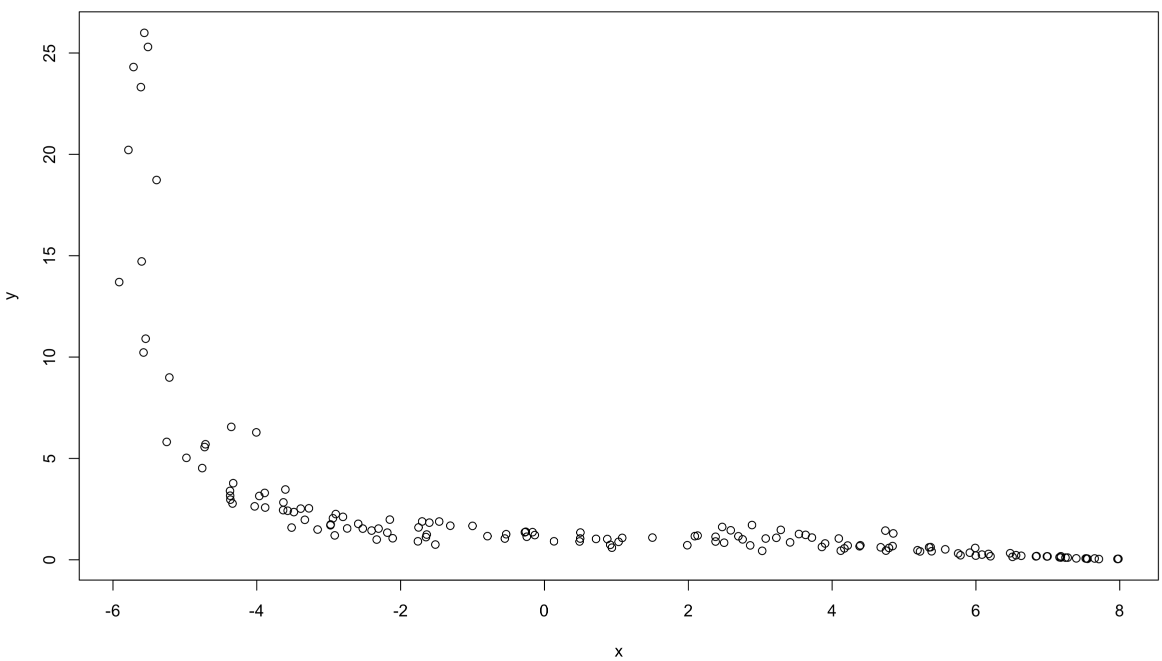
Регресійний аналіз
4 курс, статистика, Шкляр Ірина Володимирівна

Завдання 3, варіант 7

Імпортуємо дані та подивимось аналіз залежності між X та Y за допомогою діаграми розсіювання:

```
ce<-read.table(file="~/Downloads/regrasympt/c7.txt",header=T)
x<-ce[,"X"]
y<-ce[,"Y"]
plot(x,y)
```



Залежність між змінними є нелінійною, є два значення трішки відхилених від загальної кривої, але як викиди їх не вважаємо. Вид функції регресії тут гіперболічний або логарифмічний. Отже, спробуємо подивитися регресію та підгонку методом найменших квадратів:

```
# гіперболічний тип
x0<-1/x
resLm<-lm(y~x0)
summary(resLm)
```

```
Call:
lm(formula = y ~ x0)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.7861 -2.0009 -1.3837 -0.6001 23.3939 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.5651     0.3928   6.530 9.89e-10 ***
x0           -0.1932     0.3137  -0.616    0.539    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.797 on 148 degrees of freedom
Multiple R-squared:  0.002557,  Adjusted R-squared:  -0.004183 
F-statistic: 0.3794 on 1 and 148 DF,  p-value: 0.5389
```

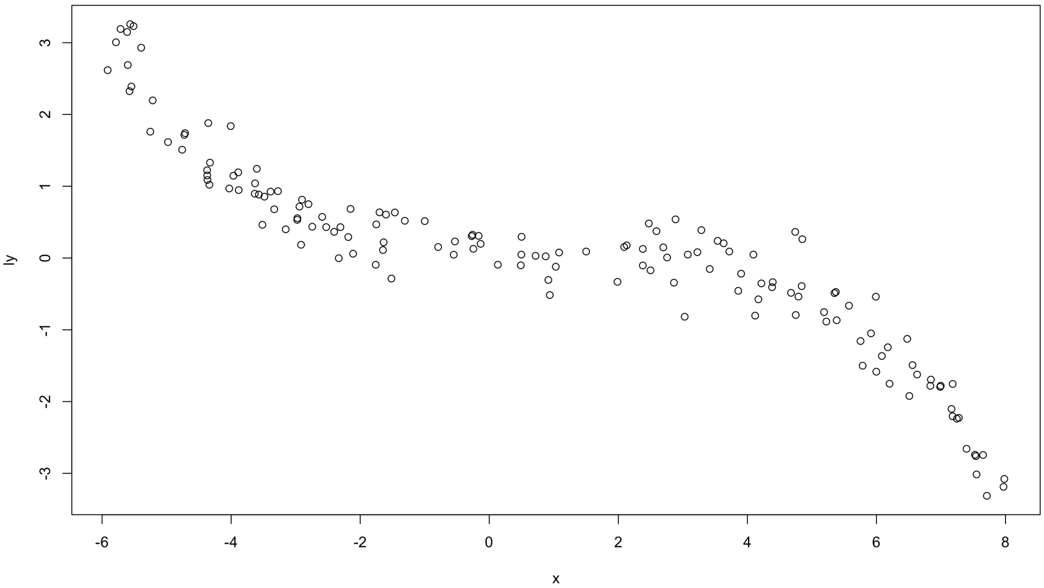
Бачимо, що показники дуже погані, тому ми не можемо описати залежність гіперболічним рівнянням. Спробуємо припустити, що вид функції регресії тут логарифмічний:

```
x4<-log(x)
resLm<-lm(y~x4)
summary(resLm)
```

```
Call:
lm(formula = y ~ x4)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.8605 -0.2635 -0.1017  0.1408  0.9933 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.08233    0.07661  14.128  < 2e-16 ***
x4          -0.34132    0.05071  -6.731 2.55e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3604 on 78 degrees of freedom
  (70 observations deleted due to missingness)
Multiple R-squared:  0.3674,    Adjusted R-squared:  0.3593 
F-statistic: 45.31 on 1 and 78 DF,  p-value: 2.546e-09
```

Показники стали кращими, але все одно ми не можемо описати залежність логарифмічним рівнянням. Тому спробуємо прологарифмувати змінну y. Отримуємо діаграму розсіювання:

ly<-log(y)
plot(x,ly)



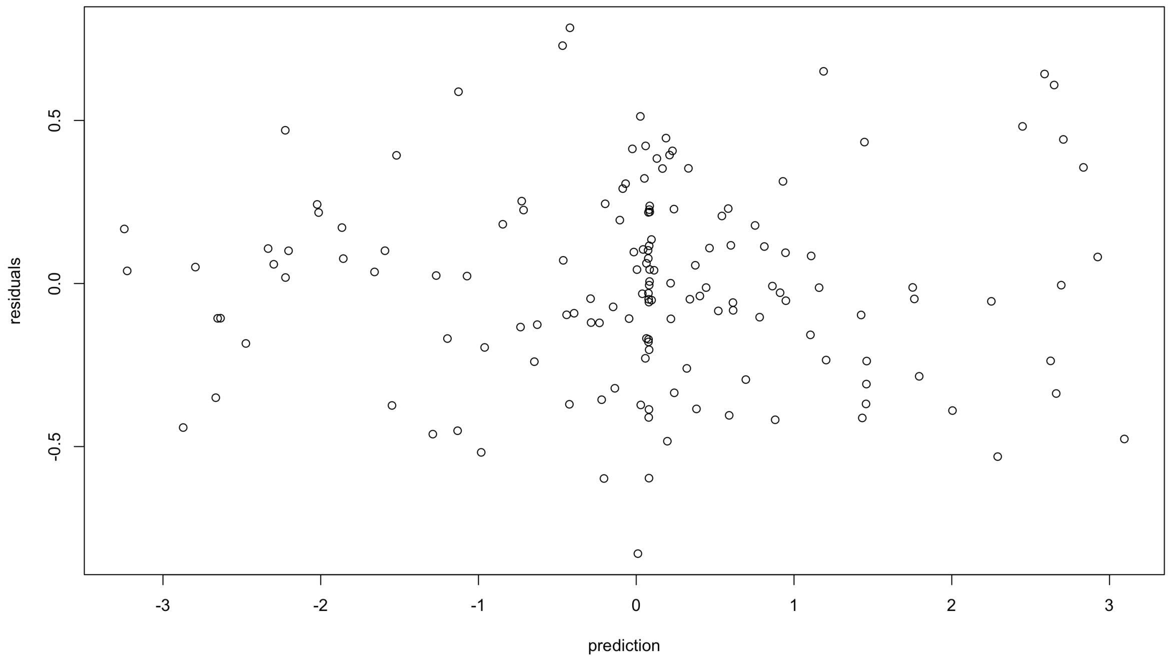
На цій діаграмі точки спостережень вкладаються на досить регулярну криву, що виглядає як графік поліному 3-го порядку. Введемо дві нові змінні x2 = x^2 та x3 = x^3 та розглянемо регресію ly на x, x2, x3. Отримуємо результати підгонки методом найменших квадратів:

x2<-x^2
x3<-x^3
resLm<-lm(ly~x+x2+x3)
summary(resLm)

```
Call:
lm(formula = ly ~ x + x2 + x3)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.8282 -0.1826 -0.0064  0.1913  0.7842 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.0793800  0.0406829   1.951   0.0529 .  
x           -0.0150394  0.0137868  -1.091   0.2771    
x2           0.0266490  0.0021810  12.219   <2e-16 ***
x3          -0.0096411  0.0004516 -21.348   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.299 on 146 degrees of freedom
Multiple R-squared:  0.9519,    Adjusted R-squared:  0.9509 
F-statistic:   963 on 3 and 146 DF,  p-value: < 2.2e-16
```

Тут вже бачимо, що показники набагато кращі. Таким чином, отримали модель з дуже високим коефіцієнтом детермінації 0.9519, досягнутий рівень значущості для перевірки гіпотези про відсутність залежності практично 0. Такі самі досягнуті рівні значущості для перевірки значущості коефіцієнтів при x, x2, x3, тобто залежність між x та log(y) дійсно описується поліномом третього ступеня. Розглянемо діаграму розсіювання прогноз-залишки:

```
plot(resLm$fitted.values,resLm$residuals, xlab="prediction",ylab="residuals")
```

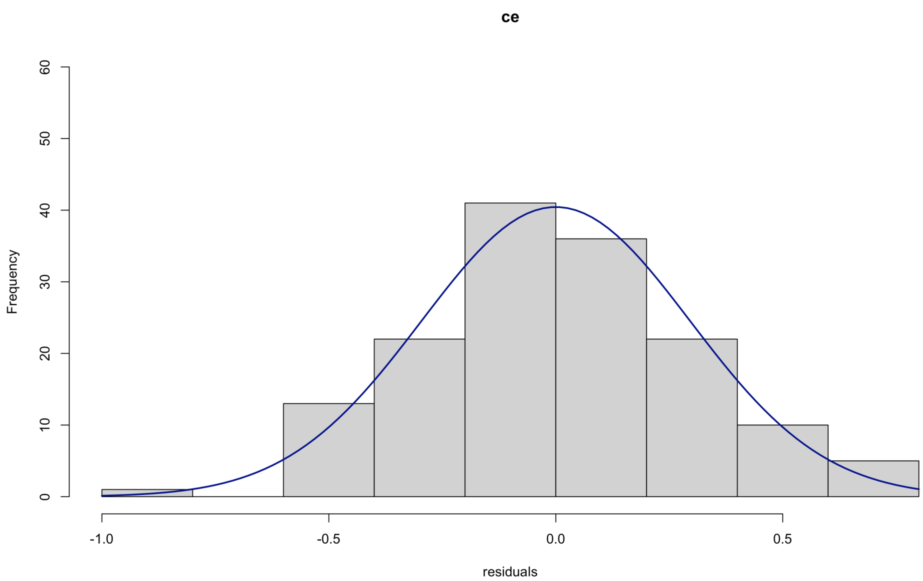


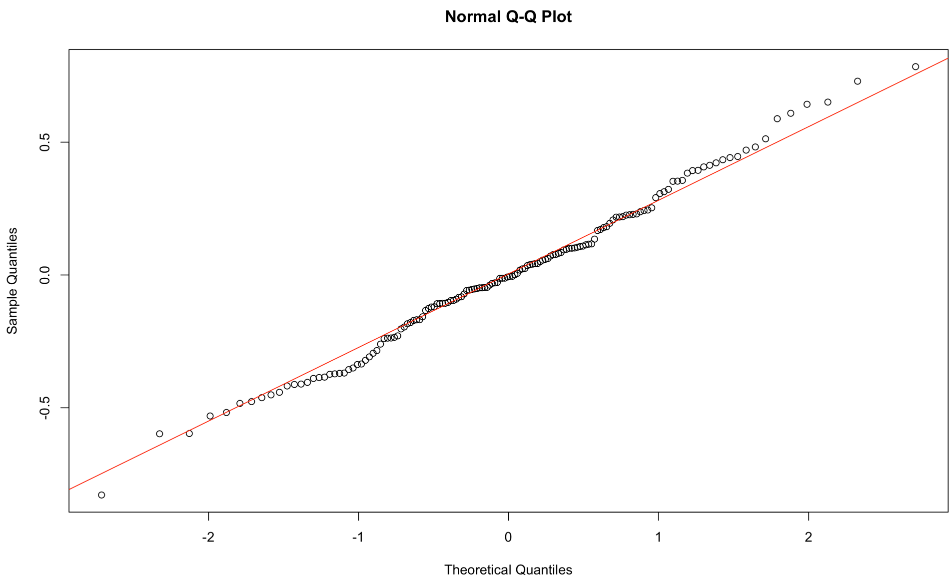
Залишки розкидані хаотично, ніяких закономірностей, що можуть свідчити про наявність невиявлених залежностей, немає.

Побудуємо остаточну модель: $y = \exp(0.08 - 0.015x + 0.03x^2 - 0.01x^3)\eta$,

де η – мультиплікативна помилка. Подивимось гістограму та діаграму квантиль проти квантиля:

```
# histogram of absolute frequencies with density curve
hi<-hist(resLm$residuals,  breaks=8, xlab="residuals", ylim=c(0, 60), main="ce")
curve(dnorm(x, mean=mean(resLm$residuals),
sd=sd(resLm$residuals))*length(resLm$residuals)*(hi$breaks[2]-hi$breaks[1]),
col="darkblue", lwd=2, add=TRUE, yaxt="n")

# QQ-diagram
qqnorm(resLm$residuals)
qqline(resLm$residuals,col="red")
```





На QQ-діаграмі відхилень від нормальності не помітно. Гістограма виглядає дещо асиметричною, чого не повинно бути при нормальному розподілі. Але це може бути результатом випадкових відхилень похибок.