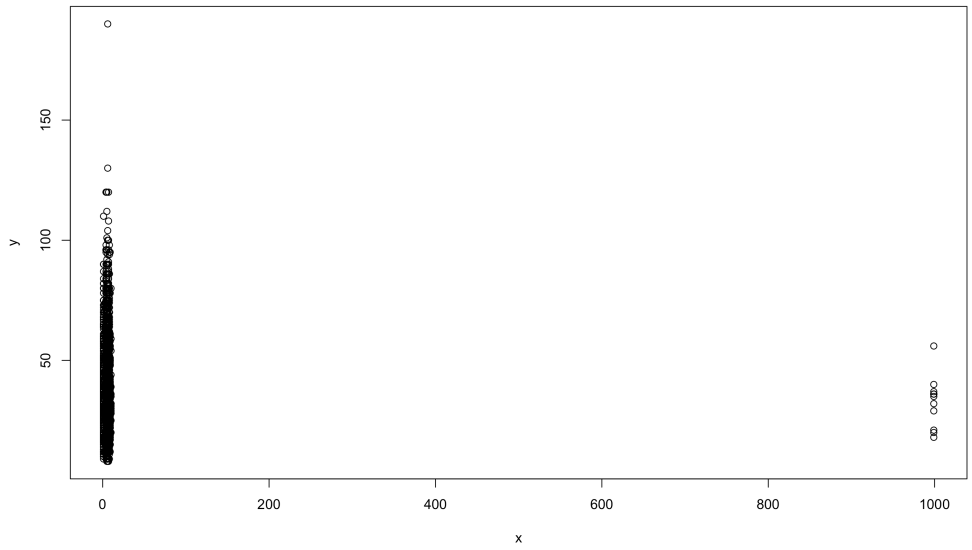
Регресійний аналіз
4 курс, статистика, Шкляр Ірина Володимирівна

Завдання 4, варіант 7

Потрібно побудувати регресійну модель залежності між рівнем освіти голови домогосподарства (L_EDUC) та житловою площею (SLIV) у Вінницькій та Одеській областях.

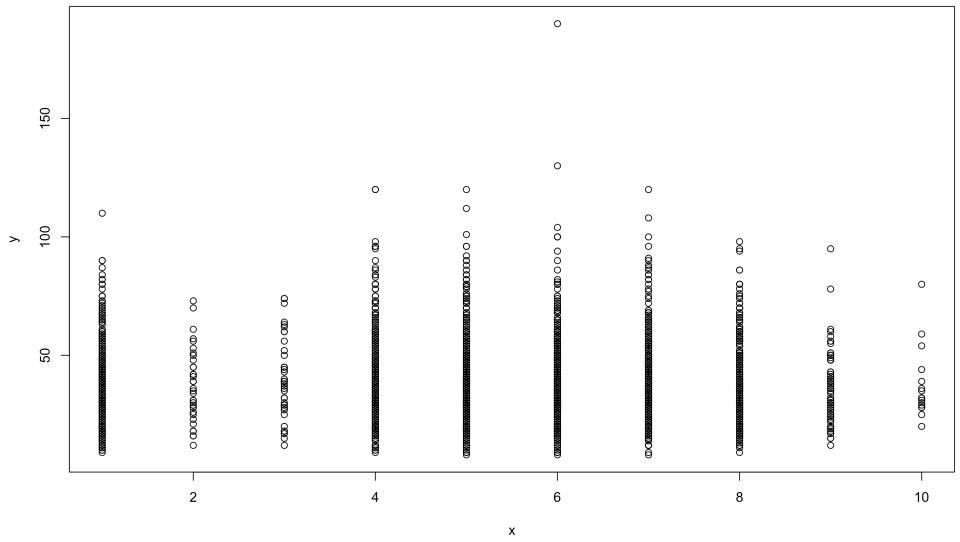
Зробимо підгонку обмеженої моделі, що відповідає відсутності залежності від області. Також зробимо діаграму розсіювання.

```
data<-read.table(file="~/Downloads/regrasympt/house01.txt",header=T)

x<-data[,"L_EDUC"]
y<-data[,"SLIV"]

plot(x,y)
```



Бачимо, що є викиди (оскільки рівень освіти не може дорівнювати 999). Приберемо їх з наших даних. Я зробила окремий файл, де немає викидів: house011.txt.

```
data<-read.table(file="~/Downloads/regrasympt/house011.txt",header=T)

x<-data[,"L_EDUC"]
y<-data[,"SLIV"]

plot(x,y)
```

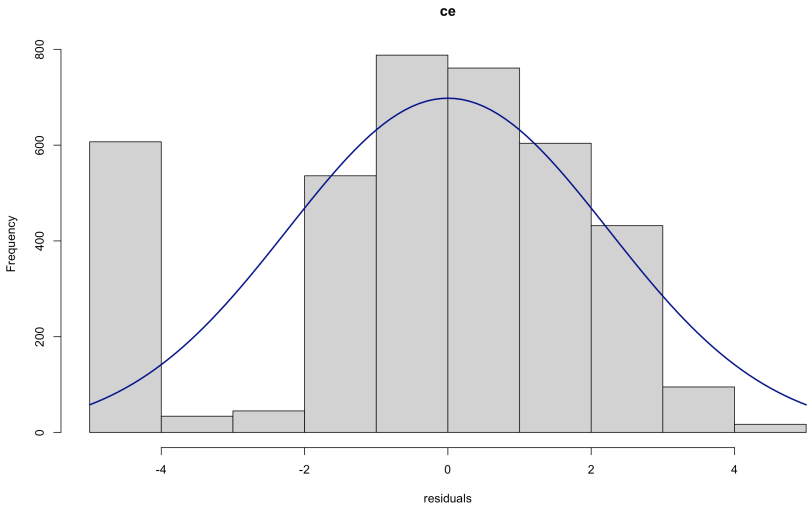



```
> resr<-lm(L_EDUC~SLIV,data=data)
> resr

Call:
lm(formula = L_EDUC ~ SLIV, data = data)

Coefficients:
(Intercept)         SLIV
   5.065543     0.002153
```

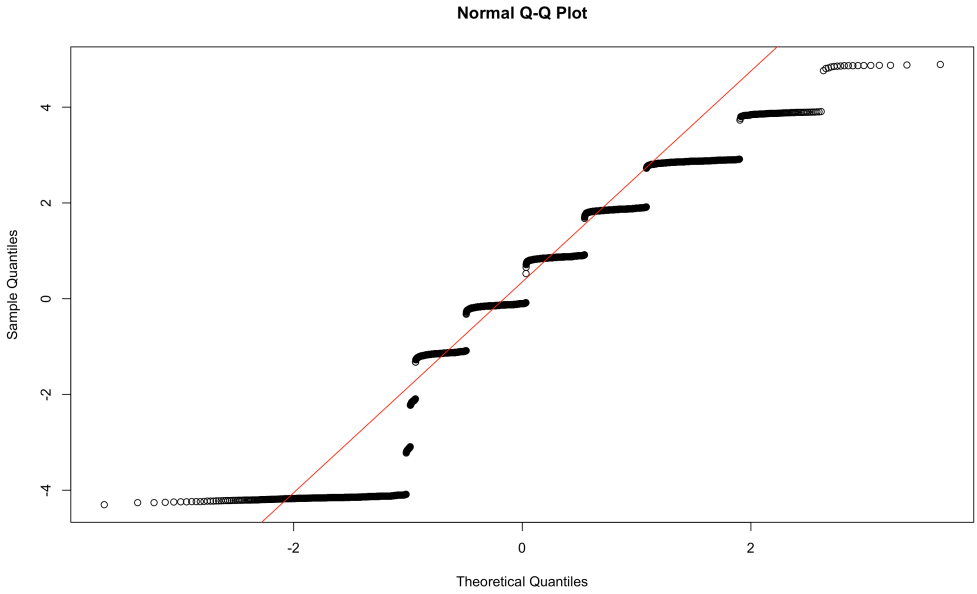
Тобто у підігнаній моделі L_EDUC ≈ 5.065543 – 0.002153 * SLIV

Подивимось гістограму та діаграму квантиль проти квантиля:

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



QQ-diagram
> qqnorm(resr$residuals)
> qqline(resr$residuals,col="red")



Дані по х є категоріальними даними, тому маємо стрибки на діаграмі.

Щоб знайти суму квадратів залишків у цій моделі, застосуємо функцію anova() до результату підгонки моделі:

```
> anovar<-anova(resr)
> anovar
Analysis of Variance Table

Response: L_EDUC
            Df  Sum Sq Mean Sq F value Pr(>F)
SLIV         1     4.7  4.7135  0.9393 0.3325
Residuals 3917 19655.2  5.0179
```

Присвоїмо значення суми квадратів залишків (19655.2) з таблиці дисперсійного аналізу змінній RSSr:

> RSSr=anova(resr)["Residuals","Sum Sq"]

Сума квадратів залишків необмеженої моделі: RSSu = RSSv + RSSo, сума квадратів залишків моделі підігнаної тільки для Вінницької області та моделі тільки для Одеської:

> RSSv<-anova(lm(L_EDUC~SLIV,data=data, subset=(COD_OBL==5)))["Residuals","Sum Sq"]
> RSSo<-anova(lm(L_EDUC~SLIV,data=data, subset=(COD_OBL==51)))["Residuals","Sum Sq"]
> RSSu=RSSv+RSSo

Підрахуємо F-відношення Фішера та досягнутий рівень значущості тесту Чоу для перевірки наявності розшарування:

```
> F<-(RSSr-RSSu)*(length(data$L_EDUC)-4)/(2*(RSSu))
> p<-1-pf(F,2,length(data$L_EDUC)-4)
> F
[1] 8116.817

> p
[1] 0
```

Значення статистики тесту F = 8116.817 та досягнутий рівень значущості p = 0. Отже p < alpha, alpha = 0.05 – стандартний рівень значущості. Тому відхиляємо гіпотезу про те, що розшарування залежності між L_EDUC та SLIV та змінною, що відповідає за область (COD_OBL) відсутнє – розшарування присутнє.