«Регрессионные уравнения»

Теоретические предпосылки

Определение: *Регрессионное уравнение* (регрессионная модель) отражает зависимость между переменными: между одной зависимой (эндогенной, объясняемой) и одной или же несколькими независимыми (экзогенными, объясняющими) переменными (факторами, регрессорами). Зависимая переменная обозначается как y, а независимые объясняющие переменные как $x_1, x_2, \dots, x_n$.

Определение: Уравнение, отражающее зависимость между математическим ожиданием (условного распределения) одной переменной и соответствующими значениями другой переменной, называется *регрессионным уравнением*.

Таким образом, регрессионное уравнение может быть записано в виде:

$$M\left(\frac{y}{x}\right) = f(x)$$

где $M\left(\frac{y}{x}\right)$ — условное математическое ожидание случайной переменной y при заданном значении x.

В частности, для i-го заданного значения уравнение регрессии записывается в виде:

$$M\left(\frac{y}{x_i}\right) = f(x_i)$$

Регрессионное уравнение есть некая регулярная часть зависимости между y и x, фактически наблюдаемое значение, состоит из этой регулярной части и случайной компоненты ε_i:

$$y_i = M\left(\frac{y}{x}\right) + \varepsilon_i$$

Наличие случайной компоненты обусловлено двумя причинами:

- любая регрессионная модель является упрощением действительности (на самом деле существуют другие факторы, от которых также зависит переменная Yi);
- присутствуют ошибки измерения показателей.

Определение: Однофакторным линейным регрессионным уравнением называется статистическая связь между зависимой переменной y и независимым фактором (регрессором) x, представленная в виде линейной зависимости:

$y = a + bx + \varepsilon$ или $y_i = a + bx_i + \varepsilon_i$

Здесь a и b неизвестные подлежащие оценке параметры регрессии (рисунок 61).

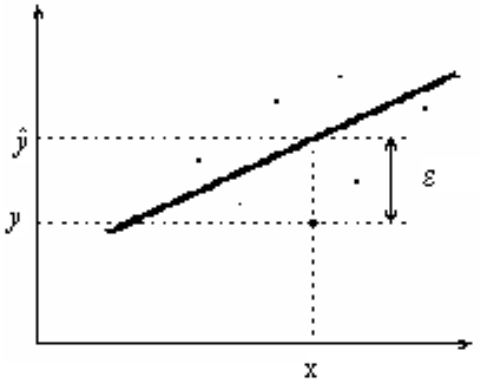


Рисунок 1 – Однофакторное регрессионное уравнение (ОЛРУ)

Случайная компонента определяется как $\varepsilon_i = y_i - \hat{y}_i$, где $\hat{y}_i = \hat{a} - \hat{b}x_i$, $\hat{y}_i$ - расчетные значения, y_i- фактические значения, $\hat{a}$ и $\hat{b}$ - оцененные значения коэффициентов ***a*** и ***b***.

Для нахождения коэффициентов регрессионного уравнения используют *метод наименьших квадратов (МНК)*:

$$F = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2 \Rightarrow \mathbf{min}$$

Запишем необходимое условие экстремума:

$$\begin{cases} \frac{\partial F}{\partial \hat{a}} = 0 \\ \frac{\partial F}{\partial \hat{b}} = 0 \end{cases} ;\Rightarrow \quad \begin{matrix} \frac{\partial F}{\partial a} = -2\sum_{i=1}^{n}(Y_i - \hat{a} - \hat{b}X_i) = 0 \\ \frac{\partial F}{\partial b} = -2\sum_{i=1}^{n}X_t(Y_i - \hat{a} - \hat{b}X_i) = 0 \end{matrix} \quad \text{или} \quad \begin{matrix} \sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}X_i) = 0 \\ \sum_{i=1}^{n}X_i(Y_i - \hat{a} - \hat{b}X_i) = 0 \end{matrix}$$

Раскрывая скобки, получим стандартную форму нормальных уравнений:

$$\hat{a}n + \hat{b}\sum X_i = \sum Y_i$$
$$\hat{a}\sum X_i + \hat{b}\sum X_i^2 = \sum X_i Y_i$$

Разрешая систему относительно $\hat{a}, \hat{b}$, получаем:

$$\hat{b} = \frac{n\sum X_i Y_i - \left(\sum X_i\right)\left(\sum Y_i\right)}{n\sum X_i^2 - \left(\sum X_i\right)^2}$$
$$\hat{a} = \frac{1}{n}\sum Y_i - \frac{1}{n}\sum X_i \cdot \hat{b}$$

Практический смысл коэффициента b: показывает прирост *y*, приходящийся на единицу прироста *x*.

Адекватность регрессионного уравнения – соответствие его реальному моделируемому процессу, достоверность его параметров.

Последовательность анализа адекватности регрессионного уравнения:

1. Оценка качества подгонки.
2. Проверка различных гипотез относительно параметров уравнения.
3. Проверка условий для получения состоятельных, несмещенных, эффективных оценок $\tilde{a}$ и $\tilde{b}$.
4. Содержательный анализ модели и корректировка модели.
5. Прогнозирование данных по модели.

1. Оценка качества подгонки.

Показатели качества подгонки:

1. Остаточная дисперсия для ОЛРУ $\sigma^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$, где **n** – число наблюдений. Чем меньше σ^2, тем лучше качество подгонки.

2. Коэффициент детерминации $R^2 = r_{yx}^2$, где r_{yx}-парный линейный коэффициент корреляции.Чем ближе R^2 к единице, тем лучше качество подгонки.

3. Скорректированный (adjusted) коэффициент детерминации для ОЛРУ:

$$R^2_{adj} = 1 - (-R^2) * \frac{n-1}{n-2} = \frac{R^2(n-1)-1}{n-2}$$

Скорректирован на число степеней свободы. R^2_{adj} позволяет сравнивать две регрессии, одна из которых является укороченной.

4. Средняя ошибка аппроксимации $A = \frac{1}{n}\sum\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\%$

Чем меньше рассеяние эмпирических точек вокруг теоретической линии регрессии, тем меньше А. Если А<5-7%, то качество модели хорошее.

2 Проверка различных гипотез относительно параметров уравнения.

1. *F-критерий.*

H_0: $\hat{a} = \hat{b} = 0$, (т.е. линейная связь между x и y отсутствует);

H_1: $\hat{a}^2 + \hat{b}^2 \neq 0$, (т.е. наличие линейной связи).

$$F_{расч} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \hat{y}_i)^2 \Big/ (n-2)} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sigma^2}$$

$F_{табл} = F^p_{1,n-2}$-табличное значение распределения Фишера для вероятности p и степеней свободы $m_1 = 1, m_2 = n-2$.

$F_{табл} > F_{расч} \Rightarrow$ принимаем H_0 с вероятностью p

$F_{табл} < F_{расч} \Rightarrow$ отвергаем H_0 в пользу H_1 с вероятностью p.

2. *t-критерий.*

$t_{расч} = \frac{\hat{b}}{\sigma_b}$, где σ_b- стандартная ошибка оценки коэффициента регрессии,

$t_{табл} = t^p_{n-2}$ - табличное значение распределения Стьюдента для вероятности p и степени свободы $m=n-k-1$.

Для коэффициента b:

$H0_b$: $\hat{b} = 0$, (т.е. фактор X незначим);

$H1_b$: $\hat{b} \neq 0$, (т.е. фактор X значим).

$$\sigma_b = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-2)\sum_{i=1}^{n}(x_i - x)^2}}$$

Для свободного члена a:

$H0_a$: $\hat{a} = 0$ (свободный член незначим)

$H1_a$: $\hat{a} \neq 0$ (свободный член значим)

$$\sigma_a = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-2)\sum_{i=1}^{n}(y_i - y)^2}}$$

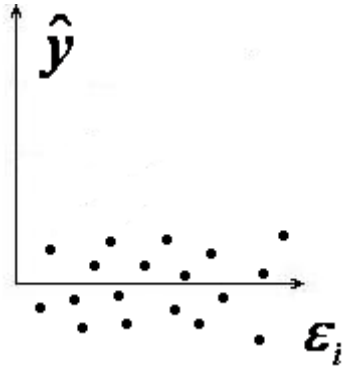
3 Проверка условий для получения состоятельных, несмещенных, эффективных оценок $\tilde{a}$ и $\tilde{b}$.

Остатки – это разность между исходными (наблюдаемыми) значениями зависимой переменной и предсказанными значениями. Исследуя остатки, мы можем оценить степень адекватности модели опытным данным.

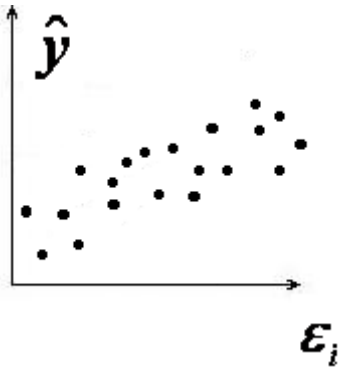
При изменении спецификации модели значение оценок остатков могут меняться. Поэтому в задачу регрессионного анализа входит не только построение самой модели, но и исследование остаточных величин.

I. Для проверки случайного характера остатков ε строят график зависимости остатков ε_i от расчетных значений зависимой переменной $\hat{y}$.

Если на графике нет направленности в расположении точек ε_i , то остатки ε случайные величины.



Если ε зависит от $\hat{y}$, то остаточная компонента ε не случайна.



Остатки – носят систематический характер. В этих случаях возможно следовало выбрать в качестве регрессионной связи нелинейную зависимость.

II. *Проверка условия* $M(\varepsilon_i) = 0$ *(рисунок 3)*

$H_0 : M(\varepsilon_i) = 0$, (математическое ожидание остатков равно нулю);

$H_1 : M(\varepsilon_i) \neq 0$, (математическое ожидание остатков отлично от нуля).

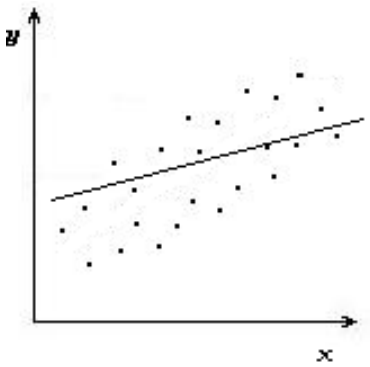


Рисунок 3. Пример смещенного оценивания.

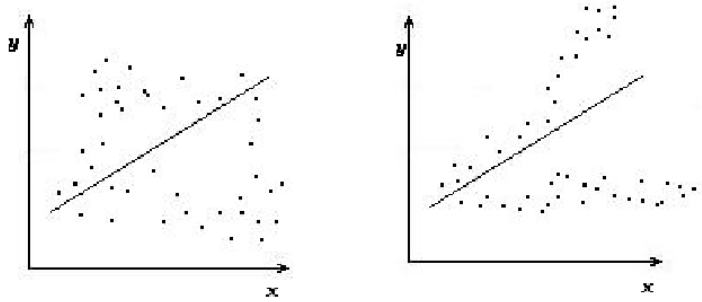
$t_{расч} = \dfrac{\mu}{\sigma/\sqrt{n}}$, где $\sigma = \sqrt{\dfrac{\sum(\varepsilon_i - \bar{\varepsilon})^2}{n-1}}$ - несмещенное выборочное стандартное отклонение, μ-выборочное среднее.

$t_{табл} = t^{p}_{n-1}$ - табличное значение распределения Стьюдента для вероятности p и степени свободы m=n-1.

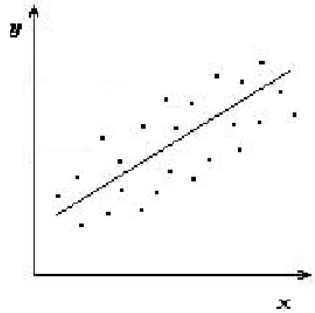
$|t_{расч}| < t_{табл} \Rightarrow$ принимаем H_0 с вероятностью p;

$|t_{расч}| > t_{табл} \Rightarrow$ отвергаем H_0 в пользу H_1 с вероятностью p.

III. Проверка условия $D(\varepsilon_i) = \sigma^2 = const$.



$D(\varepsilon_i) \neq \sigma^2$ *-гетероскедастичность*



$D(\varepsilon_i) = \sigma^2$

гомоскедастичность

Рисунок 4. Пример гомоскедастичности и гетероскедастичности в остатках.

Протестировать на наличие гетероскедастичности в остатках регрессионное уравнение можно с помощью теста Голфелда-Квандта:

H_0: $D(\varepsilon_i) = \sigma 2$ (отсутствие гетероскедастичности, наличие гомоскедастичности)

H_1: $D(\varepsilon_i) \neq \sigma 2$ (наличие гетероскедастичности)

Порядок проведения теста:

1) упорядочить y_i $(i = 1,..n)$ по возрастанию x_i;

2) исключить d средних наблюдений $(d \approx \frac{n}{6})$;

3) построить две независимые регрессии: $y_i = \tilde{a}_1 + \tilde{b}_1 x_i + e_{1i}$ для первой группы и $y_i = \tilde{a}_2 - \tilde{b}_2 x_i + e_{2i}$ для второй группы.

4) определить суммы квадратов остатков $S_1 = \sum_{i=1}^{n_1} e_{1i}^2$ и $S_2 = \sum_{i=1}^{n_2} e_{2i}^2$ (где n1, n2-число наблюдений в каждой подгруппе), а также значение статистики $F_{расч} = \frac{\max(S_1, S_2)}{\min(S_1, S_2)}$.

$F_{табл} = F^p_{n_1-2,\ n_2-2}$ - табличное значение распределения Фишера для вероятности p и степеней свободы $m_1=n_1-2$, $m_2=n_2-2$.

Для проверки нулевой гипотезы о наличие гомоскедастичности также используют тест Уайта и тест ранговой корреляции Спирмена.

IV. Проверка условия $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j$.

Тест Дарбина-Уотсона: обнаружение автокорреляции остатков вида $\varepsilon_i = \rho\varepsilon_{i-1} + e_i$ (*)

H_0:ρ=0, (т.е. автокорреляция остатков отсутствует);

H_1: ρ>0 или ρ<0, (наличие положительной или отрицательной автокорреляции остатков).

Расчетное значение статистики Дарбина-Уотсона:

$$dw = \frac{\sum_{i=2}^{n} (\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^{n} \varepsilon_i^2}$$

d_1, d_2 - табличные значения распределения Дарбина-Уотсона для степеней свободы n, и вероятности p. Области принятия соответствующих гипотез представлены графически на рисунке 5:

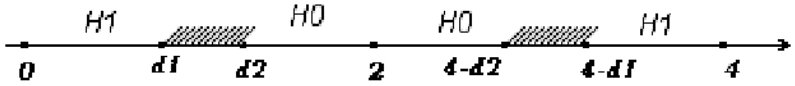


Рисунок 5. Зоны принятия решения для теста Дарбина-Уотсона

$d_1 < dw < d_2$ и $4 - d_2 < dw < 4 - d_1$ - зона неопределенности

V. Проверка условия $\varepsilon_i \sim N(0, \sigma^2)$, *т.е. согласуются ли остатки регрессии с нормальным законом.*

Существует множество критериев проверки нормальности распределения. Проверку будем производить на основе критерия Колмогорова-Смирнова.

H0: $F(\varepsilon) = F_0(\varepsilon)$, где $F_0(\varepsilon)$ - функция нормального распределения (распредеение остатков согласуется с нормальным);

H1: $F(\varepsilon) \neq F_0(\varepsilon)$, (распределение остатков не согласуется с нормальным распределением)

KS-критическое табличное значение распределения Колмогорова-Смирнова для вероятности p и объема выборки n

$KS_{\text{табл}} > KS_{\text{расч}} \Rightarrow$ принимаем H_0 с вероятностью p;

$KS_{табл} < KS_{расч} \Rightarrow$ отвергаем H_0 в пользу H_1 с вероятностью p.

Работа в R Studio

1. Создадим рабочую область в R Studio в виде отчета через команду *new file – new R Markdown.*
2. Загрузим следующие пакеты (библиотеки) в R:

library(ggplot2)
library(memisc)
library(DescTools)
library(broom)
library(caTools)
library(lmtest)
library(dplyr)
library(readxl)

- **ggplot2** — популярный графический пакет, полноценная и законченная система, наследующая идеи "Графической грамматики" (Grammar of Graphics, отсюда в названии gg).
- **memisc** — пакет, состоящий из инструментов для подготовки данных исследований, проведения имитационных исследований и представления результатов статистического анализа.
- **DescTools** — набор, содержащий основные статистические функции и удобные команды для эффективного описания данных.
- **readxl —** импорт excel данных в R.
- **broom —** преобразование данных статистических функций в формат tidy data («аккуратные данные» или «упорядоченные данные»).
- **dplyr** — расширения грамматических конструкций для манипуляций с данными.
- **lmtest** — сбор тестов, наборов данных и примеров для диагностической проверки в моделях линейной регрессии.

3. Загрузим набор данных − **diamond.**
   dmnds <- diamonds
4. Посмотрим загруженные данные
   View(dmnds)

carat	cut	color	clarity	depth	table	price
0.23	Ideal	E	SI2	61.5	55.0	326
0.21	Premium	E	SI1	59.8	61.0	326
0.23	Good	E	VS1	56.9	65.0	327
0.29	Premium	I	VS2	62.4	58.0	334
0.31	Good	J	SI2	63.3	58.0	335
0.24	Very Good	J	VVS2	62.8	57.0	336
0.24	Very Good	I	VVS1	62.3	57.0	336
0.26	Very Good	H	SI1	61.9	55.0	337
0.22	Fair	E	VS2	65.1	61.0	337
0.23	Very Good	H	VS1	59.4	61.0	338
0.30	Good	J	SI1	64.0	55.0	339
0.23	Ideal	J	VS1	62.8	56.0	340

5. Разделим выборку на тестовую и обучающую.
   set.seed(56)
   split <- sample.split(dmnds$price, SplitRatio = 0.75)

train <- subset(dmnds, split == TRUE)
test <- subset(dmnds, split == FALSE)

6. Построим модель линейной регрессии. В качестве зависимой переменной выступает – *price*, независимой – *carat.*

model_1 <- lm (data = train, price ~ carat)
summary(model_1)

```
Call:
lm(formula = price ~ carat, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-19246.6   -859.8    -23.7    566.4  12643.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2317.63      15.48  -149.7   <2e-16 ***
carat        7900.65      16.11   490.5   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1626 on 42124 degrees of freedom
Multiple R-squared:  0.851,    Adjusted R-squared:  0.851
F-statistic: 2.406e+05 on 1 and 42124 DF,  p-value: < 2.2e-16
```

Согласно полученным результатам уравнение регрессии имеет вид:

$$\boldsymbol{y = -2317.63 + 7900.65 * x}$$

Команда *summary()* выдает полную информацию о построенной модели:

- значения остатков (*residuals* − разность модельных и истинных значений переменной *y*). Если объем выборки большой, то печатается оценка распределения остатков (квартили);
- коэфициенты модели;
- t(df) and p-value —значение t-критерия и уровня значимости p. t-критерий используется для проверки гипотезы о значимости коэффициентов (в примере все коэфициенты значимы, поскольку все вероятности (2e-16, то есть $2*10^{-16}$);
- Std. Error — стандартная ошибка оценки коэффицинтов регрессии;
- F-statistic — значение критерия Фишера для проверки гипотезы о наличие или отсутствие линейной связи;
- Adjusted R-squared — квадрат коэфициента скоректированной корреляции (R-squared), который показывает долю дисперсии y, объясненной с использованием модели;
- Multiple R-squared — коэффициент множественной корреляции (эта статистика полезна во множественной регрессии, когда вы хотите описать зависимости между переменными) – характеризует тесноту связи зависимой переменной с совокупностью независимых переменных.

— Residual standard error − стандартная ошибка оценки. Эта статистика является мерой рассеяния наблюдаемых значений относительно регрессионной прямой

Для нахождения коэффициентов детерминации необходимо применить следующие команды:

```{r}
summary(lm(data = train, price ~ carat))$r.squared
summary(lm(data = train, price ~ carat))$adj.r.squared
```

```
[1] 0.851
[1] 0.8509965
```

Для нахождения остаточной дисперсии необходимо выполнить следующую команду:

```{r}
summary(lm(data = train, price ~ carat))$sigma^2
```

```
[1] 2643664
```

Также можно применить функцию glance() для нахождения вышеописанных параметров:

```{r}
glance(model_1)
```

```
```

r.squared <dbl>	adj.r.squared <dbl>	sigma <dbl>	statistic <dbl>	p.value <dbl>	df <int>	logLik <dbl>	AIC <dbl>	BIC <dbl>	deviance <dbl>
0.851	0.8509965	1625.935	240587.4	0	2	-371246	742498.1	742524	111361717552

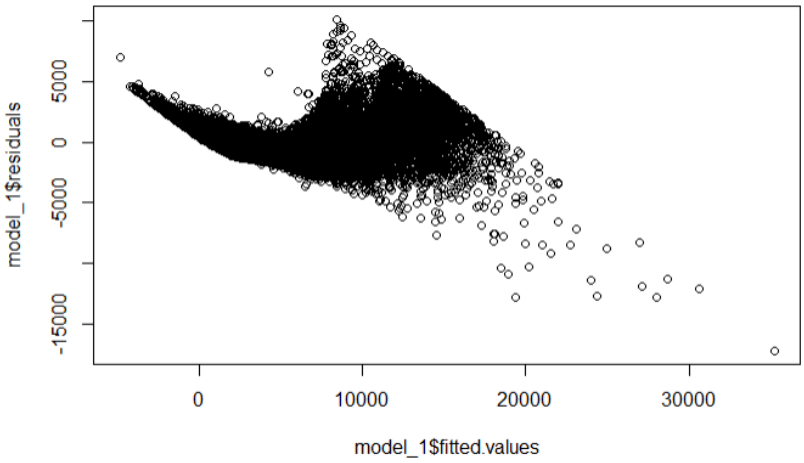
1 row | 1-10 of 11 columns

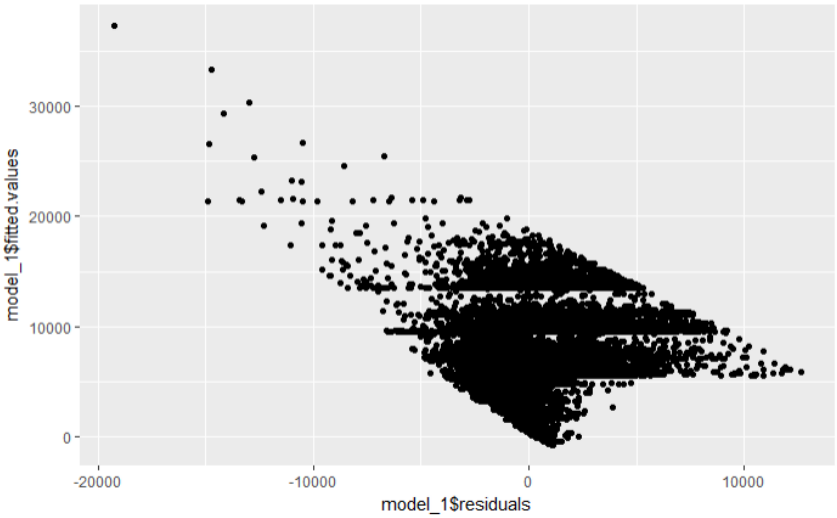
7. Для проверки условий для получения состоятельных, несмещенных, эффективных оценок необходимо выполнить следующие расчеты в R.

7.1. Проверка случайности остаточной компоненты

Для выполнения проверки данной предпосылки необходимо в R построить график зависимости остатков от спрогнозированных значений модели.

```{r}
plot(model_1$fitted.values, model_1$residuals)
qplot(y = model_1$fitted.values, x = model_1$residuals)
```





7.2. Проверка условия $M(\varepsilon_i) = 0$

```{r}
mean(model_1$residuals)
```

```
[1] 4.28474e-13
```

Для вычисления $t_{расч}$ воспользуемся следующими возможностями R:

```{r}
a<-mean(model_1$residuals)
b<-sd(model_1$residuals, na.rm = FALSE)
n <- sqrt(42126)
tm <- a/b*n
str(tm)

```

```
num -3.27e-14
```

Для вычисления $t_{табл}$ воспользуемся командой qt() со степенью свободы $df = n - 1 = 42126 - 1 = 42125$ и с уровнем доверительной вероятности $p = 0{,}95$. В качестве уровня доверительной вероятности

```{r}
qt(0.975,42125)
```

```
[1] 1.96002
```

7.3.Проверка условия $D(\varepsilon_i) = \sigma^2 = const$.

H_0: отсутствие гетероскедастичности

H_1: наличие гетероскедастичности

Для проверки условия о постоянстве дисперсии могут быть использованы следующие тесты:

- Тест Гольфелда-Квандта
- Тест Уайта
- Тест Спирмена
- Тест Бройша-Погана
- Тест Глейзера

Проведем в R Тест Бройша-Погана

```{r}
bptest(model_1)
```

```
studentized Breusch-Pagan test

data:  model_1
BP = 6734, df = 1, p-value < 2.2e-16
```

Так как значение p-value меньше 0.05, нулевая гипотеза о гомоскедастичности остатков отвергается.

Проведем в R тест Спирмена

Наличие гетероскедастичности в остатках регрессии можно проверить с помощью теста ранговой корреляции Спирмена. Суть теста заключается в определении наличия связи между ростом остаточной компоненты и ростом независимого фактора, то есть определение роста дисперсии остатков. Такая зависимость проверяется на основе расчета коэффициента ранговой корреляции Спирмена ρ между остатками модели ε и независимым фактором x. Проверка статистической значимости коэффициента Спирмена на основе соответствующего *t*-критерия аналогична проверке нулевой гипотезы об отсутствии гетероскедастичности в остатках.

```{r}
cor.test (train$carat, model_1$residuals, method = "spearman")
```

```
        Spearman's rank correlation rho

data:  train$carat and model_1$residuals
S = 1.4815e+13, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.1890263
```

7.4. Проверка условия $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j$.

H_0: ρ=0 (т.е. автокорреляция остатков отсутствует);

H_1: ρ>0 или ρ<0 (наличие положительной или отрицательной автокорреляции остатков).

— Тест Дарбина-Уотсона

— Тест Бройша- Годфри

Проведем в R тесты Бройша- Годфри и Дарбина-Уотсона

```{r}
bgtest(model_1)
dwtest(model_1)
```

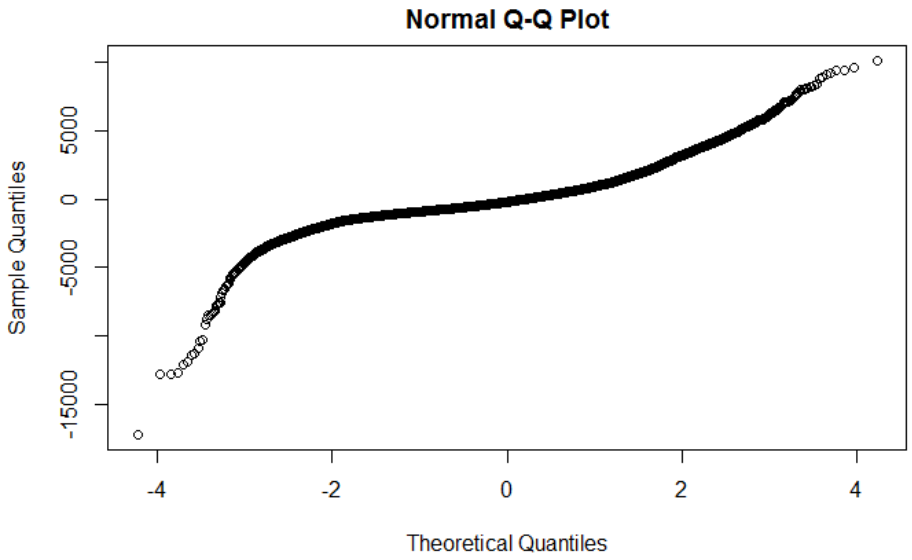
```
Breusch-Godfrey test for serial correlation of order up to 1

data:  model_1
LM test = 9899, df = 1, p-value < 2.2e-16


        Durbin-Watson test

data:  model_1
DW = 1.0539, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

Поскольку значение p-value меньше 0.05 нулевая гипотеза об отсутствии автокорреляции остатков отвергается.

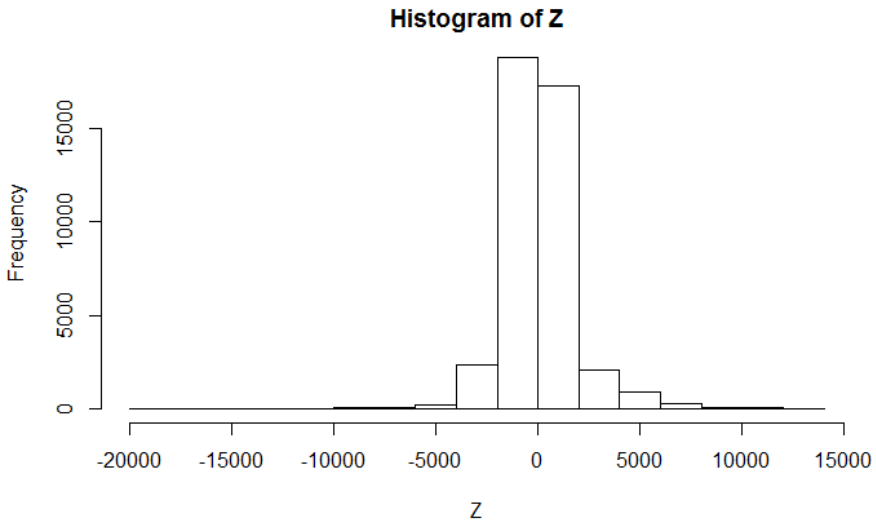
7.5 Проверка условия $\varepsilon_i \sim N(0, \sigma^2)$, т.е. согласуются ли остатки регрессии с нормальным законом.

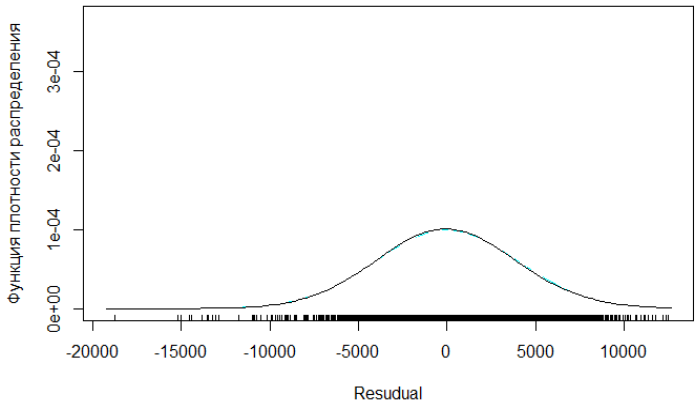
Рассмотрим графические тесты:

```{r}
qqnorm(model_1$residuals)
```




```{r}
library(sm)
Z<- model_1$residuals
hist(Z)
sm.density(model_1$residuals, model = "Normal",xlab = "Resudual", ylab = "Функция плотности распределения", xlim=c(-19246.62,12642.95))
```





Рассмотрим параметрические тесты:

```{r}
library(nortest)
lillie.test(model_1$residuals)
```

```
Lilliefors (Kolmogorov-Smirnov) normality test

data:  model_1$residuals
D = 0.14802, p-value < 2.2e-16
```

Так как значение p-value меньше 0.05, нулевая гипотеза о согласие распределения остатков с нормальным законом распределения отвергается.

Рассмотрим также тест Шапиро-Уилка. Данный тест применяется для выборки объема не менее 3 наблюдений и не более 5000. Для рассмотрения данного теста на текущей выборке выберем 4990 элементов выборки. Данная процедура применена только для демонстрации метода. Если ваша выборка включает больше 5000 тысяч значения, то данный тест применять нецелесообразно.

```{r}
sh <- model_1$residuals[1:4990]
View(sh)
shapiro.test(sh)
```

Функция shapiro.test () выполняет тест Шапиро-Уилка. Обънм

```
        Shapiro-Wilk normality test

data:  sh
W = 0.97111, p-value < 2.2e-16
```

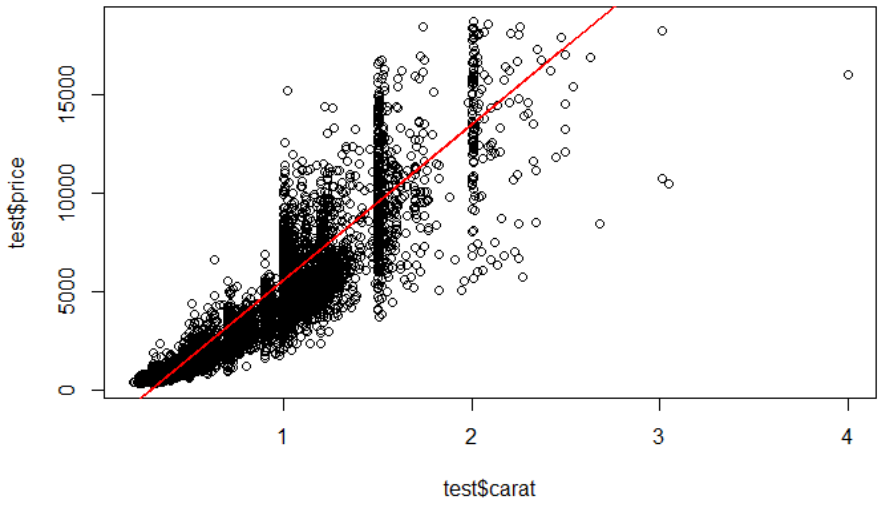
Так как значение p-value меньше 0.05, нулевая гипотеза о согласии распределения остатков с нормальным законом распределения отвергается.

8. Прогнозирование значений по полученной модели на тестовой выборке

```{r}
test$model_1 <- predict(model_1, test)
View(test$model_1)
```

```
Посмотрим результат модели на графике

```{r}
plot(test$carat, test$price)
lines(test$carat, predict(model_1, test), col = 'red')
```



Корректировка модели.

1. Если свободный член модели незначим, то необходимо исключить его из рассмотрения и проанализировать модель снова.

Для этого необходимо применить следующую команду

model_2 <- lm (data = train, price ~ carat - 1), где -1 в формуле модели используется для того, чтобы исключить свободный член в регрессионной модели.

После построения новой модели необходимо выполнить следующие шаги:

1. Оценка качества подгонки.
2. Проверка различных гипотез относительно параметров уравнения.
3. Проверка условий для получения состоятельных, несмещенных, эффективных оценок $\widetilde{a}$ и $\widetilde{b}$.
4. Содержательный анализ модели и корректировка модели.
5. Прогнозирование данных по модели.

2. Если не выполняется одна или несколько предпосылок МНК, необходимо сделать выводы о корректировке модели.
```