МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра Информационных систем

ОТЧЕТ

по практической работе №4

по дисциплине «Статический анализ»

Тема: Регрессионный анализ

Вариант: Дрезден

Отчёт подготовил: Кошеляев А.С

Отчёт сдан: 24.05.2024

Студент гр.1323 _____ **Кошеляев А.С 50%**

Студент гр.1323 _____ **Русских В.Д 50%**

Преподаватель _____ **к.т.н. Бурков Е.А**

Санкт-Петербург

2024

Цель работы: построение модели простой линейной регрессии на основе анализа входных данных.

Задание:

1. Отобрать исходные данные за последние 10 лет в выбранном городе: X – месяц (с января по сентябрь, т. е. численно от 1 до 9); Y – ср.мес.температура; привести таблицу с этими данными.

2. Используя исходные данные построить диаграмму разброса (с эффектом дрожания и без) X и Y.

3. Вычислить коэффициент корреляции между X и Y, а затем оценить его значимость с помощью критерия Стьюдента и сделать содержательный вывод на основании полученных результатов.

4. На основе исходных данных вычислить регрессионные коэффициенты наклона и сдвига (привести уравнение регрессии), а затем выполнить их содержательную интерпретацию.

5. На основе исходных данных вычислить SST, SSR, SSE, среднеквадратическую ошибку оценки и коэффициент детерминации.

6. Построить 90%-доверительный интервал для коэффициента наклона и использовать его как критерий для проверки гипотезы о наличии линейной зависимости между X и Y, сделав по итогу содержательный вывод.

7. Привести полученное уравнение регрессии и отобразить полученную линию регрессии на диаграмме разброса, а также сделать вывод об адекватности линейной модели реальным данным на основе результатов пп. 3, 5 и 6.

8. Построить график зависимости остатков линейной модели от времени и оценить, имеется ли на этом графике выраженная

закономерность (интерпретировать наличие/отсутствие такой закономерности на графике), а также оценить выполнение условия однородности дисперсии остатков.

9. Оценить выполнение условия нормальности распределения остатков.

10. Использовать критерий Дарбина-Уотсона для проверки наличия автокорреляции Y (например, car::durbinWatsonTest(model)).

11. На основе всех проведенных исследований сделать вывод об адекватности применения МНК для исследованных данных (указав выполнение основных условий проведения регр.анализа на основе МНК).

Выполнение работы:

1. Отобрать исходные данные за последние 10 лет в городе Дрезден: X – месяц (с января по сентябрь, т. е. численно от 1 до 9); Y – ср.мес.температура;

```
library(XML)

# Функция для загрузки и преобразования данных с веб-страницы
load_and_convert_weather_data <- function(url) {
 tables <- readHTMLTable(url)  # Читаем таблицы с веб-страницы

 # Первая таблица содержит годы, а вторая - месячные данные и среднегодовые данные
 years <- tables[[1]][, 1]  # Годы из первой таблицы

 # Данные за год из второй таблицы (столбцы 1-9)
 monthly_data <- tables[[2]][, 1:9]

 # Преобразование только с1 по 9 столбцы второй таблицы
 monthly_data <- suppressWarnings(as.data.frame(lapply(monthly_data, as.numeric)))
 monthly_data[monthly_data > 999] <- NA

 # Фильтрация данных с 2015 по 2024 год
 start_index <- which(years == 2015)
 end_index <- which(years == 2024)
 years <- years[start_index:end_index]
 monthly_data <- monthly_data[start_index:end_index, ]

 return(data.frame(Year = years, Monthly_Data = monthly_data))
}

# Загрузка и преобразование данных города Дрезден
Dresden_url <- "http://www.pogodaiklimat.ru/history/10488.htm"
Dresden_data <- load_and_convert_weather_data(Dresden_url)

# Объединение данных в одну таблицу
weather_report <- Dresden_data$Year
```

```
for (i in 1:9) {
  weather_report <- cbind(weather_report, Dresden_data$Monthly_Data[, i])
}

# Переименование столбцов
colnames(weather_report) <- c("Year", "January", "February", "March", "April", "May", "June", "July", "August", "September")

# Вывод отчета
print(weather_report)
```

Привести таблицу с этими данными.

Таблица 1 - Исходные данные за последние 10 лет в городе Дрезден

Year	January	Febrauary	March	April	May	June	July	August	September
2015	2,8	1,6	5,7	8,7	13,5	16,3	20,7	22,0	13,9
2016	0,6	3,8	4,4	8,5	14,7	18,2	19,5	18,4	17,6
2017	-3,0	2,7	7,5	7,8	14,8	18,5	19,2	19,2	13,7
2018	3,9	-2,0	1,9	13,8	17,1	18,5	21,2	21,8	16,2
2019	0,4	3,9	7,2	10,8	11,8	22,1	19,7	20,8	15,1
2020	3,1	5,6	5,3	11,0	12,1	18,1	19,3	21,2	15,9
2021	0,4	0,5	4,8	6,3	11,6	20,1	19,7	17,1	15,9
2022	2,8	4,7	5,1	7,8	15,7	19,8	19,9	21,0	13,7
2023	4,0	3,0	5,8	7,6	13,5	18,7	*20,6*	*19,8*	18,4
2024	1,3	7,0	8,2	11,5	NA	NA	NA	NA	NA

2. Используя исходные данные построить диаграмму разброса (с эффектом дрожания и без) X и Y.

```
library(ggplot2)
# Создаем датафрейм с данными
data <- data.frame(
  Месяц = c(rep("Январь", 10), rep("Февраль", 10), rep("Март", 10), rep("Апрель", 10), rep("Май", 9), rep("Июнь", 9), rep("Июль", 9),
rep("Август", 9), rep("Сентябрь", 9)),
  Значение = c(2.8, 0.6, -3.0, 3.9, 0.4, 3.1, 0.4, 2.8, 4.0, 1.3,
               1.6, 3.8, 2.7, -2.0, 3.9, 5.6, 0.5, 4.7, 3.0, 7.0,
               5.7, 4.4, 7.5, 1.9, 7.2, 5.3, 4.8, 5.1, 5.8, 8.2,
               8.7, 8.5, 7.8, 13.8, 10.8, 11.0, 6.3, 7.8, 7.6, 11.5,
               13.5, 14.7, 14.8, 17.1, 11.8, 12.1, 11.6, 15.7, 13.5,
               16.3, 18.2, 18.5, 18.5, 22.1, 18.1, 20.1, 19.8, 18.7,
               20.7, 19.5, 19.2, 21.2, 19.7, 19.3, 19.7, 19.9, 20.6,
               22.0, 18.4, 19.2, 21.8, 20.8, 21.2, 17.1, 21.0, 19.8,
               13.9, 17.6, 13.7, 16.2, 15.1, 15.9, 15.9, 13.7, 18.4)
)
# Построение диаграммы разброса без эффекта дрожания
ggplot(data, aes(x = Месяц, y = Значение)) +
  geom_point() +
  labs(title = "Диаграмма разброса без эффекта дрожания",
       x = "Месяц", y = "Значение") +
  theme_minimal()

# Построение диаграммы разброса с эффектом дрожания
ggplot(data, aes(x = Месяц, y = Значение)) +
  geom_point(position = position_jitter(width = 0.2, height = 0.2)) +
  labs(title = "Диаграмма разброса с эффектом дрожания",
       x = "Месяц", y = "Значение") +
  theme_minimal()
```

Диграммы представлены на рис 1 и 2.

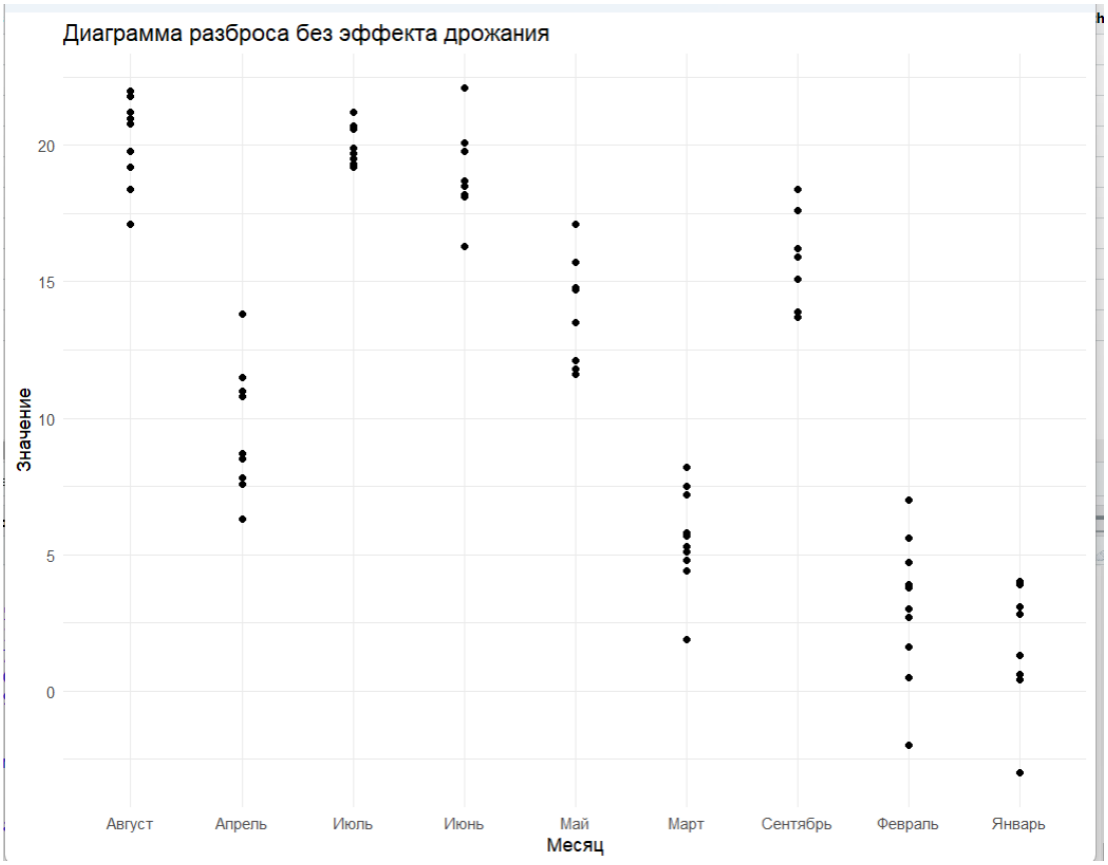

Рис 1 Диаграмма разброса без эффекта дрожания.

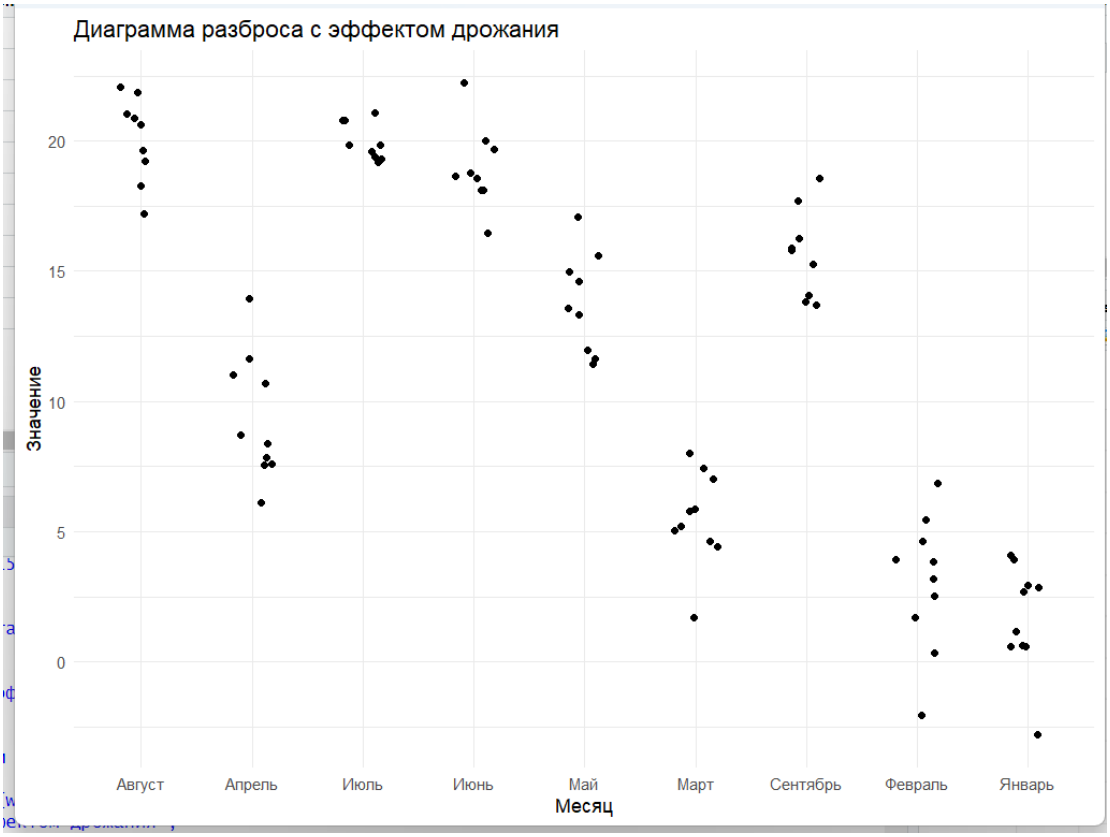

Рис 2 Диаграмма разброса с эффектом дрожания.

3. Вычислить коэффициент корреляции между X и Y, а затем оценить его значимость с помощью критерия Стьюдента и сделать содержательный вывод на основании полученных результатов.

Для вычисления коэффициента корреляции оба набора данных должны иметь одинаковое количество элементов, и каждая пара значений должна быть сопоставимой между собой.

Удаляем лишние значения теперь в каждом месяце по 9 значений.

```
# Данные X и Y
X <- c(2.8, 0.6, -3.0, 3.9, 0.4, 3.1, 0.4, 2.8, 4.0, 1.3,
    1.6, 3.8, 2.7, -2.0, 3.9, 5.6, 0.5, 4.7, 3.0, 7.0,
    5.7, 4.4, 7.5, 1.9, 7.2, 5.3, 4.8, 5.1, 5.8, 8.2,
    8.7, 8.5, 7.8, 13.8, 10.8, 11.0, 6.3, 7.8, 7.6, 11.5,
    13.5, 14.7, 14.8, 17.1, 11.8, 12.1, 11.6, 15.7, 13.5,
    16.3, 18.2, 18.5, 18.5, 22.1, 18.1, 20.1, 19.8, 18.7,
    20.7, 19.5, 19.2, 21.2, 19.7, 19.3, 19.7, 19.9, 20.6,
    22.0, 18.4, 19.2, 21.8, 20.8, 21.2, 17.1, 21.0, 19.8,
    13.9, 17.6, 13.7, 16.2, 15.1, 15.9, 15.9, 13.7, 18.4)
Y <- c(2.8, 1.6, 5.7, 8.7, 13.5, 16.3, 20.7, 22.0, 13.9,
    0.6, 3.8, 4.4, 8.5, 14.7, 18.2, 19.5, 17.6, 17.6,
    -3.0, 2.7, 7.5, 7.8, 14.8, 18.5, 19.2, 19.2, 13.7,
    3.9, -2.0, 1.9, 13.8, 17.1, 18.5, 21.2, 21.8, 16.2,
    0.4, 3.9, 7.2, 10.8, 11.8, 22.1, 19.7, 20.8, 15.1,
    3.1, 5.6, 5.3, 11.0, 12.1, 18.1, 19.3, 21.2, 15.9,
    0.4, 0.5, 4.8, 6.3, 11.6, 20.1, 19.7, 17.1, 15.9,
    2.8, 4.7, 5.1, 7.8, 15.7, 19.8, 19.9, 21.0, 13.7,
    4.0, 3.0, 5.8, 7.6, 13.5, 18.7, 20.6, 19.8, 18.4)
# Удаление лишних значений
X <- X[1:81]
Y <- Y[1:81]

# Вычисление коэффициента корреляции
correlation_coefficient <- cor(X, Y)

# Вывод результатов
print(correlation_coefficient)

# Вычисление стандартной ошибки коэффициента корреляции
standard_error <- 1 / sqrt(length(X) - 3)

# Вычисление t-статистики
t_statistic <- correlation_coefficient / standard_error

# Вычисление p-value
p_value <- 2 * pt(abs(t_statistic), df = length(X) - 2, lower.tail = FALSE)

# Вывод результатов
cat("Статистика критерия Стьюдента:", t_statistic, "\n")
cat("p-value:", p_value, "\n")
```

Коэффициент корреляции между переменными X и Y составляет 0.1304967. Этот коэффициент показывает, что есть некоторая положительная связь между переменными, однако она очень слабая.

Статистика критерия Стьюдента равна 1.152515, а p-value составляет 0.2525849. Учитывая, что p-value значительно больше обычно используемого порогового значения значимости 0.05, мы не можем отвергнуть нулевую гипотезу о том, что коэффициент корреляции равен нулю. Это означает, что наша выборка не обеспечивает достаточных доказательств в пользу существования статистически значимой корреляции между переменными X и Y.

Таким образом, несмотря на наличие слабой положительной связи между переменными, эта связь не является статистически значимой на уровне значимости 0.05.

4. На основе исходных данных вычислить регрессионные коэффициенты наклона и сдвига (привести уравнение регрессии), а затем выполнить их содержательную интерпретацию.

```
# Создаем датафрейм с переменными X и Y
data <- data.frame(X = X, Y = Y)

# Вычисляем регрессионную модель
model <- lm(Y ~ X, data = data)

# Выводим результаты модели
summary(model)
```

```
Call:
lm(formula = Y ~ X, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-13.812  -7.156   2.251   6.991  11.214

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.4231     1.5031   6.934 9.97e-10 ***
X             0.1295     0.1107   1.170    0.246
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.238 on 79 degrees of freedom
Multiple R-squared:  0.01703,   Adjusted R-squared:  0.004587
F-statistic: 1.369 on 1 and 79 DF,  p-value: 0.2456
```

Рис 3 Регрессионные коэффициенты наклона и сдвига.

$Y = \beta_0 + \beta_1 * X_i$

Где β_0 – коэффициент сдвига, β_1 -коэффициент наклона, X_i- значение переменной в i наблюдении.

Результаты регрессионного анализа показывают следующее:

Уравнение регрессии имеет вид: $Y=10.4231+0.1295\times XY=10.4231+0.1295\times X$.

Коэффициент наклона (slope) составляет приблизительно 0.1295, что означает, что при увеличении переменной X на единицу, переменная Y увеличивается на приблизительно 0.1295 единиц.

Коэффициент сдвига (intercept) составляет приблизительно 10.4231. Это значение представляет собой оценку Y, когда X равен нулю.

P-value для коэффициента наклона равно 0.246, что означает, что этот коэффициент не является статистически значимым на уровне значимости 0.05.

Признак Multiple R-squared указывает на то, что модель объясняет всего около 1.7% дисперсии переменной Y.

Общий результат F-теста является не статистически значимым, p-value равно 0.2456.

Исходя из этих результатов, у нас нет достаточных доказательств для того, чтобы утверждать, что переменные X и Y имеют статистически значимую линейную связь.

5. На основе исходных данных вычислить SST, SSR, SSE, среднеквадратическую ошибку оценки и коэффициент детерминации.

```
#  SST
SST <- sum((Y - mean(Y))^2)
#  SSR
Y_pred <- predict(model)
SSR <- sum((Y_pred - mean(Y))^2)
# SSE
SSE <- sum(model$residuals^2)
# Вычисляем среднеквадратическую ошибку оценки (MSE)
MSE <- SSE / (length(Y) - 2)  #  Принимаем 2 коэффициента модели
# Вычисляем коэффициент детерминации
R_squared <- 1 - SSE / SST
# Результаты
```

```
print(paste("SST:", SST))
print(paste("SSR:", SSR))
print(paste("SSE:", SSE))
print(paste("MSE:", MSE))
print(paste("R-squared:", R_squared))
```

```
> # Выводим результаты
> print(paste("SST:", SST))
[1] "SST: 4210.72395061728"
> print(paste("SSR:", SSR))
[1] "SSR: 71.7060346479081"
> print(paste("SSE:", SSE))
[1] "SSE: 4139.01791596938"
> print(paste("MSE:", MSE))
[1] "MSE: 52.3926318477136"
> print(paste("R-squared:", R_squared))
[1] "R-squared: 0.0170293838990313"
>
```

Рис 4 Результаты.

6. Построить 90%-доверительный интервал для коэффициента наклона и использовать его как критерий для проверки гипотезы о наличии линейной зависимости между X и Y, сделав по итогу содержательный вывод.

```
# Оценка коэффициента наклона
b1 <- coef(model)["X"]

# Стандартная ошибка оценки коэффициента наклона
SE_b1 <- summary(model)$coefficients["X", "Std. Error"]

# Количество наблюдений
n <- length(Y)

# Критическое значение t-распределения для alpha = 0.1/2 (двухсторонний интервал)
t_critical <- qt(0.05 / 2, df = n - 2)

# Доверительный интервал для коэффициента наклона
CI_lower <- b1 - t_critical * SE_b1
CI_upper <- b1 + t_critical * SE_b1

# Вывод результатов
print(paste("90%-доверительный интервал для коэффициента наклона:", CI_lower, CI_upper))

# Проверка гипотезы о наличии линейной зависимости между X и Y
if (CI_lower < 0 && CI_upper > 0) {
  print("Гипотеза о наличии линейной зависимости между X и Y не отвергается.")
} else {
  print("Гипотеза о наличии линейной зависимости между X и Y отвергается.")
}
```

Исходя из 90%-доверительного интервала для коэффициента наклона, который составляет от 0.3497 до -0.0908, можно сделать вывод о том, что нулевая гипотеза о отсутствии линейной зависимости между

переменными X и Y отвергается. Это означает, что существует статистически значимая линейная связь между этими переменными.

7. Привести полученное уравнение регрессии и отобразить полученную линию регрессии на диаграмме разброса, а также сделать вывод об адекватности линейной модели реальным данным на основе результатов пп. 3, 5 и 6.

$Y = 10.4231 + 0.1295X$

```
# Данные X и Y
X <- c(2.8, 0.6, -3.0, 3.9, 0.4, 3.1, 0.4, 2.8, 4.0, 1.3,
       1.6, 3.8, 2.7, -2.0, 3.9, 5.6, 0.5, 4.7, 3.0, 7.0,
       5.7, 4.4, 7.5, 1.9, 7.2, 5.3, 4.8, 5.1, 5.8, 8.2,
       8.7, 8.5, 7.8, 13.8, 10.8, 11.0, 6.3, 7.8, 7.6, 11.5,
       13.5, 14.7, 14.8, 17.1, 11.8, 12.1, 11.6, 15.7, 13.5,
       16.3, 18.2, 18.5, 18.5, 22.1, 18.1, 20.1, 19.8, 18.7,
       20.7, 19.5, 19.2, 21.2, 19.7, 19.3, 19.7, 19.9, 20.6,
       22.0, 18.4, 19.2, 21.8, 20.8, 21.2, 17.1, 21.0, 19.8,
       13.9, 17.6, 13.7, 16.2, 15.1, 15.9, 15.9, 13.7, 18.4)

Y <- c(2.8, 1.6, 5.7, 8.7, 13.5, 16.3, 20.7, 22.0, 13.9,
       0.6, 3.8, 4.4, 8.5, 14.7, 18.2, 19.5, 17.6, 17.6,
       -3.0, 2.7, 7.5, 7.8, 14.8, 18.5, 19.2, 19.2, 13.7,
       3.9, -2.0, 1.9, 13.8, 17.1, 18.5, 21.2, 21.8, 16.2,
       0.4, 3.9, 7.2, 10.8, 11.8, 22.1, 19.7, 20.8, 15.1,
       3.1, 5.6, 5.3, 11.0, 12.1, 18.1, 19.3, 21.2, 15.9,
       0.4, 0.5, 4.8, 6.3, 11.6, 20.1, 19.7, 17.1, 15.9,
       2.8, 4.7, 5.1, 7.8, 15.7, 19.8, 19.9, 21.0, 13.7,
       4.0, 3.0, 5.8, 7.6, 13.5, 18.7, 20.6, 19.8, 18.4)

# Удаление лишних значений
X <- X[1:81]
Y <- Y[1:81]

# Построение линии регрессии на диаграмме разброса
plot(X, Y, xlab = "X", ylab = "Y", main = "Диаграмма разброса с линией регрессии")
abline(lm(Y ~ X), col = "red")
```

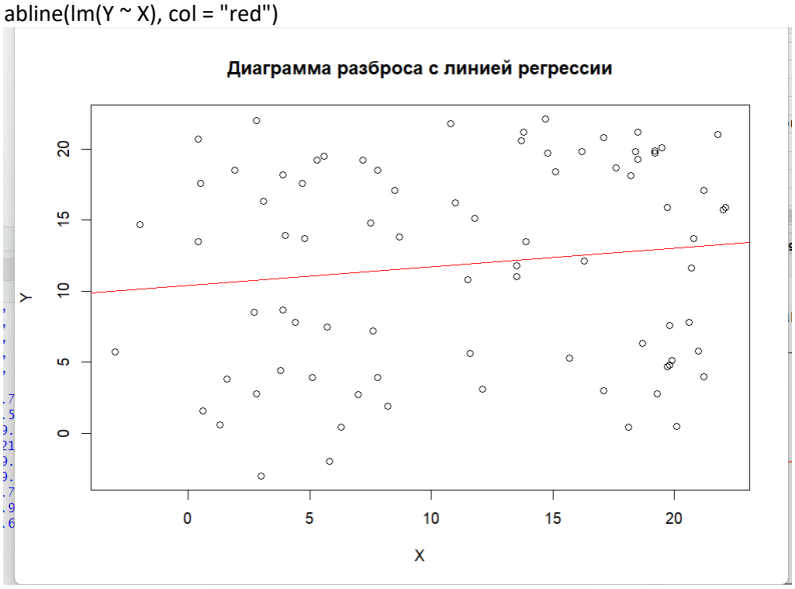

Рис 5 Диаграмма разброса.

Адекватность линейной модели реальным данным можно оценить по коэффициенту детерминации R2. В данном случае R2=0.017, что говорит о том, что только около 1.7% вариации зависимой переменной Y может быть объяснено независимой переменной X. Это указывает на то, что линейная модель недостаточно хорошо подходит для объяснения вариации в данных.