Статистический анализ

Переверзев Дмитрий

23.12.2020

Индивидуальное домашнее задание №3

Вариант 18

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

Таблица 1. $\alpha_1 = 0.10; h = 2.80$

```
##    [,1]  [,2]  [,3]  [,4]  [,5]  [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
## x  5.00  5.00  4.00  3.00  6.00  2.00 2.00 4.00 0.00  0.00   2.0  1.00  5.00
## y 18.13 35.07 11.95 12.34 15.09 25.97 6.23 6.98 9.33  1.72   3.3 20.49  6.83
##   [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25] [,26]
## x  0.00  5.00  6.00  4.00  0.00  3.00  5.00   3.0  3.00  2.00  1.00  0.00  5.00
## y  3.13 22.18  5.01 15.22 22.14  6.78 12.35   2.1  9.48 12.63  8.68 28.42 39.03
##   [,27] [,28] [,29] [,30] [,31] [,32] [,33] [,34] [,35] [,36] [,37] [,38] [,39]
## x  2.00  0.00  6.00  3.00  4.00   2.0  2.00  1.00  0.00  4.00  1.00  6.00  4.00
## y 17.77 19.11 12.83 17.75  4.25  15.1  0.14 18.32 13.17 12.72 16.93 23.14 23.04
##   [,40] [,41] [,42] [,43] [,44] [,45] [,46] [,47] [,48] [,49] [,50]
## x  0.00  5.00  4.00  3.00   5.0  2.00  5.00  4.00  3.00  1.00  3.00
## y 20.18  3.57  7.98 21.42  10.5  5.87 19.63 17.04 17.21 28.83  9.45
```

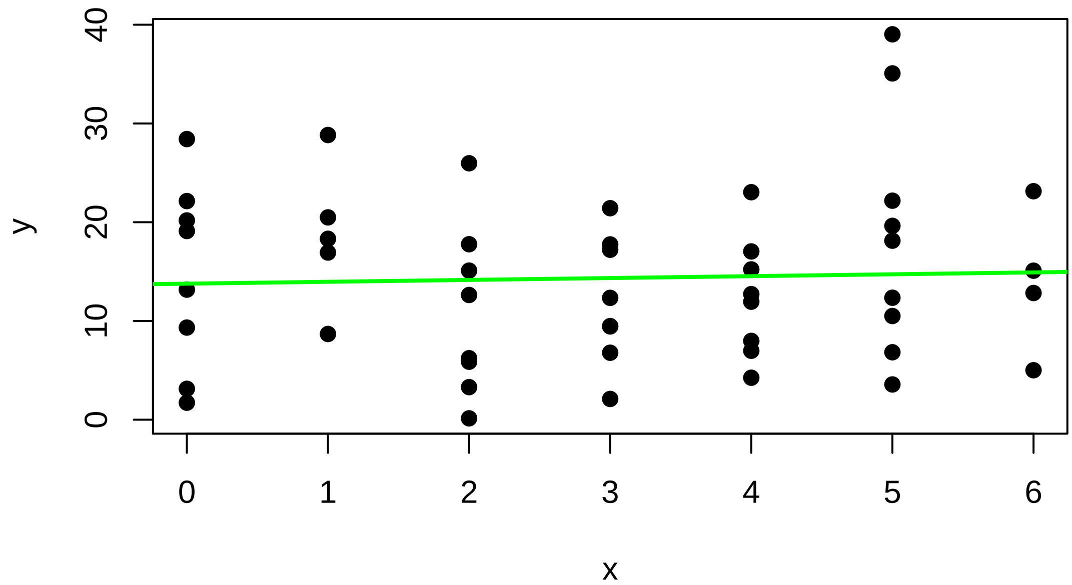
Пункт (а)

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.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \varepsilon_i - N(0, \sigma^2) \tag{1}$$

$$\beta_1 = \frac{\bar{xy} - \bar{x}\bar{y}}{\bar{x^2} - \bar{x}^2} = 0.1902404; \beta_0 = \bar{Y} - \beta_1 \bar{X} = 13.775098 \tag{2}$$

$$Y = \hat{\beta}_1 X + \beta_0 = 0.1902404X + 13.775098 \tag{3}$$



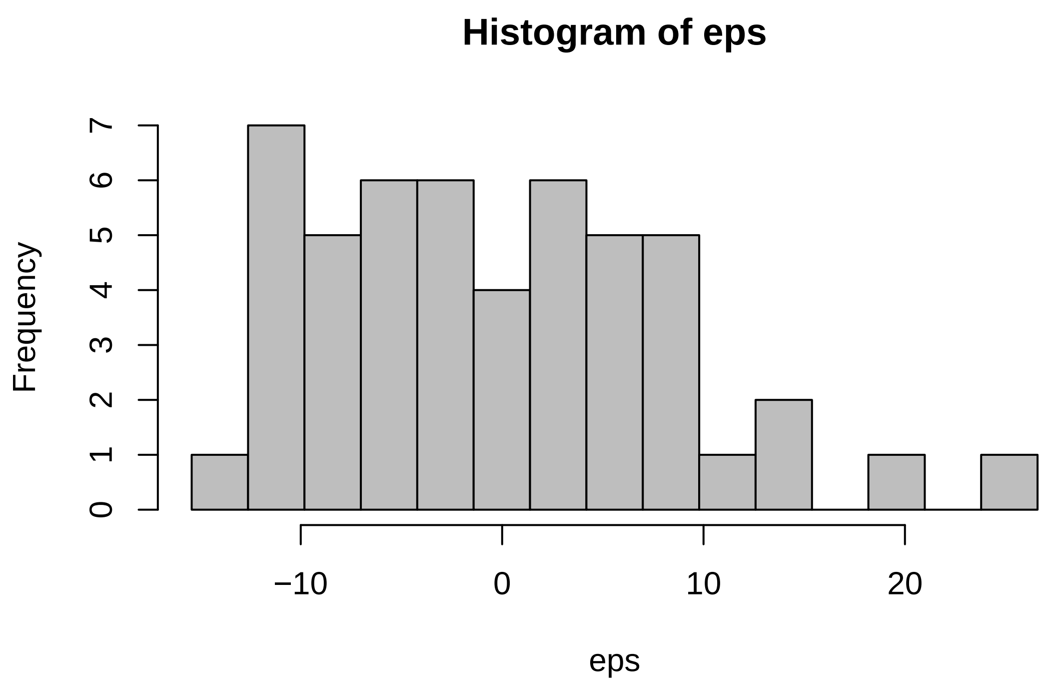
Пункт (b)

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.

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_1^n (Y-\bar{Y})^2 = 76.2468392; \varepsilon_i = Y_i - \hat{\beta}_1 X_i - \hat{\beta}_0 \tag{4}$$

$\varepsilon_1, ..., \varepsilon_n (sorted):$

```
##  [1] -14.0155788 -12.2458192 -12.0550980 -11.1563001 -10.8555788 -10.6450980
##  [7] -10.2860597  -9.9065405  -8.2855788  -7.9255788  -7.8963001  -7.5658192
## [13]  -7.5560597  -6.5560597  -5.2853384  -4.8958192  -4.8658192  -4.4450980
## [19]  -4.2263001  -2.5860597  -2.3763001  -2.0865405  -2.0058192  -1.8160597
## [25]  -1.5255788  -0.6050980   0.1734595   0.6839403   0.9444212   2.5039403
## [31]   2.8641808   2.9646616   3.4036999   3.4041808   3.6144212   4.3546616
## [37]   4.9036999   5.3349020   6.4049020   6.5246616   7.0741808   7.4536999
## [43]   8.2234595   8.3649020   8.5039403  11.8144212  14.6449020  14.8646616
## [49]  20.3436999  24.3036999
```



$H_0 : \varepsilon_1, ..., \varepsilon_n \sim N(0, \sigma^2)$;

Минимизация хи-квадрат: $\underset{\sigma^2}{argmin} \sum_{i=1}^{r} \frac{n_i - np_i(0,\sigma^2))^2}{np_i(0,\sigma^2)}$

Делим выборку на 6 интервалов

Интервал	$(-\infty; -10.0)$	$(-10.0; -5.0)$	$(-5.0; 0)$	$(0; 5.0)$	$(5.0; 10.0)$	$(10.0; \infty)$
Частота	7	8	11	11	8	5

```
r <- 6
P <- function(a){
  p<-0
  p[1] <- pnorm(-10.0, 0, a)
  p[2] <- pnorm(-5.0, 0, a) - sum(p)
  p[3] <- pnorm(0, 0, a) - sum(p)
  p[4] <- pnorm(5.0, 0, a) - sum(p)
  p[5] <- pnorm(10.0, 0, a) - sum(p)
  p[6] <- 1 - sum(p)
  p}
X2 <- function(a){g<-n*P(a); f<-(nu-g)^2/g;sum(f)}
nu <- c(7, 8, 11, 11, 8, 5)
a <- c(sqrt(s))
XM <- nlm(X2, a)
XM$estimate^2; XM$minimum
```

```
## [1] 73.39046
```

```
## [1] 0.3331803
```

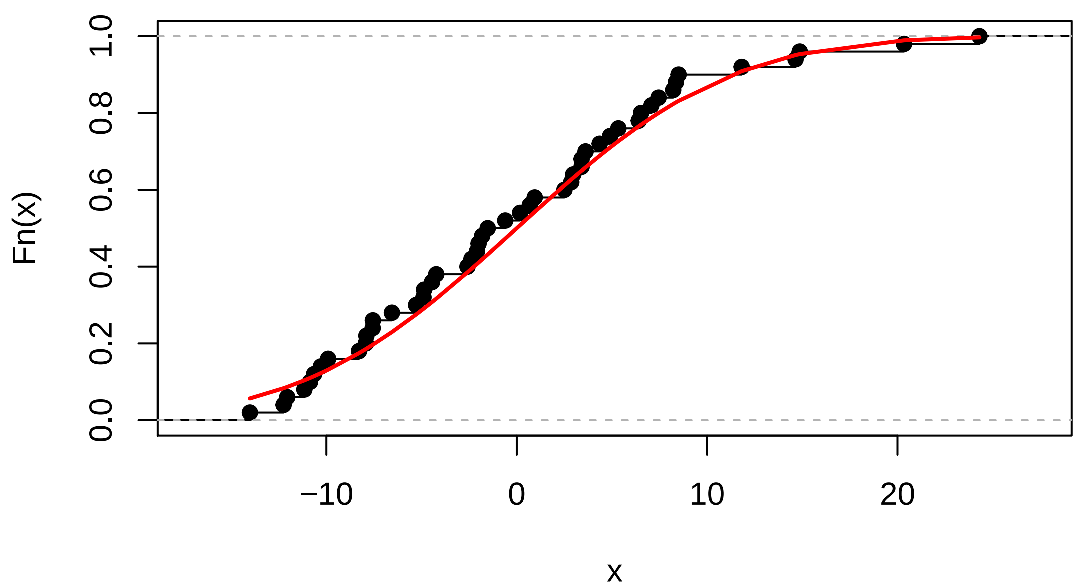
```
xa <- qchisq(1- alpha1, r-1-1)
```

Получили $\hat{\sigma}^2 = 73.3904593$ и $\chi^2 = 0.3331803$
$\chi^2 = 0.3331803 < \chi^2 = 7.7794403 \Rightarrow$ гипотеза H_0 принимается

Оценим расстояние оценки до класса нормальных распределений по Колмогорову.
Минимизируем статистику Колмогорова с помощью 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, XM$estimate)
```

Получили расстояние $D = 0.068403$ и $\tilde{\sigma}^2 = 78.3880132$.



Пункт (c)

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).

$\psi = C^T\beta$
$\hat{\psi} \sim N(\psi, \sigma^2 b), \sigma^2 b = var(\hat{\psi}) = \sigma^2 C^T (XX^T)^{-1} C$

$\frac{\hat{\psi}-\psi}{\sigma\sqrt{b}} \sim N(0,1); \frac{(n-r)s^2}{\sigma^2} \sim \chi^2_{n-r} \Rightarrow \frac{\hat{\psi}-\psi}{s\sqrt{b}} \sim S_{n-r}$
$x_\alpha : S_{n-r}(x_\alpha) = 1 - \frac{\alpha}{2}; x_\alpha = 1.6772242$
$P(-x_\alpha \leqslant \frac{\hat{\psi}-\psi}{s\sqrt{b}} \leqslant x_\alpha) = P(\hat{\psi} - x_\alpha s\sqrt{b} \leqslant \psi \leqslant \hat{\psi} + x_\alpha s\sqrt{b})$

$$\beta_0 \in (9.9714223; 17.5787737)$$
$$\beta_1 \in (-0.9023369; 1.2828177)$$

$\frac{(\hat{\psi}-\psi)^T(C^T(XX^T)^{-1}C)^{-1}(\hat{\psi}-\psi)}{\sigma^2} \sim \chi^2_q$

$x_\alpha : F_{q,n-r}(x_\alpha) = 1 - \alpha$
$\left\{\psi : (\hat{\psi} - \psi)^T(C^T(XX^T)^{-1}C)^{-1}(\hat{\psi} - \psi) \leqslant s^2 q x_\alpha\right\}$
$x_\alpha = 2.4166601$
$\begin{pmatrix}\beta_1 - \hat{\beta}_1 & \beta_0 - \hat{\beta}_0\end{pmatrix}\begin{pmatrix}0.0055654 & -0.0162511\\ -0.0162511 & 0.0674533\end{pmatrix}\begin{pmatrix}\beta_1 - \hat{\beta}_1\\ \beta_0 - \hat{\beta}_0\end{pmatrix} \leqslant 368.5253897$

Собственные числа матрицы: 0.0714611, 0.0015576
После ортогонального преобразования получаем:

$0.0714611(\beta_1^* - 0.1902404)^2 + 0.0015576(\beta_0^* + 13.775098)^2 \leqslant 368.5253897$

$$\frac{(\beta_1^* - 0.1902404)^2}{71.8123122^2} + \frac{(\beta_0^* + 13.775098)^2}{486.4109861^2} \leqslant 1$$

Полуоси эллипса: 71.8123122; 486.4109861.

Пункт (d)

Сформулировать гипотезу независимости переменной Y от переменной X. Провести проверку значимости.
$\psi = C^T\beta; \hat{\psi} \sim N(\psi, \sigma^2 C^T(XX^T)^{-1}C)$
$H_0 : \psi = 0$
Статистика F-критерия: $F = \frac{\psi^T(C^T(X\hat{X^T})^{-1}C)^{-1}\hat{\psi}}{qs^2} \overset{H_0}{\sim} F_{q,n-r}$
$F = 0.0852871$
$x_\alpha : F_{q,n-r}(x_\alpha) = 1 - \alpha; x_\alpha = 2.813081$
$H_1 : \hat{\psi} = \beta_1$

$F < x_\alpha \Rightarrow$ Принимаем гипотезу H_0.
0.7715149 – наибольшее значение уровня значимости, на котором нет оснований отвергнуть данную гипотезу.

Пункт (e)

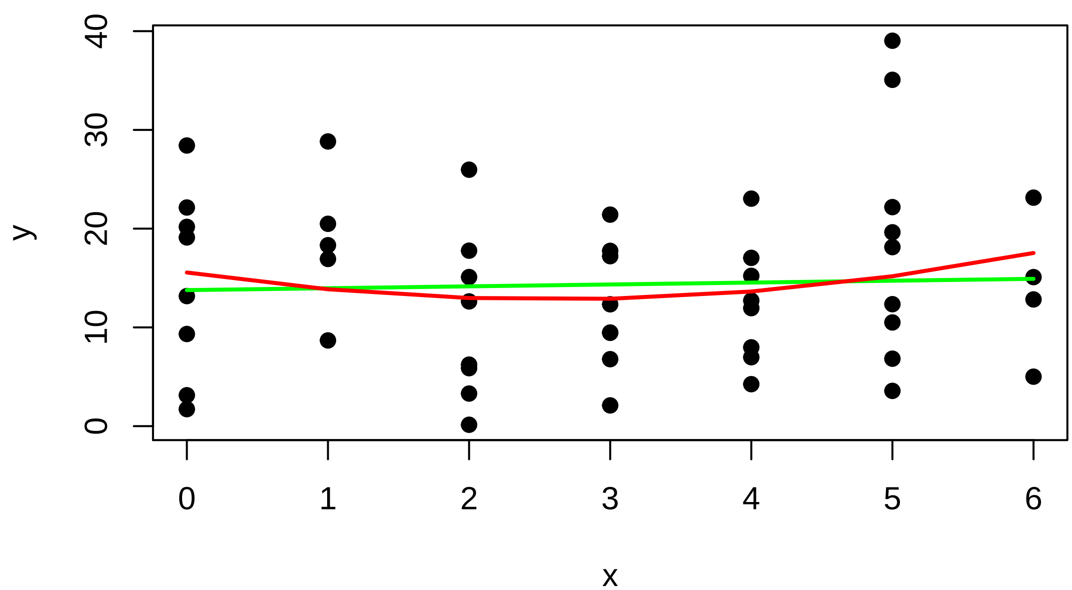
Сформулировать модель, включающую дополнительный член с X^2. Построить МНК оценки параметров $\beta_1, \beta_2, \beta_3$ в данной модели. Изобразить графически полученную регрессионную

зависимость.

$$Y_i = \beta_3 + \beta_2 X_i + \beta_1 X_i^2 + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2) \tag{5}$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{6}$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \begin{pmatrix} 0.405228 \\ -2.1022026 \\ 15.5576681 \end{pmatrix} \tag{7}$$

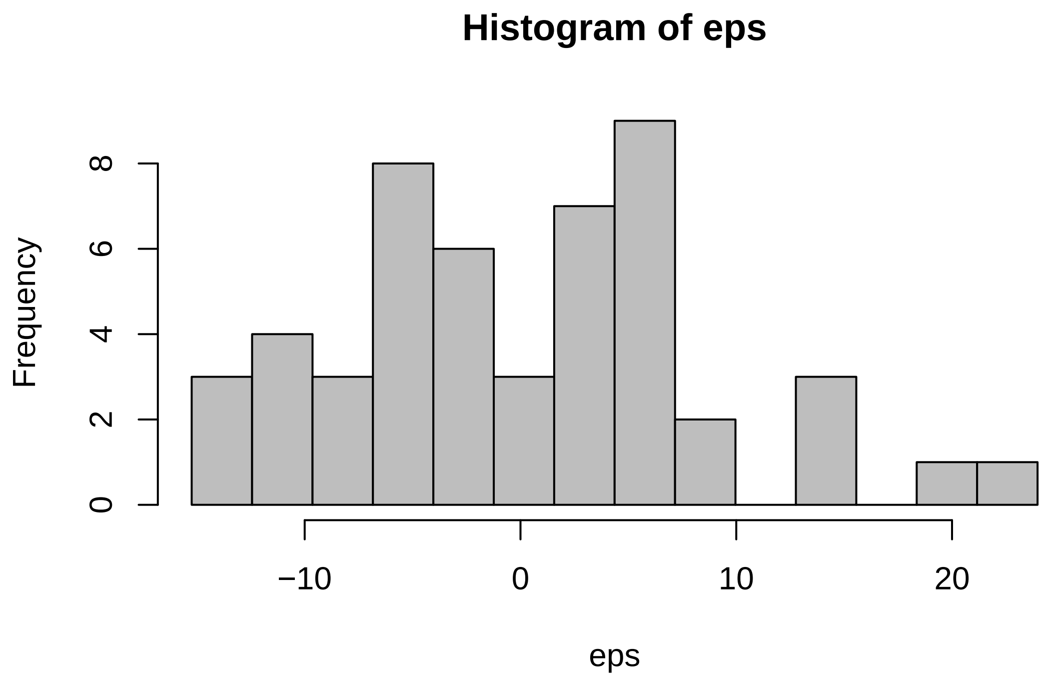


Пункт (f)

Построить несмещенную оценку дисперсии. Провести исследование нормальности ошибок как в п.3.
$\hat{\sigma}^2 = \frac{1}{n-3}\sum\limits_1^n (Y_i - \hat{\beta}_1 X_i^2 - \hat{\beta}_2 X_i - \hat{\beta}_3)^2 = 75.8340207; \varepsilon_i = \hat{Y} - X^T\hat{\beta}$

$\varepsilon_1, ..., \varepsilon_n (sorted):$

```
##  [1] -13.8376681 -12.8341750 -12.5226607 -12.4276681 -11.6073553 -10.7981124
##  [7]  -9.6741750  -9.3825059  -8.3473553  -7.1041750  -6.7441750  -6.6525059
## [13]  -6.2276681  -6.1181124  -5.6525059  -5.1806936  -4.7026607  -4.6773553
## [19]  -3.4481124  -3.4181124  -2.8273553  -2.4426607  -2.3876681  -1.6825059
## [25]  -0.9125059  -0.5581124  -0.3441750   1.5874941   2.1258250   2.9526447
## [31]   3.0693064   3.4074941   3.5523319   4.3118876   4.4526447   4.4593064
## [37]   4.6223319   4.7958250   4.8518876   5.6073393   6.5823319   6.6293064
## [43]   7.0026447   8.5218876   9.4074941  12.8623319  12.9958250  14.9693064
## [49]  19.8926447  23.8526447
```



$H_0 : \varepsilon_1, ..., \varepsilon_n \sim N(0, \sigma^2)$

Минимизация хи-квадрат: $\underset{\sigma^2}{argmin} \sum\limits_{i=1}^{r} \frac{n_i - np_i(0,\sigma^2))^2}{np_i(0,\sigma^2)}$

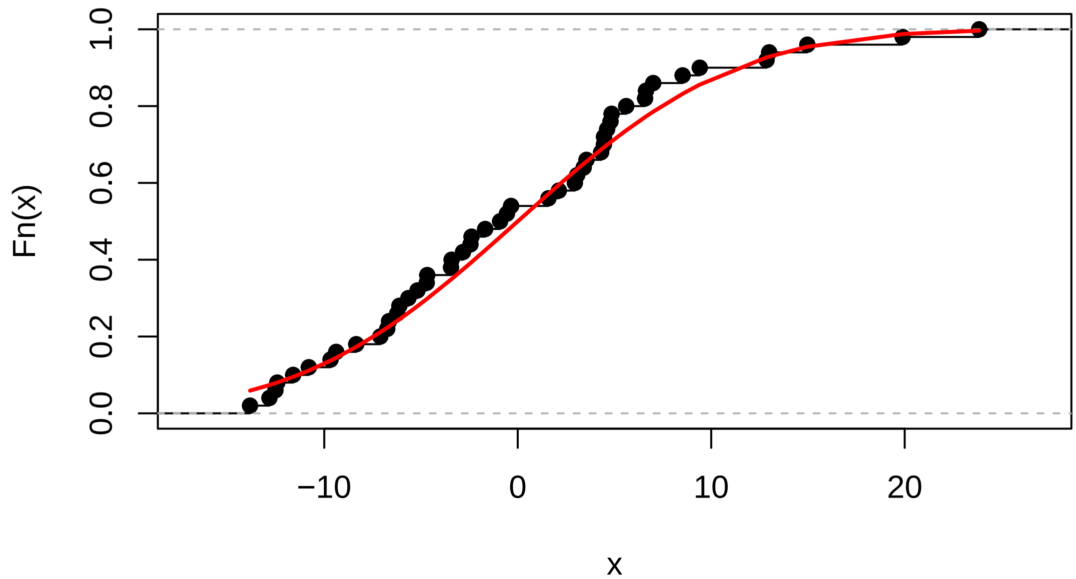
Делим выборку на 6 интервалов

Интервал	$(-\infty; -10.0)$	$(-10.0; -5.0)$	$(-5.0; 0)$	$(0; 5.0)$	$(5.0; 10.0)$	$(10.0; \infty)$
Частота	6	10	11	12	6	5

Получили $\hat{\sigma}^2 = 66.4715803$ и $\chi^2 = 1.1353156$
$\chi^2 = 1.1353156 < \chi^2 = 7.7794403 \Rightarrow$ гипотеза H_0 принимается

Оценим расстояние оценки до класса нормальных распределений по Колмогорову.
Минимизируем статистику Колмогорова.

Получили расстояние $D = 0.068403$ и $\tilde{\sigma}^2 = 78.3880132$.



Пункт (g)

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 $1 - \alpha_1$.
$\psi = C^T\beta; \hat{\psi} \sim N(\psi, \sigma^2 C^T (XX^T)^{-1} C)$
$x_\alpha : S_{n-r}(x_\alpha) = 1 - \frac{\alpha}{2}; x_\alpha = 1.6779267$
$P(\hat{\psi} - x_\alpha s\sqrt{b} \leqslant \psi \leqslant \hat{\psi} + x_\alpha s\sqrt{b})$

$$\beta_1 \in (-0.2218946; 1.0323506)$$
$$\beta_2 \in (-5.8136314; 1.6092263)$$
$$\beta_3 \in (10.8659807; 20.2493555)$$

$x_\alpha : F_{q,n-r}(x_\alpha) = 1 - \alpha; x_\alpha = 2.2041824$
$\left\{\psi : (\hat{\psi} - \psi)^T (C^T (XX^T)^{-1} C)^{-1} (\hat{\psi} - \psi) \leqslant s^2 q x_\alpha\right\}$
$\begin{pmatrix} \beta_1 - \hat{\beta}_1 & \beta_2 - \hat{\beta}_2 & \beta_3 - \hat{\beta}_3 \end{pmatrix} \begin{pmatrix} 0.001842 & -0.0104206 & 0.0081029 \\ -0.0104206 & 0.0645167 & -0.0620908 \\ 0.0081029 & -0.0620908 & 0.1030975 \end{pmatrix} \begin{pmatrix} \beta_1 - \hat{\beta}_1 \\ \beta_2 - \hat{\beta}_2 \\ \beta_3 - \hat{\beta}_3 \end{pmatrix} \leqslant 501.4560391$

Собственные числа матрицы: $0.1499161, 0.0194699, 7.0244607 \times 10^{-5}$
После ортогонального преобразования получаем:

$0.1499161(\beta_1^* - 0.405228)^2 + 0.0194699(\beta_2^* - (-2.1022026))^2 + 7.0244607 \times 10^{-5}(\beta_3^* - 15.5576681)^2 \leqslant 501.4560391$

$$\frac{(\beta_1^* - 0.405228)^2}{57.8352148^2} + \frac{(\beta_2^* - (-2.1022026))^2}{160.4849927^2} + \frac{(\beta_3^* - 15.5576681)^2}{2671.8368734^2} \leqslant 1$$

Полуоси эллипсоида: 57.8352148; 160.4849927; 2671.8368734.

Пункт (h)

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

$\psi = C^T\beta; \hat{\psi} \sim N(\psi, \sigma^2 C^T (XX^T)^{-1} C)$
$H_0 : \psi = 0$
Статистика F-критерия:
$F = \frac{\psi^T (C^T (\hat{XX^T})^{-1} C)^{-1} \hat{\psi}}{qs^2} \overset{H_0}{\sim} F_{q,n-r}$
$F = 1.1755467$
$x_\alpha : F_{q,n-r}(x_\alpha) = 1 - \alpha; x_\alpha = 2.8154381$
$H_1 : \hat{\psi} = \beta_1$

$F < x_\alpha \Rightarrow$ Принимаем гипотезу H_0.
0.283795 – наибольшее значение уровня значимости, на котором нет оснований отвергнуть данную гипотезу.