

Статистический анализ

Индивидуальное домашнее задание №3

Вариант 8

Киреев Константин

23.12.2020

Результаты статистического эксперимента приведены в таблице 1. Требуется оценить характер (случайной) зависимости переменной Y от переменной X .

Таблица 1. $\alpha_1 = 0.05; h = 1.70$

##	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]
## x	3.00	2.00	1.00	2.00	0.0	2.00	3.00	1.00	1.00	2.00	3.00	3.0	3.00	3.00
## y	8.89	10.33	8.22	12.15	6.6	7.65	8.64	4.05	6.68	2.27	10.63	5.7	8.92	4.59
##	[,15]	[,16]	[,17]	[,18]	[,19]	[,20]	[,21]	[,22]	[,23]	[,24]	[,25]	[,26]	[,27]	
## x	2.00	0.00	3.00	1.00	1.00	2.00	2.00	3.00	2.00	1.00	1.00	4.00	2.00	
## y	10.82	11.09	2.14	6.63	8.02	6.99	8.28	7.01	4.09	8.52	5.48	9.92	7.35	
##	[,28]	[,29]	[,30]	[,31]	[,32]	[,33]	[,34]	[,35]	[,36]	[,37]	[,38]	[,39]	[,40]	
## x	2.00	2.00	1.00	1.0	1.00	1.00	3.00	2.00	2.00	3.00	4.00	4.00	3.00	
## y	9.92	10.16	10.49	9.1	7.29	5.29	5.06	8.32	7.64	10.25	9.02	8.92	4.98	
##	[,41]	[,42]	[,43]	[,44]	[,45]	[,46]	[,47]	[,48]	[,49]	[,50]				
## x	3.00	4.00	3.00	2.00	2.00	3.00	2.00	3.00	1.00	2.00				
## y	8.33	6.61	9.19	11.52	9.74	6.55	5.67	6.01	4.57	2.29				

Задание 1

Построить графически результаты эксперимента. Сформулировать линейную регрессионную модель переменной Y по переменной X . Построить МНК оценки параметров сдвига β_0 и масштаба β_1 . Построить полученную линию регрессии. Оценить визуально соответствие полученных данных и построенной оценки.

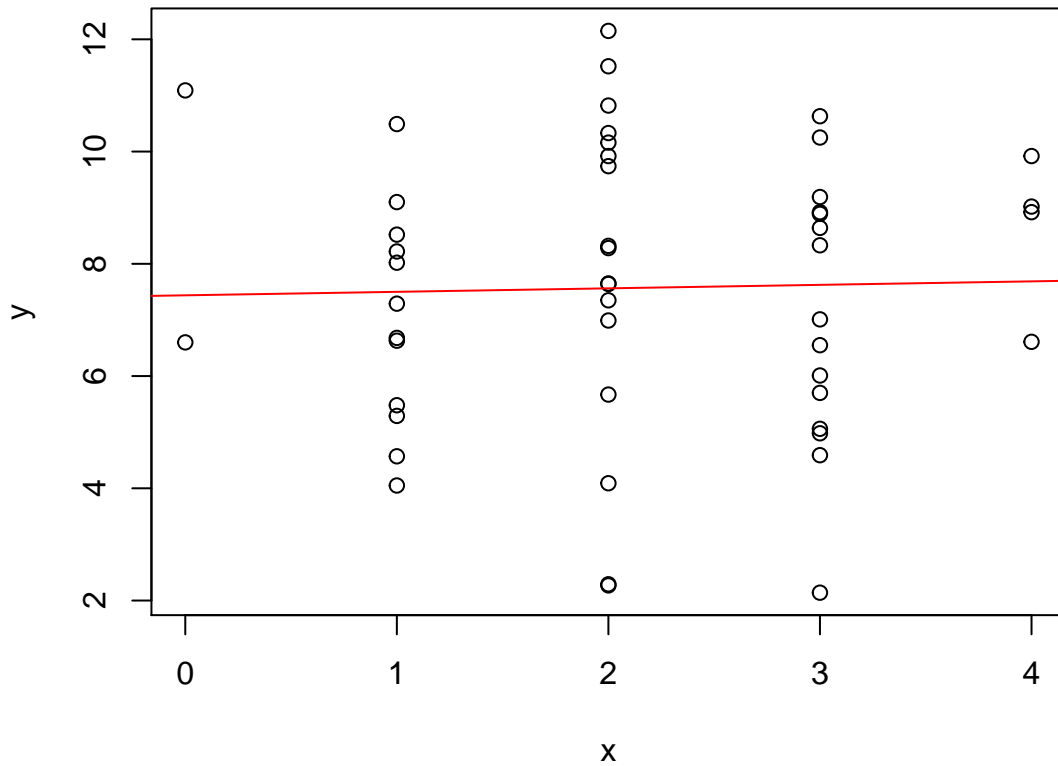
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2) \quad (1)$$

$$\beta_1 = \frac{\bar{xy} - \bar{x}\bar{y}}{\bar{x^2} - \bar{x}^2} = 0.0621511 \quad (2)$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X} = 7.4385966 \quad (3)$$

$$Y = \hat{\beta}_1 X + \beta_0 = 0.0621511X + 7.4385966 \quad (4)$$

Linear regression



Задание 2

Построить и интерпретировать несмещенную оценку дисперсии. На базе ошибок построить гистограмму с шагом h . Проверить гипотезу нормальности ошибок на уровне α_1 по χ^2 . Оценить расстояние полученной оценки до класса нормальных распределений по Колмогорову. Визуально оценить данный факт.

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \bar{Y})^2 = 6.1827473 \quad (5)$$

$$\varepsilon_i = Y_i - \hat{\beta}_1 X_i - \hat{\beta}_0 \quad (6)$$

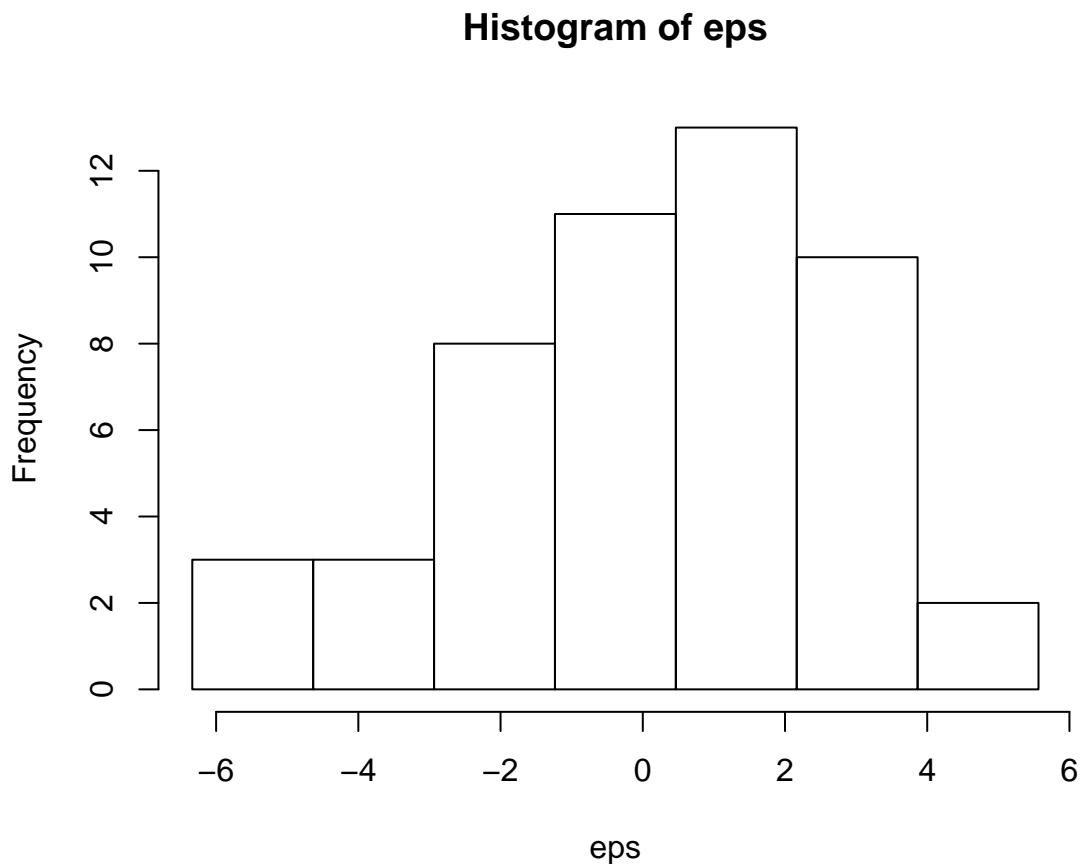
$\varepsilon_1, \dots, \varepsilon_n$:

```
## [1] 1.26495002 2.76710116 0.71925230 4.58710116 -0.83859656 0.08710116
## [7] 1.01495002 -3.45074770 -0.82074770 -5.29289884 3.00495002 -
1.92504998
## [13] 1.29495002 -3.03504998 3.25710116 3.65140344 -5.48504998 -
0.87074770
```

```

## [19]  0.51925230 -0.57289884  0.71710116 -0.61504998 -3.47289884  1.01925230
## [25] -2.02074770  2.23279888 -0.21289884  2.35710116  2.59710116  2.98925230
## [31]  1.59925230 -0.21074770 -2.21074770 -2.56504998  0.75710116  0.07710116
## [37]   2.62495002   1.33279888   1.23279888  -2.64504998   0.70495002  -
1.07720112
## [43]   1.56495002   3.95710116   2.17710116  -1.07504998  -1.89289884  -
1.61504998
## [49] -2.93074770 -5.27289884

```



$$H_0 : \varepsilon_1, \dots, \varepsilon_n \sim N(0, \sigma^2) \quad (7)$$

$$\sum_{i=1}^r \frac{n_i - np_i(0, \sigma^2))^2}{np_i(0, \sigma^2)} \rightarrow \chi^2 \quad (8)$$

Метод минимизации хи-квадрат

$$\underset{\sigma^2}{\operatorname{argmin}} \sum_{i=1}^r \frac{n_i - np_i(0, \sigma^2))^2}{np_i(0, \sigma^2)} \quad (9)$$

Разделим выборку на 6 зон:

Интервал	$(-\infty; -3.4)$	$(-3.4; -1.7)$	$(-1.7; 0)$	$(0; 1.7)$	$(1.7; 3.4)$	$(3.4; \infty)$
m_i	5	8	10	15	9	3

Задача реализована в R с помощью скрипта:

```
csq <- function(sgm.sq) {
  prob <- pnorm(up, 0, sgm.sq) - pnorm(lw, 0, sgm.sq)
  return (sum((m-n*prob)^2)/prob/n)
}
XM <- nlm(csq, sqrt(s))
XM$estimate^2; XM$minimum
```

Получаем $\hat{\sigma}^2 = 6.0184705$ и $\chi^2 = 1.5922697$

$$l = r - d - 1 = 4 \quad (10)$$

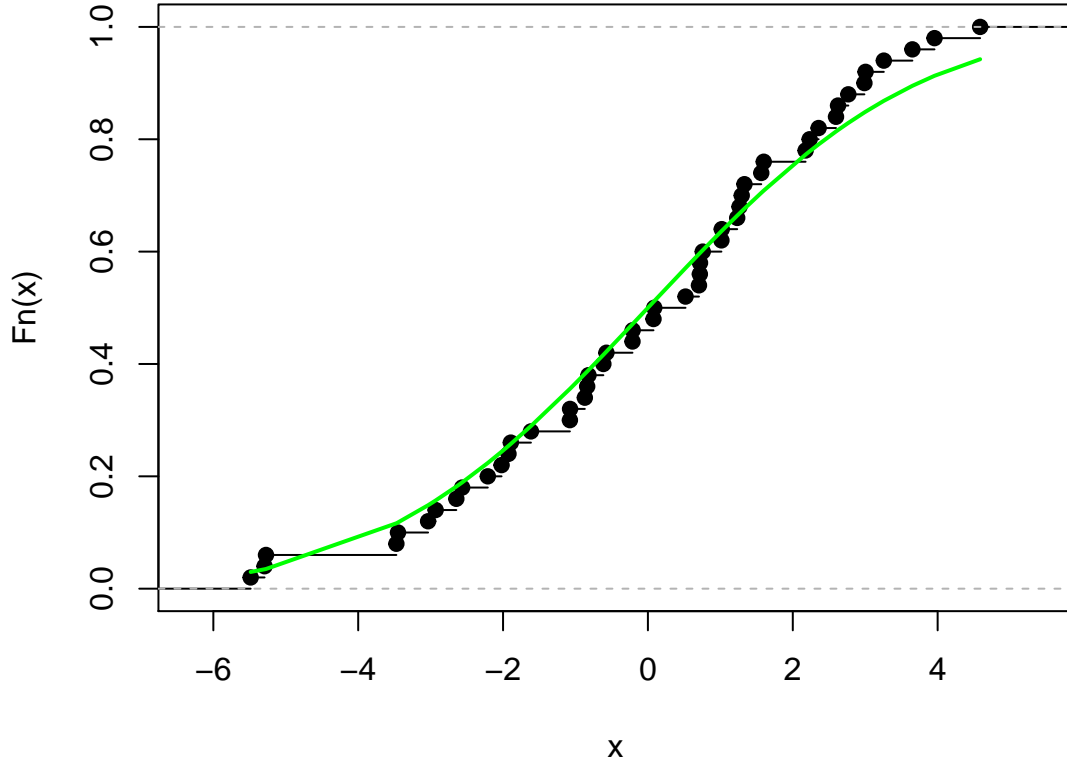
$$\chi^2 = 9.487729 \quad (11)$$

$$\chi^2 = 1.5922697 < \chi^2 = 9.487729 \Rightarrow \text{гипотеза } H_0 \text{ принимается} \quad (12)$$

Оценим расстояние оценки до класса нормальных распределений по Колмогорову. Минимизируем статистику Колмогорова с помощью следующего скрипта:

```
kolm.stat<-function(s) {
  sres<-sort(eps)
  fdistr<-pnorm(sres, 0, s)
  max(abs(c(0:(n-1))/n-fdistr), abs(c(1:n)/n-fdistr))
}
ks.dist<-nlm(kolm.stat, 2.453257)
```

Получаем расстояние $D = 0.0756747$ и $\tilde{\sigma}^2 = 8.4739449$. Ниже представлены эмпирическая функция распределения ошибок и функция распределения $N(0, \tilde{\sigma}^2)$.



Задание 3

В предположении нормальности ошибок построить доверительные интервалы для параметров β_0 и β_1 уровня доверия $1 - \alpha_1$. Построить доверительный эллипс уровня доверия $1 - \alpha_1$ для (β_0, β_1) (вычислить его полуоси).

$$\psi = C^T \beta \quad (13)$$

$$\hat{\psi} \sim N(\psi, \sigma^2 b), \sigma^2 b = \text{var}(\hat{\psi}) = \sigma^2 C^T (X X^T)^{-1} C \quad (14)$$

$$\frac{\hat{\psi} - \psi}{\sigma \sqrt{b}} \sim N(0, 1); \frac{(n-r)s^2}{\sigma^2} \sim \chi_{n-r}^2 \Rightarrow \frac{\hat{\psi} - \psi}{s \sqrt{b}} \sim S_{n-r} \quad (15)$$

$$x_\alpha : S_{n-r}(x_\alpha) = 1 - \frac{\alpha}{2} \quad (16)$$

$$P(-x_\alpha \leq \frac{\hat{\psi} - \psi}{s \sqrt{b}} \leq x_\alpha) = P(\hat{\psi} - x_\alpha s \sqrt{b} \leq \psi \leq \hat{\psi} + x_\alpha s \sqrt{b}) \quad (17)$$

$$x_\alpha = 2.0106348 \quad (18)$$

$$\beta_0 \in (5.7687784; 9.1084148) - \text{ДИ с уровнем доверия } 1 - \alpha \quad (19)$$

$$\beta_1 \in (-0.6447393; 0.7690416) - \text{ДИ с уровнем доверия } 1 - \alpha \quad (20)$$

$$\frac{(\hat{\psi} - \psi)^T (C^T (X X^T)^{-1} C)^{-1} (\hat{\psi} - \psi)}{\sigma^2} \sim \chi_q^2 \quad (21)$$

$$\frac{((\hat{\psi} - \psi)^T (C^T (X X^T)^{-1} C)^{-1} (\hat{\psi} - \psi) q \sigma^2)}{\frac{n-r}{n+r} \cdot \frac{s^2}{\sigma^2}} = \frac{(\hat{\psi} - \psi)^T (C^T (X X^T)^{-1} C)^{-1} (\hat{\psi} - \psi)}{q s^2} \sim F_{q, n-r} \quad (22)$$

$$x_\alpha : F_{q, n-r}(x_\alpha) = 1 - \alpha \quad (23)$$

$$\left\{ \psi : (\hat{\psi} - \psi)^T (C^T (X X^T)^{-1} C)^{-1} (\hat{\psi} - \psi) \leq s^2 q x_\alpha \right\} \quad (24)$$

$$x_\alpha = 3.1907273 \quad (25)$$

$$\begin{pmatrix} \beta_1 - \hat{\beta}_1 & \beta_0 - \hat{\beta}_0 \end{pmatrix} \begin{pmatrix} 0.019992 & -0.0427829 \\ -0.0427829 & 0.1115554 \end{pmatrix} \begin{pmatrix} \beta_1 - \hat{\beta}_1 \\ \beta_0 - \hat{\beta}_0 \end{pmatrix} \leq 39.4549219 \quad (26)$$

Собственные числа матрицы: **0.1284342, 0.0031132**

После ортогонального преобразования получаем:

$$0.1284342(\beta_1^* - 0.0621511)^2 + 0.0031132(\beta_0^* + 7.4385966)^2 \leq 39.4549219 \quad (27)$$

$$\frac{(\beta_1^* - 0.0621511)^2}{17.5271082^2} + \frac{(\beta_0^* + 7.4385966)^2}{112.5765063^2} \leq 1 \quad (28)$$

Полуоси эллипса: **17.5271082; 112.5765063**.

Задание 4

Сформулировать гипотезу независимости переменной Y от переменной X. Провести проверку значимости.

$$\psi = C^T \beta; \hat{\psi} \sim N(\psi, \sigma^2 C^T (X X^T)^{-1} C) \quad (29)$$

$$H_0 : \psi = 0 \quad (30)$$

$$\text{Статистика F-критерия:} \quad (31)$$

$$F = \frac{\hat{\psi}^T (C^T (X \hat{X}^T)^{-1} C)^{-1} \hat{\psi}}{q s^2} \stackrel{H_0}{\sim} F_{q, n-r} \quad (32)$$

$$F = 0.0312507 \quad (33)$$

$$x_\alpha : F_{q, n-r}(x_\alpha) = 1 - \alpha; \phi(Y, X) = 1_{F > x_\alpha} \quad (34)$$

$$x_\alpha = 4.0426521 \quad (35)$$

$$H_1 : \hat{\psi} = \beta_1 \quad (36)$$

$F < x_\alpha \Rightarrow$ Принимаем гипотезу H_0 , переменная Y независима с переменной X на уровне значимости α (37)

0.8604259 – наибольшее значение уровня значимости, на котором нет оснований отвергнуть данную гипотезу. (38)

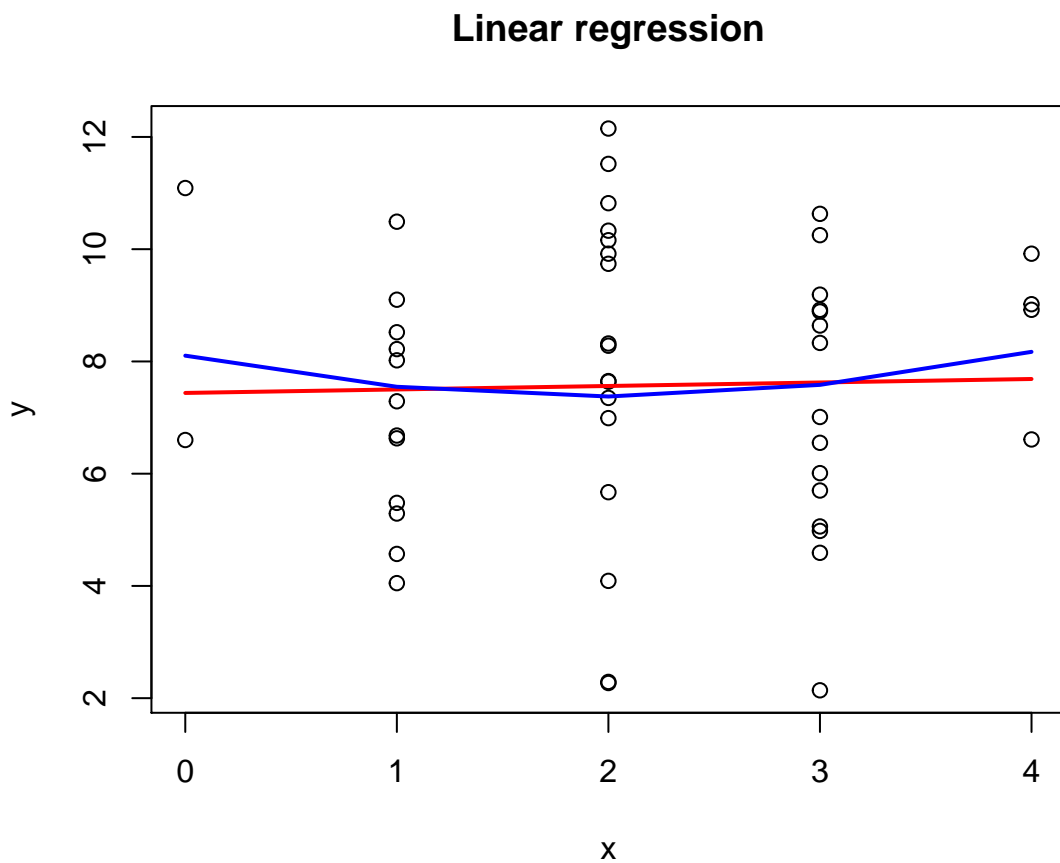
Задание 5

Сформулировать модель, включающую дополнительный член с X^2 . Построить МНК оценки параметров $\beta_1, \beta_2, \beta_3$ в данной модели. Изобразить графически полученную регрессионную зависимость.

$$Y_i = \beta_3 + \beta_2 X_i + \beta_1 X_i^2 + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2) \quad (39)$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (40)$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \begin{pmatrix} 0.190431 \\ -0.744725 \\ 8.1027066 \end{pmatrix} \quad (41)$$



Задание 6

Построить несмещенную оценку дисперсии. Провести исследование нормальности ошибок как в п.3.

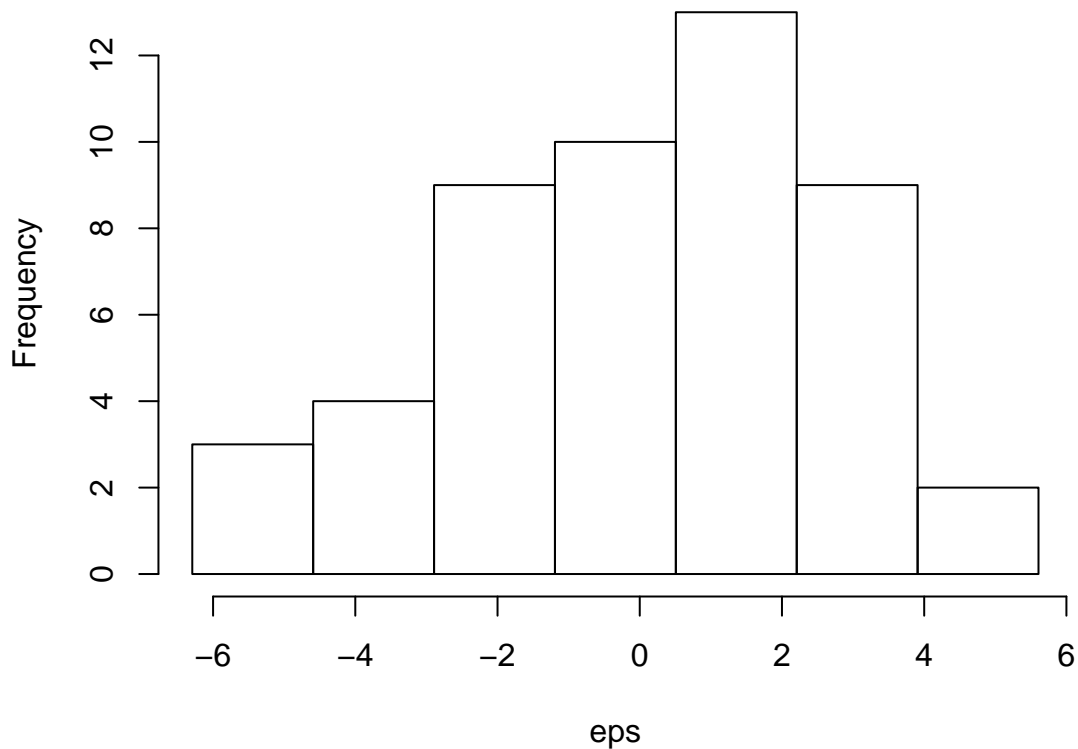
$$\hat{\sigma}^2 = \frac{1}{n-3} \sum_1^n (Y_i - \hat{\beta}_1 X_i^2 - \hat{\beta}_2 X_i - \hat{\beta}_3)^2 = 6.2575876 \quad (42)$$

$$\varepsilon_i = \hat{Y} - X^T \hat{\beta} \quad (43)$$

$\varepsilon_1, \dots, \varepsilon_n$:

```
## [1] 1.30758951 2.95501947 0.67158744 4.77501947 -1.50270658 0.27501947
## [7] 1.05758951 -3.49841256 -0.86841256 -5.10498053 3.04758951 -
1.88241049
## [13] 1.33758951 -2.99241049 3.44501947 2.98729342 -5.44241049 -
0.91841256
## [19] 0.47158744 -0.38498053 0.90501947 -0.57241049 -3.28498053 0.97158744
## [25] -2.06841256 1.74929755 -0.02498053 2.54501947 2.78501947 2.94158744
## [31] 1.55158744 -0.25841256 -2.25841256 -2.52241049 0.94501947 0.26501947
## [37] 2.66758951 0.84929755 0.74929755 -2.60241049 0.74758951 -
1.56070245
## [43] 1.60758951 4.14501947 2.36501947 -1.03241049 -1.70498053 -
1.57241049
## [49] -2.97841256 -5.08498053
```

Histogram of eps



$$H_0 : \varepsilon_1, \dots, \varepsilon_n \sim N(0, \sigma^2) \quad (44)$$

$$\sum_{i=1}^r \frac{(n_i - np_i(0, \sigma^2))^2}{np_i(0, \sigma^2)} \rightarrow \chi^2 \quad (45)$$

Метод минимизации хи-квадрат

$$\underset{\sigma^2}{\operatorname{argmin}} \sum_{i=1}^r \frac{(n_i - np_i(0, \sigma^2))^2}{np_i(0, \sigma^2)} \quad (46)$$

Разделим выборку на 6 зон:

Интервал	$(-\infty; -3.4)$	$(-3.4; -1.7)$	$(-1.7; 0)$	$(0; 1.7)$	$(1.7; 3.4)$	$(3.4; \infty)$
m_i	4	9	10	15	9	3

Получаем $\hat{\sigma}^2 = 5.5559278$ и $\chi^2 = 1.4170828$

$$l = r - d - 1 = 4 \quad (47)$$

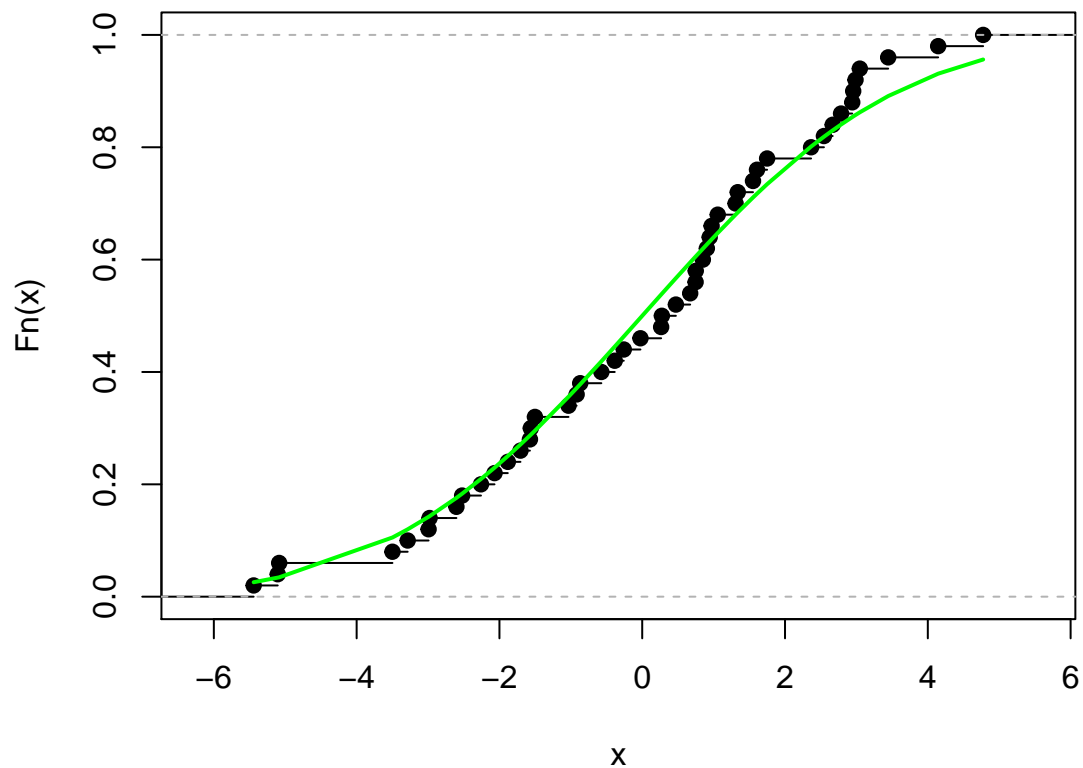
$$\chi^2 = 9.487729 \quad (48)$$

$$\chi^2 = 1.4170828 < \chi^2 = 9.487729 \Rightarrow \text{гипотеза } H_0 \text{ принимается} \quad (49)$$

Оценим расстояние оценки до класса нормальных распределений по Колмогорову. Минимизируем статистику Колмогорова с помощью следующего скрипта:

```
kolm.stat<-function(s) {
  sres<-sort(eps)
  fdistr<-pnorm(sres, 0, s)
  max(abs(c(0:(n-1))/n-fdistr), abs(c(1:n)/n-fdistr))
}
ks.dist<-nlm(kolm.stat, 2.357102)
```

Получаем расстояние $D = 0.0777714$ и $\tilde{\sigma}^2 = 7.8117778$. Ниже представлены эмпирическая функция распределения ошибок и функция распределения $N(0, \tilde{\sigma}^2)$.



Задание 7

В предположении нормальности ошибок построить доверительные интервалы для параметров $\beta_1, \beta_2, \beta_3$ уровня $1 - \alpha_1$. Написать уравнение доверительного эллипсоида уровня доверия $1 - \alpha_1$.

$$\psi = C^T \beta; \hat{\psi} \sim N(\psi, \sigma^2 C^T (X X^T)^{-1} C) \quad (50)$$

$$x_\alpha : S_{n-r}(x_\alpha) = 1 - \frac{\alpha}{2} \quad (51)$$

$$P(\hat{\psi} - x_\alpha s \sqrt{b} \leq \psi \leq \hat{\psi} + x_\alpha s \sqrt{b}) \quad (52)$$

$$x_\alpha = 2.0117405 \quad (53)$$

$$\beta_1 \in (-0.4190854; 0.7999474) - \text{ДИ с уровнем доверия } 1 - \alpha \quad (54)$$

$$\beta_2 \in (-3.4235389; 1.9340889) - \text{ДИ с уровнем доверия } 1 - \alpha \quad (55)$$

$$\beta_3 \in (5.3928258; 10.8125874) - \text{ДИ с уровнем доверия } 1 - \alpha \quad (56)$$

$$x_\alpha : F_{q,n-r}(x_\alpha) = 1 - \alpha \quad (57)$$

$$\{\psi : (\hat{\psi} - \psi)^T (C^T (X X^T)^{-1} C)^{-1} (\hat{\psi} - \psi) \leq s^2 q x_\alpha\} \quad (58)$$

$$x_\alpha = 2.8023552 \quad (59)$$

$$\begin{pmatrix} \beta_1 - \hat{\beta}_1 & \beta_2 - \hat{\beta}_2 & \beta_3 - \hat{\beta}_3 \end{pmatrix} \begin{pmatrix} 0.0146697 & -0.0621569 & 0.0511159 \\ -0.0621569 & 0.2833572 & -0.2595491 \\ 0.0511159 & -0.2595491 & 0.2899676 \end{pmatrix} \begin{pmatrix} \beta_1 - \hat{\beta}_1 \\ \beta_2 - \hat{\beta}_2 \\ \beta_3 - \hat{\beta}_3 \end{pmatrix} \leq 52.607949 \quad (60)$$

Собственные числа матрицы: **0.5580359, 0.0296035, 3.5505965×10^{-4}**

После ортогонального преобразования получаем:

$$0.5580359(\beta_1^* - 0.190431)^2 + 0.0296035(\beta_2^* - (-0.744725))^2 + 3.5505965 \times 10^{-4}(\beta_3^* - 8.1027066)^2 \leq 52.607949 \quad (61)$$

$$\frac{(\beta_1^* - 0.190431)^2}{9.7094494^2} + \frac{(\beta_2^* - (-0.744725))^2}{42.1555085^2} + \frac{(\beta_3^* - 8.1027066)^2}{384.9240317^2} \leq 1 \quad (62)$$

Полуоси эллипсоида: **9.7094494; 42.1555085; 384.9240317.**

Задание 8

Сформулировать гипотезу линейной регрессионной зависимости переменной Y от переменной X и проверить ее значимость на уровне α_1 .

$$\psi = C^T \beta; \hat{\psi} \sim N(\psi, \sigma^2 C^T (X X^T)^{-1} C) \quad (63)$$

$$H_0 : \psi = 0 \quad (64)$$

$$\text{Статистика F-критерия:} \quad (65)$$

$$F = \frac{\hat{\psi}^T (C^T (X X^T)^{-1} C)^{-1} \hat{\psi}}{q s^2} \stackrel{H_0}{\sim} F_{q,n-r} \quad (66)$$

$$F = 0.3950467 \quad (67)$$

$$x_\alpha : F_{q,n-r}(x_\alpha) = 1 - \alpha; \phi(Y, X) = 1_{F > x_\alpha} \quad (68)$$

$$x_\alpha = 4.0470999 \quad (69)$$

$$H_1 : \hat{\psi} = \beta_1 \quad (70)$$

$F < x_\alpha \Rightarrow$ Принимаем гипотезу H_0 , переменная Y линейно регрессионно зависима с переменной X на уровне значимости α (71)

0.5327012 – наибольшее значение уровня значимости, на котором нет оснований отвергнуть данную гипотезу. (72)