Домашнее задание №3

datasalad

31/07/2017

Дано X и Y (X общий, а зависимая переменная у каждого подписана). Необходимо построить диаграмму рассеяния (график Y от X), гистограмму распределения Y. Также нужно визуально оценить вид функциональной зависимости и выбрать одну (можно и не одну) из предложенных моделей. Для выбранной модели создать столбец расчитанных Y для заданных X. Найти разницу между заданным Y и расчитанным (ошибку модели), построить график изменения ошибки от X.

Загрузка набора данных в R

```
setwd("~/Desktop/dz/visualization")
dataset <- read.csv("dz.csv", header = TRUE, stringsAsFactors = FALSE)

str(dataset)
```

```
## 'data.frame':    60 obs. of  2 variables:
##  $ X: num  0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 ...
##  $ Y: num  -30.1 -38.2 -33.5 -27.8 -27.2 ...
```

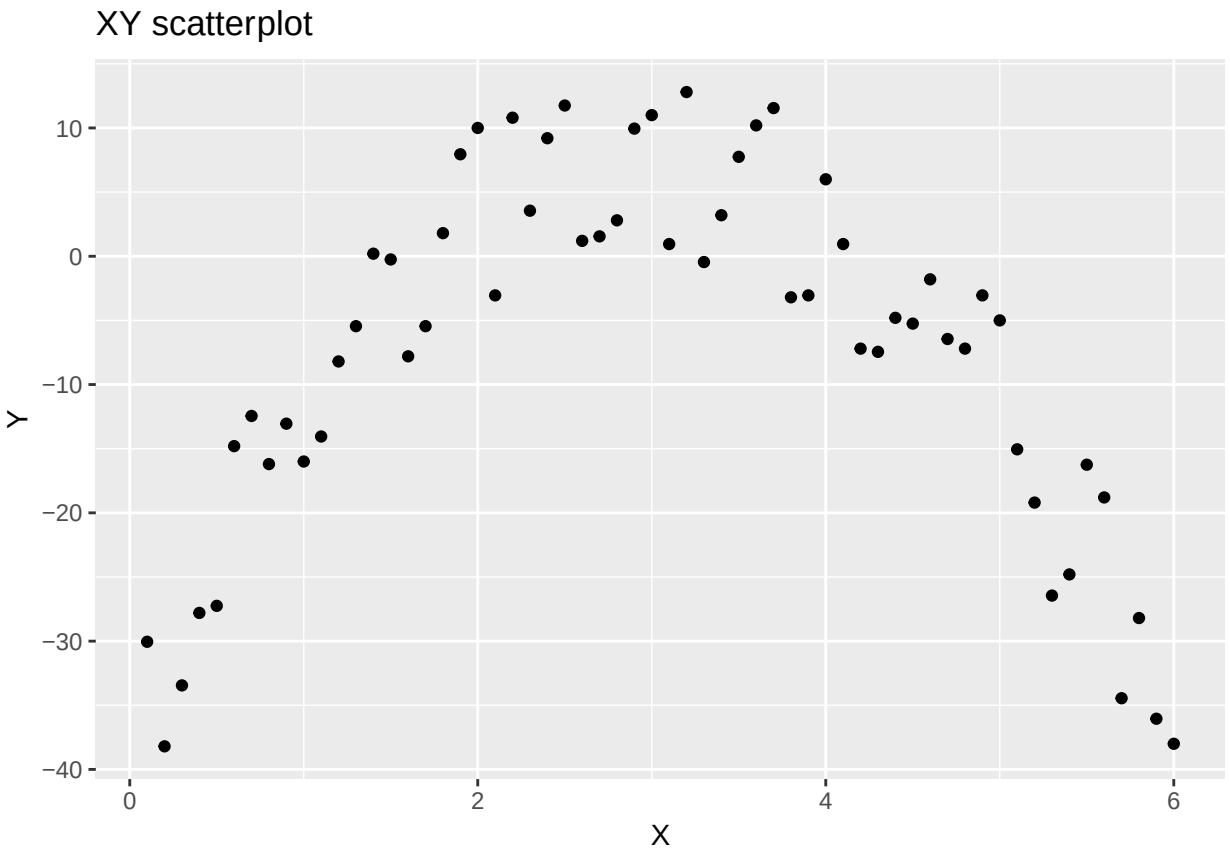
```
summary(dataset)
```

```
##        X                Y         
##  Min.   :0.100    Min.   :-38.200  
##  1st Qu.:1.575    1st Qu.:-16.050  
##  Median :3.050    Median : -5.125  
##  Mean   :3.050    Mean   : -7.508  
##  3rd Qu.:4.525    3rd Qu.:  2.050  
##  Max.   :6.000    Max.   : 12.800
```

Диаграмма рассеяния (график Y от X)

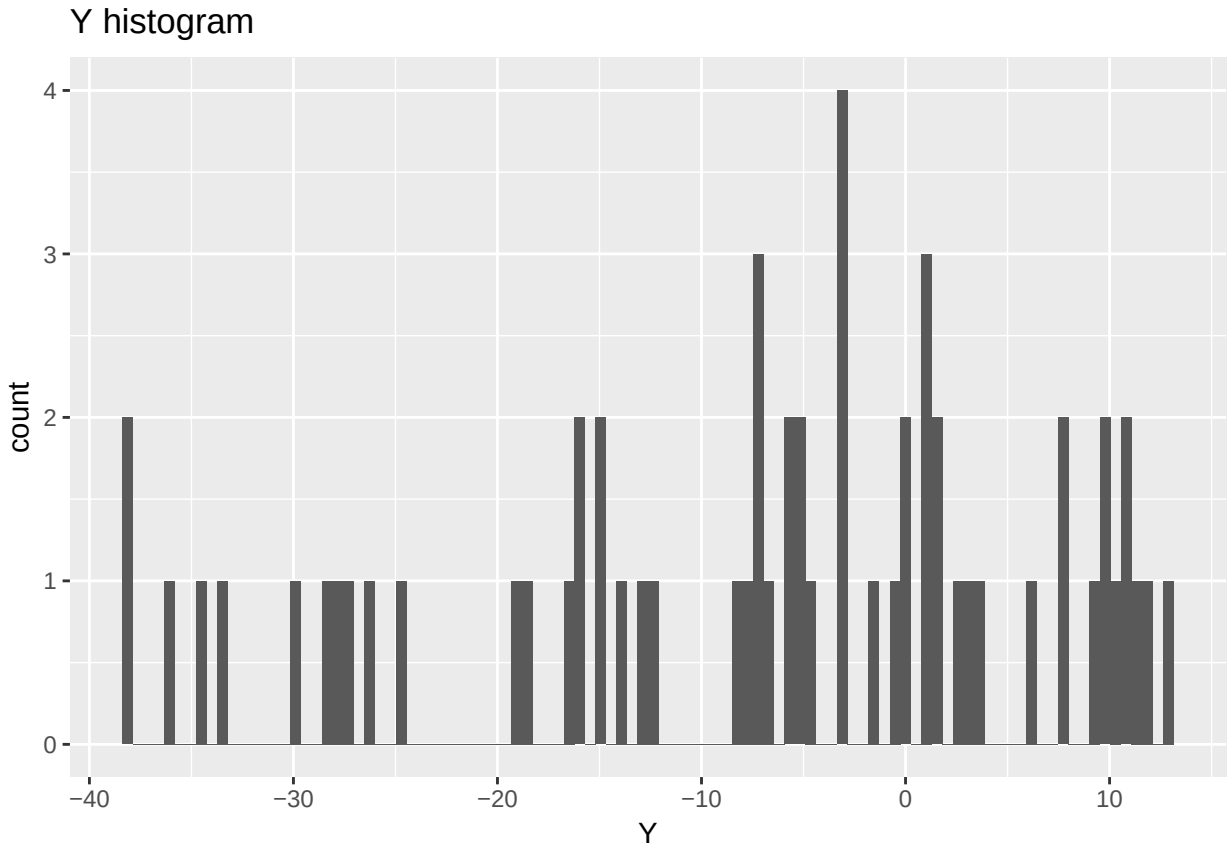
```
library(ggplot2)

plot <- ggplot(dataset, aes(x = X, y = Y))
plot + geom_point() + ggtitle("XY scatterplot")
```



Гистограмма распределения Y

```
qplot(Y, data = dataset, geom = "histogram", bins = 100, main = "Y histogram")
```



Выбор модели

- y = 40,5*x-2,5
- y = 0,5*x+10

- y = 405,5*x-413
- y = 10,11*x^0,08

- y = 103e^(0,5x)

- y = -0,0015*x+0,0076

- y = 0,01e^(-0,5x)

- y = 58x^2-206x+262

- y = -5x^2+30x-37

Расположение точек на диаграмме рассеивания напоминает параболу, уходящую ветвями вниз, поэтому выберем последнюю модель: y = -5x^2+30x-37

Для выбранной модели создаем столбец расчитанных Y для заданных X.

```
f9 <- function(x) { -5 * x^2 + 30 * x - 37 }
```

```
library(dplyr)
```

```
data <- tbl_df(dataset)
data <- mutate(data, y9 = f9(X))
```

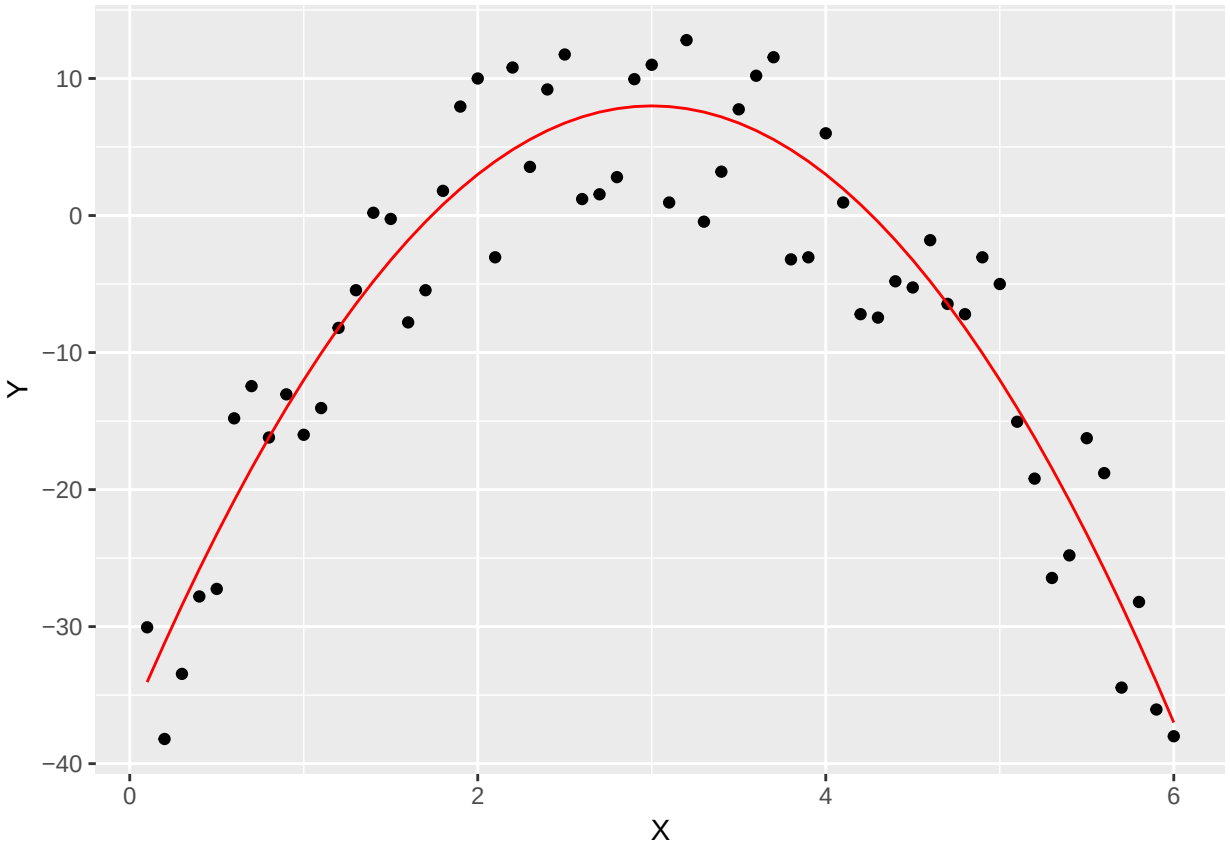
```
head(data)
```

```
## # A tibble: 6 x 3
##       X      Y     y9
##   <dbl>  <dbl>  <dbl>
## 1   0.1 -30.05 -34.05
## 2   0.2 -38.20 -31.20
## 3   0.3 -33.45 -28.45
## 4   0.4 -27.80 -25.80
## 5   0.5 -27.25 -23.25
## 6   0.6 -14.80 -20.80
```

```
summary(data)
```

```
##        X                Y                 y9
##  Min.   :0.100   Min.   :-38.200   Min.   :-37.000
##  1st Qu.:1.575   1st Qu.:-16.050   1st Qu.:-16.762
##  Median :3.050   Median : -5.125   Median : -3.250
##  Mean   :3.050   Mean   : -7.508   Mean   : -7.008
##  3rd Qu.:4.525   3rd Qu.:  2.050   3rd Qu.:  4.987
##  Max.   :6.000   Max.   : 12.800   Max.   :  8.000
```

```
qplot(x = X, y = Y, data = data)  + geom_line(mapping = aes(x = X, y = y9), col = "red")
```



Найти разницу между заданным Y и расчитанным (ошибку модели), построить график изменения ошибки от X.

```
## Ошибка модели
data <- mutate(data, diff = y9-Y)
summary(data$diff)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -7.00   -3.25    0.50    0.50    5.00    8.00
```

```
## График изменения ошибки от X

qplot(X, diff, data = data, main = "Model error by X") + geom_line()
```