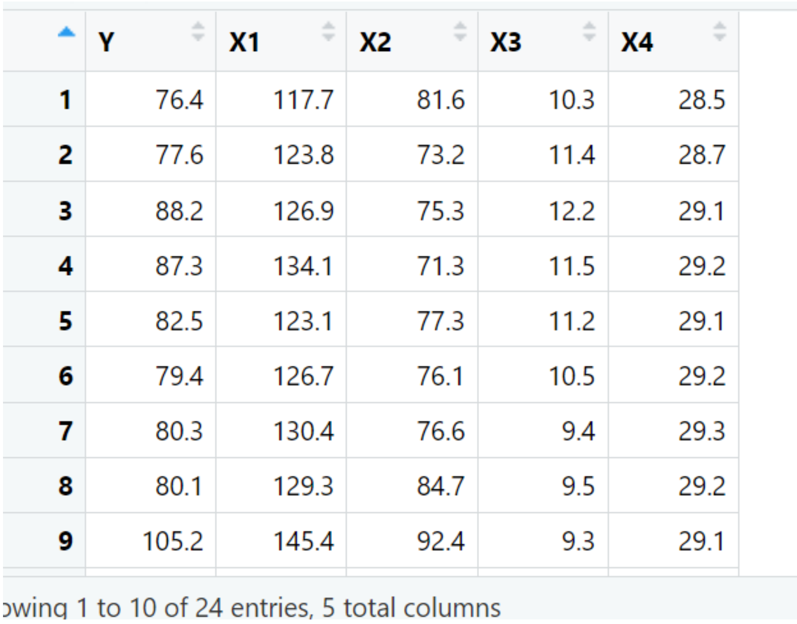


Рисунок 1 – Выбор файла с исходными данными

В RStudio появится вкладка с этой таблицей (рисунок 2).



	Y	X1	X2	X3	X4
1	76.4	117.7	81.6	10.3	28.5
2	77.6	123.8	73.2	11.4	28.7
3	88.2	126.9	75.3	12.2	29.1
4	87.3	134.1	71.3	11.5	29.2
5	82.5	123.1	77.3	11.2	29.1
6	79.4	126.7	76.1	10.5	29.2
7	80.3	130.4	76.6	9.4	29.3
8	80.1	129.3	84.7	9.5	29.2
9	105.2	145.4	92.4	9.3	29.1

ɔwing 1 to 10 of 24 entries, 5 total columns

Рисунок 2 – Вид файла с исходными данными в RStudio

Далее вызываем функцию lm() и заполняем ее аргументы, как показано на рисунке 3. Формулу в функции lm() можно было записать как Y~.

```
#Загружаем таблицу
df <- read.csv2("К ПЗ множ регр.csv")
# Показать таблицу
View(df)

#Расчет параметров уравнения регрессии с помощью lm()
model<-lm(data=df, Y~X1+X2+X3+X4)
summary(model)
```

Рисунок 3 – Построение модели множественной регрессии в R-studio

Нажимаем Ctrl+Enter для выполнения кода. Создается список с именем model, в котором содержатся вычисленные коэффициенты и другая информация по построенной модели. С помощью функции summary() можно вывести детальную информацию о подогнанной модели.

Результаты вычислений представлены на рисунке 4.

```
4:1    (Top Level) ⬍
Console  Terminal ×  Jobs ×
R 4.1.0 · ~/

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4, data = LR1)

Residuals:
   Min     1Q Median     3Q    Max 
-8.495 -3.019 -1.616  1.562 15.571 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -203.55241   82.87974  -2.456  0.02385 *  
X1             0.66553    0.06578  10.117 4.36e-09 ***
X2             1.23927    0.33098   3.744  0.00137 ** 
X3             6.98020    2.52993   2.759  0.01249 *  
X4             1.09074    2.84643   0.383  0.70583    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.392 on 19 degrees of freedom
Multiple R-squared:  0.9477,	Adjusted R-squared:  0.9367 
F-statistic: 86.08 on 4 and 19 DF,  p-value: 6.693e-12

> |
```

Рисунок 4 – Результаты регрессионного анализа

Поясним полученные результаты на рисунке 4.

В таблице *Residuals* приведены квартили остатков модели (разницы между фактическими значениями переменной Y и рассчитанными с помощью построенной модели значениями $\hat{Y}$), минимальный остаток равен -

8,495, 25% значений меньше -3,019, 50% меньше -1,616, 75% меньше 1,562, максимальный остаток равен 15,571.

В таблице Coefficients в столбце Estimate приведены значения оцененных параметров модели, по ним можно выписать уравнение множественной регрессии:

$$\hat{Y} = -203{,}55 + 0{,}67X_1 + 1.24X_2 + 6.98X_3 + 1.09X_4. \qquad (3)$$

В столбце Std.Error указаны стандартные ошибки параметров модели (их среднеквадратические отклонения), в столбце t value – расчетные значения t-критерия Стъюдента для оценки значимости отдельных параметров модели.

Последний столбец таблицы содержит для каждого параметра P-значение: если P-значение меньше заданного уровня значимости, то нулевая гипотеза $H_0 : b_j = 0$ отклоняется, то есть коэффициент b_j значимо отличается от нуля (с достаточно большой вероятностью).

Для нашего примера, если принять уровень значимости 0,05 (это вероятность того, что ошибочно отвергается нулевая гипотеза о том, что на самом деле коэффициент b_j равен 0), то значимыми оказываются параметры при всех переменных, кроме X_4. Переменную X_4 следует исключить из модели, так как скорее всего она не влияет на Y.

Multiple R-squared – множественный *R-квадрат* – это *коэффициент детерминации*, который показывает, какая часть (доля) вариации (изменчивости) объясняемой переменной Y обусловлена вариацией объясняющих переменных ($0 \le R^2 \le 1$). Чем ближе R^2 к единице, тем лучше уравнение регрессии аппроксимирует эмпирические данные.

В нашем случае $R^2 \approx 0{,}948$, что свидетельствует о том, что изменения зависимой переменной Y(оборот розничной торговли) в основном (на 94,8%) можно объяснить изменениями включенных в модель объясняющих переменных – X_1, X_2, X_3, X_4. Такое значение свидетельствует об **адекватности** модели.

Adjasted R-squared – скорректированный коэффициент детерминации.

$$R^2_{Adj} = 1 - \left(1 - R^2\right)\frac{n-1}{n-p-1},$$

где n – число наблюдений, p – число объясняющих переменных. Недостатком коэффициента детерминации R^2 является то, что он увеличивается при добавлении новых объясняющих переменных, хотя это и не обязательно означает улучшение качества регрессионной модели. В этом смысле предпочтительнее использовать R^2_{Adj}. В отличие от R^2 скорректированный коэффициент R^2_{Adj} может уменьшаться при введении в модель новых объясняющих переменных, не оказывающих существенное влияние на зависимую переменную.

Residual standart error – *стандартная ошибка* регрессии $S_{ост} = \sqrt{S^2_{ост}}$,

где $S^2_{ост} = \sum_{i=1}^{n} \frac{e_i^2}{n-p-1} = \sum_{i=1}^{n} \frac{\left(y_i - \hat{y}_i\right)^2}{n-p-1}$ – необъясненная дисперсия (мера разброса значений зависимой переменной вокруг линии регрессии), *p* – число независимых переменных (в нашем случае 4), *n* – объем выборки (в нашем случае 24).

F-statistic – вычисленное значение критерия Фишера (F-статистики); С помощью него проверяется гипотеза о том, что одновременно все коэффициенты при независимых переменных равны 0 (говорят, что построенное уравнение регрессии не значимо). Уравнение регрессии считается значимым на уровне значимости α, если F больше табличного значения F-критерия, или если величина p-value меньше принятого уровня значимости α, например 0,05.

В нашем примере расчетное значение F- критерия Фишера равно 86,08. Значимость F = 6,693E-12=6,693*10^{-12}, что намного меньше 0,05. Таким образом, полученное уравнение в целом значимо.

Как было сказано ранее из уравнения регрессии следует исключить переменную X_4. Исключим несущественный фактор X_4 (официальный курс рубля по отношению к доллару США) и построим уравнение зависимости

$$Y = f(X_1, X_2, X_3).$$

Получим модель на рисунке 5. Объект, в котором будет храниться результат моделирования, назовем model_1.

```
Console   Terminal ×   Jobs ×
R 4.1.0 · ~/
> summary(model)

Call:
lm(formula = Y ~ X1 + X2 + X3, data = LR1)

Residuals:
   Min      1Q Median     3Q    Max 
-8.486 -3.132 -1.657  2.109 15.749 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -178.4539    49.6900  -3.591 0.001825 ** 
X1             0.6644     0.0643  10.333 1.82e-09 ***
X2             1.2721     0.3128   4.067 0.000602 ***
X3             7.3519     2.2862   3.216 0.004336 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.254 on 20 degrees of freedom
Multiple R-squared:  0.9473,    Adjusted R-squared:  0.9394 
F-statistic: 119.8 on 3 and 20 DF,  p-value: 5.966e-13

> |
```

Рисунок 5 – Результаты регрессионного анализа со статистически значимыми коэффициентами

Модель зависимости оборота розничной торговли запишется в следующем виде:

$$\hat{Y} = -178{,}45 + 0{,}66X_1 + 1.27X_2 + 7{,}35X_3. \qquad (4)$$

Оценим адекватность и значимость уравнения (4) по сравнению с уравнением (3).

Из таблицы (рисунок 5) видно, что значение коэффициента детерминации (R^2) осталось примерно на прежнем уровне, немного возросло значение скорректированного коэффициента детерминации (нормированный R^2), следовательно, можно сделать вывод об адекватности модели.

Стандартная ошибка регрессии для второго уравнения меньше, чем для первого (6,254<6,392). Расчетное значение F-критерия Фишера увеличилось примерно на 33. Значимость F=5,966E-13, что меньше 0,05, таким образом, полученное уравнение в целом значимо. Также можно сделать вывод о том, что все включенные в модель факторы являются значимыми, так как их *P*-значение < 0,05.

Экономическая интерпретация параметров модели.

Коэффициент b_1 = 0,66, означает, что при увеличении только денежных доходов населения (X_1) на 1 тыс. руб. оборот розничной торговли возрастет в среднем на 0,66 тыс. руб., а то что коэффициент b_2 = 1,27, означает, что увеличение только доли доходов, используемых на покупку товаров и оплату услуг на 1 тыс. руб., оборот розничной торговли возрастет в среднем на 1,27 тыс. руб. при условии неизменности других двух факторов, коэффициент b_3 = 7,35, означает, что при увеличении только уровня инфляции на 1%, оборот розничной торговли возрастет в среднем на 7,35 млрд. руб. при условии неизменности других двух факторов.

С помощью элемента списка model_1: fitted.values определим расчетные значения зависимой переменной Y: $\hat{Y}$ и добавим с помощью cbind() столбец со значениями $\hat{Y}$ к таблице df. При этом fitted – это имя столбца. Код на рисунке 6.

```
# Добавление в таблицу df столбца fitted с расчетными значениями Y
df<-cbind(df,fitted=model_1$fitted.values)
```

Рисунок 6

Далее построим график фактических и расчетных значений Y и $\hat{Y}$:

```
#Построение графиков фактических и расчетных значений Y
plot(1:24, df$Y, type="l", xlab="Номер наблюдения", ylab="Y")
lines(1:24,df$fitted, type="l", col=2)

#Добавление легенды
legend("bottomright", legend = c("Y", "fitted Y"), lwd=1, col = c(2, 1))
```

Рисунок 7

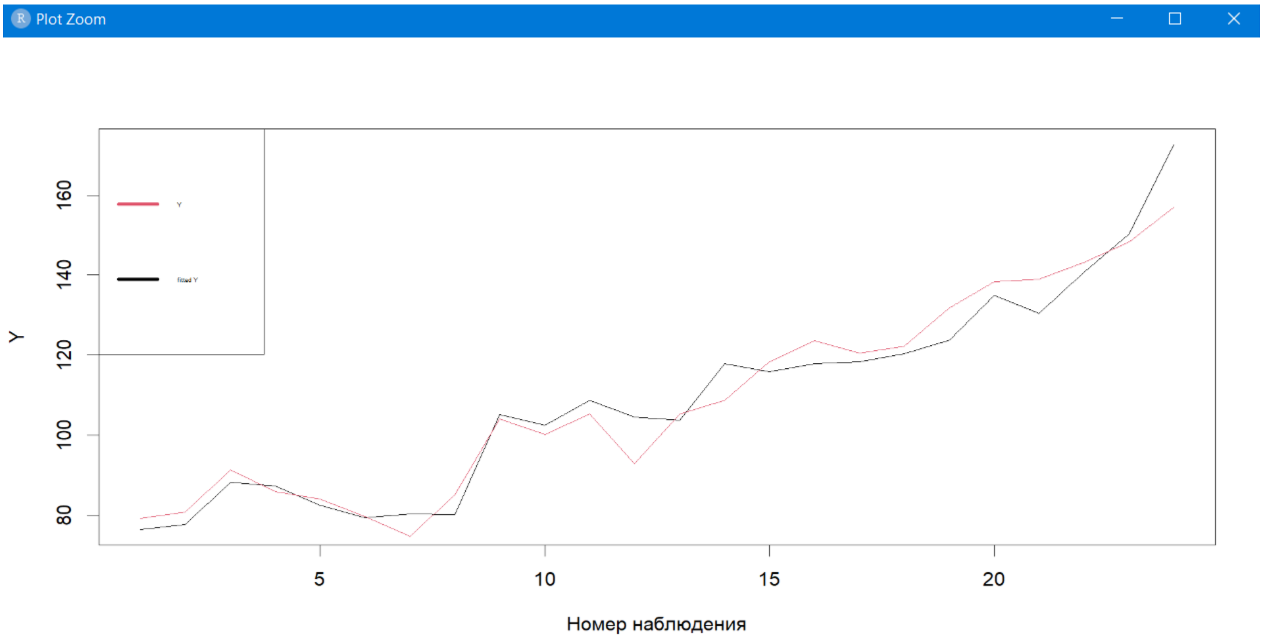


Рисунок 8

Или можно построить график по слоям с помощью пакета ggplot2 (см. рисунок 9)

```
# Построение графиков с помощью пакета ggplot2
library(ggplot2)
ggplot(NULL, aes(x,y))+geom_line(data=data.frame(x=1:24,y=df$fitted),
                                aes(color='расчетные значения Y'))+
  geom_line(data=data.frame(x=1:24, y=df$Y),
            aes(color="Фактические значения Y"))+
   labs(x = "Номер наблюдения", y = "Оборот розничной торговли")
```

Рисунок 9

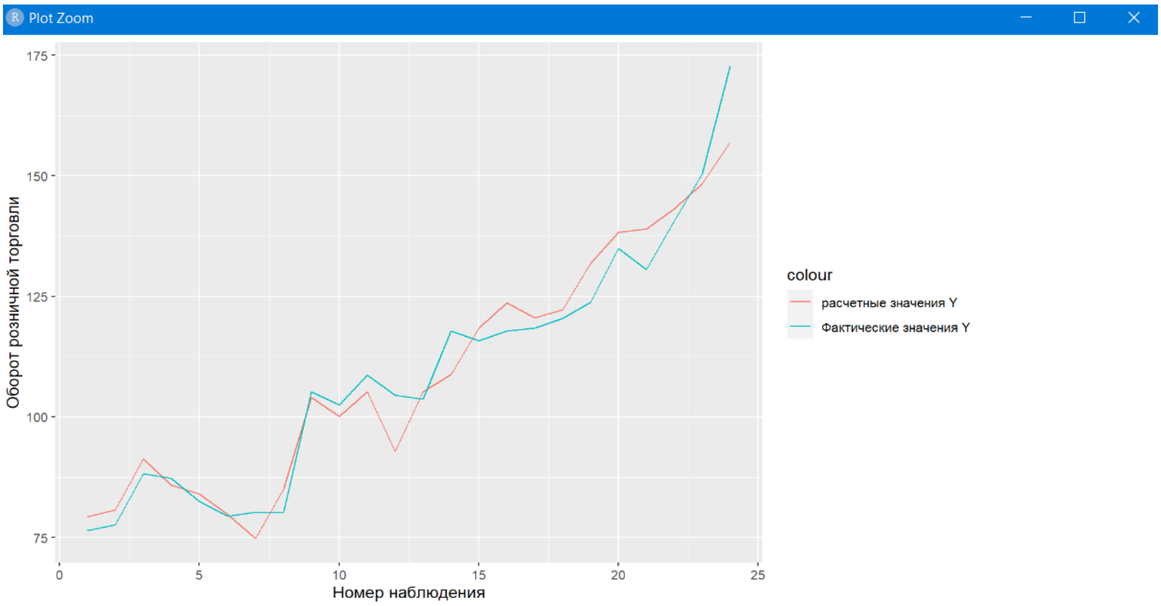


Рисунок 10

Для выполнения задания 6 воспользуемся функцией predict(), ее синтаксис

predict(object, newdata, type=”response”, interval),

где **object –** результат выполнения функции lm();

newdata – дата фрейм со значениями независимых переменных для прогноза;

type – тип прогноза;

interval – тип расчета интервала.

Для варианта задания набора значений переменных X_1=130, X_2=90, X_3=10, создадим дата фрейм

```
dp<-data.frame(X1=130, X2=90, X3=10)
```

и присвоим его аргументу newdata

```
predict(object=model_1, newdata=dp)
```

В результате Y=95.927 млрд.руб.

Для таблицы значений переменных

X_1	X_2	X_3
100	200	12
150	190	9
120	175	10

переменной dp присвоим дата фрейм вида (с помощью c() задают векторы)

```
dp<-data.frame(X1=c(100, 150, 120), X2=c(200, 190, 175), X3=c(12, 9, 10))
```

и присвоим его аргументу newdata

```
predict(object=model_1, newdata=dp)
```

В результате получим

```
> predict(model_1, newdata=dp)
       1        2        3 
230.6316 229.0744 197.4128 
```

Для выполнения заданий 7–8, потребуется дать следующие пояснения.

Регрессионный анализ включает не только построение самой регрессионной модели, но и исследование остатков $\varepsilon = Y - \hat{Y}$. При этом сами остатки ε_i следует рассматривать как случайные величины.

Каждый оцененный коэффициент регрессии является случайной величиной, свойства которой зависят от свойств остаточного члена ε в регрессионной модели.

Для того чтобы регрессионный анализ, основанный на методе наименьших квадратов, давал наилучшие и обоснованные результаты, необходимо выполнение некоторых предположений относительно поведения остатков ε_i.

Для регрессионных моделей, линейных относительно независимых переменных $X_1, X_2, \ldots, X_p$, должны удовлетворяться четыре **условия Гаусса-Маркова**:

1)Матрица X – детерминированная матрица, имеющая максимальный ранг, равный p+1 .

2) Математические ожидания возмущения $\varepsilon_i \quad (i = 1, 2, \ldots, n)$ равны нулю:

$$M(\varepsilon_i) = 0$$

или в матричном виде

$$M(\varepsilon) = 0_n$$ – нулевая матрица-столбец

3) Дисперсия возмущения ε_i постоянна для любого i (условие гомоскедастичности):

$$D(\varepsilon_i) = \sigma_\varepsilon^2$$

Свойство гомоскедастичности на практике проверяется на самом деле для остатков модели, а не для истинных ошибок и может выполнятся лишь приближенно. Если условие гомоскедастичности не выполнено (то есть дисперсия ошибок не постоянна), то говорят, что имеет место условие **гетероскедастичности**.

Свойство гомоскедастичности и гетероскедастичности можно проиллюстрировать рисунками:

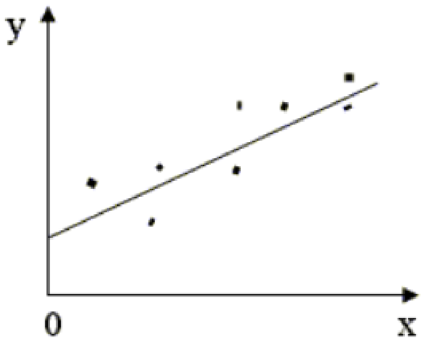


Рисунок 11. Гомоскедастичность остатков

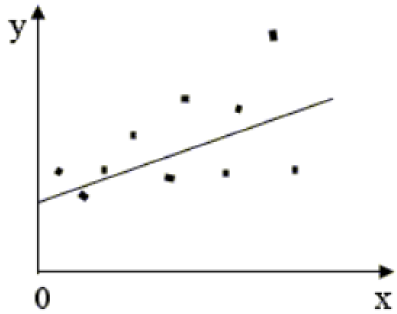


Рисунок 12. Гетероскедостичность остатков

4) Возмущения ε_i и ε_j не коррелированы (отсутствие автокорреляции остатков):

$$\operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0 \quad (i \neq j).$$

Если все эти условия выполняются, то полученные с помощью МНК оценки коэффициентов регрессии будут обладать свойствами хороших статистических оценок: несмещенностью[1], состоятельностью[2] и эффективностью[3].

Нарушение 3-го условия – гомоскедастичности остатков приводит к следующим последствиям:

1. МНК-оценки коэффициентов останутся **несмещенными**.
2. МНК-оценки коэффициентов больше **не являются эффективными**.
3. **Стандартные ошибки оценок** коэффициентов, рассчитанные по формуле для случая гомоскедастичности, оказываются **смещенными и несостоятельными**. Следовательно, их использование для **тестирования гипотез и построения доверительных интервалов** может привести к **некорректным выводам**.

Первые два перечисленных последствия говорят о том, что МНК-оценки коэффициентов в условиях гетероскедастичности хотя и теряют в точности, однако остаются в среднем правильными. Третье же последствие весьма критично, так как увеличивает вероятность неверной интерпретации результатов моделирования.

Для преодоления 3-го последствия можно посчитать стандартные ошибки МНК-коэффициентов уравнения регрессии по скорректированным формулам, в этом случае, стандартные ошибки станут состоятельными (их называют робастными стандартными ошибками). Для этого можно воспользоваться пакетом **sandwich** и функцией **vcovHC()**, с помощью которой можно найти робастные стандартные ошибки для параметров модели регрессии **model**. Аргумент **type** функции **vcovHC()** может принимать значения от HC0 до HC5, и отвечает за способ оценки

---

[1] Статистическая оценка некоторого параметра называется **несмещенной**, если ее математическое ожидание равно истинному значению этого параметра.

[2] Свойство **состоятельности** оценок заключается в том, что при неограниченном возрастании объема выборки, значение оценки должно стремиться (по вероятности) к истинному значению параметра, а дисперсии оценок должны уменьшаться и в пределе стремиться к нулю.

[3] Оценка называется **эффективной**, если она имеет минимальную дисперсию по сравнению с другими оценками заданного класса.

ковариационной матрицы параметров модели регрессии. Корни квадратные из **диагональных элементов матрицы**, полученной с помощью **vcovHC()** и есть робастные стандартные ошибки соответствующих коэффициентов модели, начиная с b_0.

Можно выполнить проверку значимости коэффициентов модели на основе робастных стандартных ошибок с помощью функции **coeftest() (**для этого надо загрузить библиотеку **lmtest)**

Код приведен ниже:

```
install.packages("sandwich")
library("sandwich")
install.packages("lmtest")
library("lmtest")
#Проверка значимости коэффициентов уравнения на основании
#робастных стандартных ошибок коэффициентов (функция из пакета lmtest)
coeftest(model_1, vcov = vcovHC(model_1,type="HC3"))
# Скорректированная ковариационная матрица коэффициентов модели
vcovHC(model_1, type="HC3")
#Корни квадратные из диагональных элементов скорректированной кавариационной матрицы
# (робастные стандартные ошибки)
v<-diag(vcovHC(model_1, type="HC3")^0.5)
```

Так же можно выполнить тесты на наличие гетероскедастичности, например, с помощью теста Гольдфельда-Квандта или Бреуша-Пагана.

```
# Тест Гольдфельда-Квандта на гомоскедастичность
install.packages("lmtest")
library ("lmtest")
gqtest(model_1, order.by=~X3, data=df, fraction=8)


#Тест Бреуша-Пагана
```

```
bptest(model_1)
```

Результаты теста Гольфельда-Квандта:

```
	Goldfeld-Quandt test

data:  model_1
GQ = 0.079181, df1 = 4, df2 = 4, p-value = 0.9846
alternative hypothesis: variance increases from segment 1 to 2
```

Вывод: наблюдается гомоскедастичность относительно переменной X_3.

Результат применения теста Бреуша-Пагана:

```
	studentized Breusch-Pagan test

data:  model_1
BP = 3.6928, df = 3, p-value = 0.2966
```

Так как p-value > 0.05, то принимаем нулевую гипотезу об отсутствии гетероскедастичности.

Нарушение 4-го условия – отсутствия автокорреляции остатков

Те же последствия, что и при нарушении 3-го условия.

Для выявления наличия автокорреляции остатков воспользуемся тестом Дарбина-Уотсона – это функция dwtest() из пакета lmtest

```
#Тест Дарбина-Уотсона на автокорреляцию остатков
dwtest (model_1)
```

Этот тест используется для обнаружения автокорреляции первого порядка, т.е. проверяется некоррелированность не любых, а только соседних величин ε_i. Соседними обычно считаются соседние во времени (при рассмотрении временных рядов) или по возрастанию объясняющей переменной X значения ε_i.

$$\varepsilon_i = \rho \cdot \varepsilon_{i-1} + v_i.$$

Гипотеза $H_0 : \rho = 0$ (автокорреляция отсутствует).

Результат

```
data:  model_1
DW = 1.3522, p-value = 0.01929
alternative hypothesis: true autocorrelation is greater than 0
```

Так как p-value < 0.05, то гипотезу о том, что коэффициент корреляции равен нулю следует отвергнуть. То есть наблюдается автокорреляция остатков 1-го порядка.