Регресійний аналіз
4 курс, статистика, Шкляр Ірина Володимирівна

Завдання 2

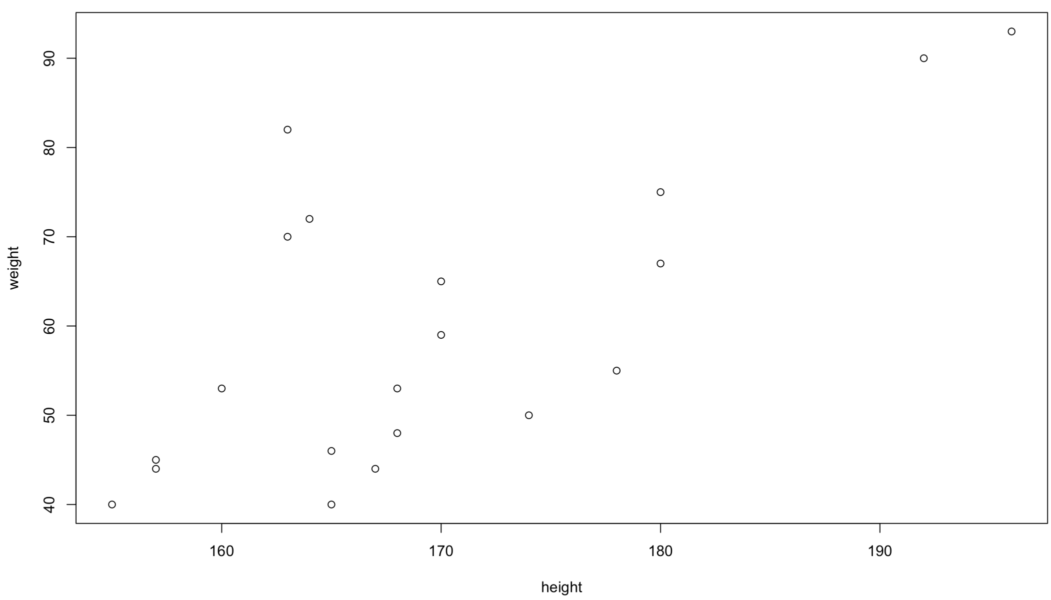
Я взяла дані для роботи частково з життя, і частково з інтернету на тему: вага і зріст 20 людей.

people.txt

N	HEIGHT	WEIGHT
1	163	82
2	168	53
3	157	44
4	160	53
5	164	72
6	167	44
7	155	40
8	163	70
9	170	65
10	165	40
11	180	75
12	174	50
13	192	90
14	168	48
15	170	59
16	157	45
17	178	55
18	180	67
19	196	93
20	165	46

Розпочнемо з діаграми розсіювання, щоб побачити залежність між змінними height і weight:

```
peo <- read.table(file="~/Downloads/regrasympt/people.txt", header=T)
plot(peo[,"HEIGHT"], peo[,"WEIGHT"], xlab = "height", ylab="weight")
```



Залежність не сильно помітна. Також її лінійність сумнівна.

Спробуємо застосувати лінійну регресію:

resLm<-lm(WEIGHT~HEIGHT,data=peo)
summary(resLm)

```
Call:
lm(formula = WEIGHT ~ HEIGHT, data = peo)

Residuals:
    Min      1Q  Median      3Q     Max 
-14.777  -9.055  -1.976   5.290  29.298 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -116.4336    42.8710  -2.716 0.014165 *  
HEIGHT         1.0376     0.2523   4.113 0.000653 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.1 on 18 degrees of freedom
Multiple R-squared:  0.4845,	Adjusted R-squared:  0.4559 
F-statistic: 16.92 on 1 and 18 DF,  p-value: 0.0006528
```

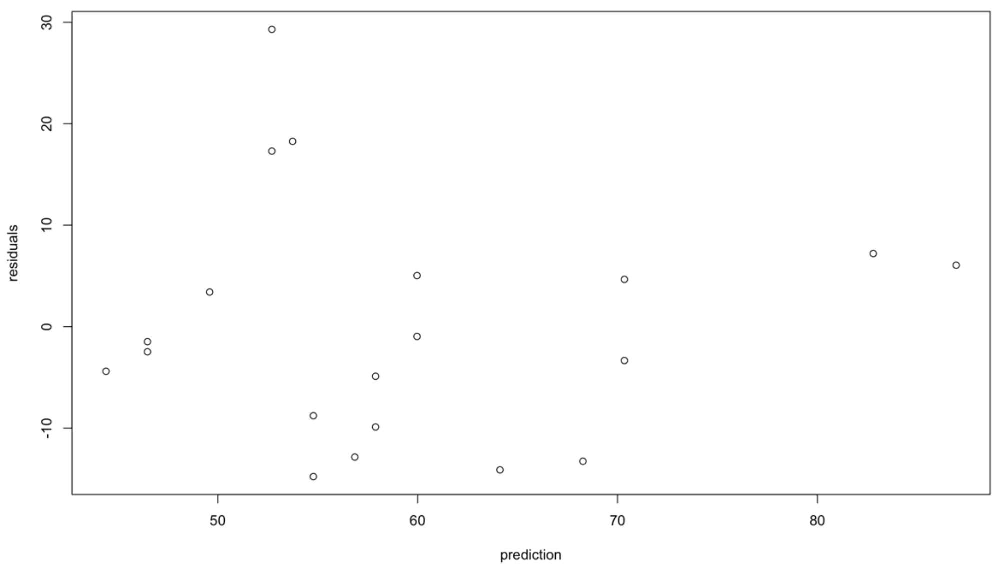
Наша модель має коефіцієнт детермінації 0.4845, це мало для практичної мети прогнозування однак досягнутий рівень значущості для перевірки гіпотези про наявність залежності від хоча б одного з регресорів p = 0.0006528, тобто значуща залежність виявлена. Модель залежності, підігнана за методом найменших квадратів, має вигляд:

WEIGHT = -116.4336 + 1.0376*HEIGHT

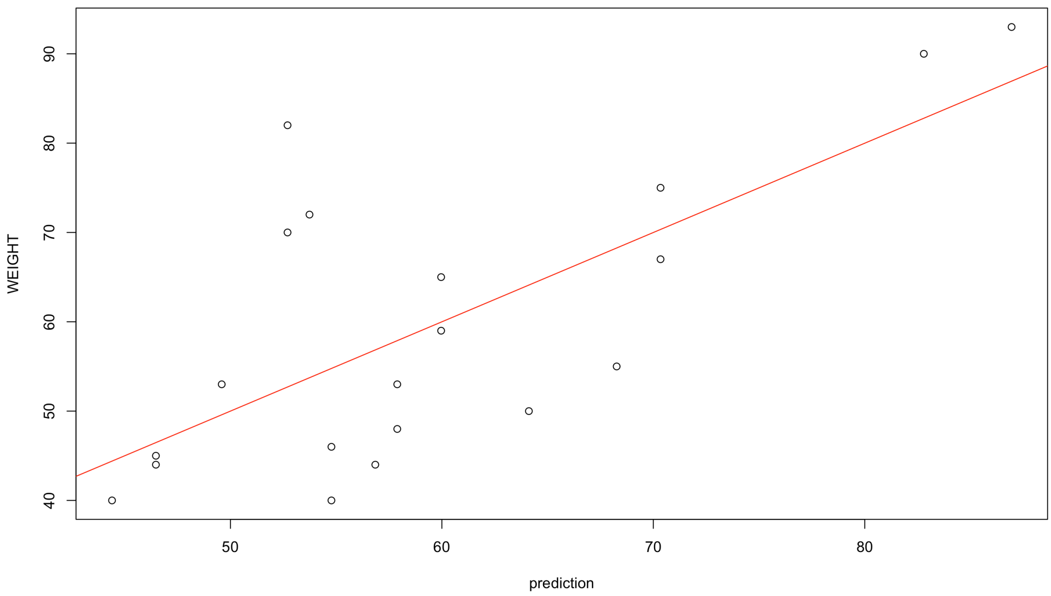
У цій моделі вільний член не значущо відрізняється від 0 (p= 0.014165), коефіцієнт при HEIGHT – не значущо відрізняється від 0 (p= 0.000653).

Проведемо аналіз залишків. Почнемо з діаграми прогноз-залишки:

resLm<-lm(WEIGHT~HEIGHT,data=peo)
plot(resLm$fitted.values,resLm$residuals, xlab="prediction",ylab="residuals")

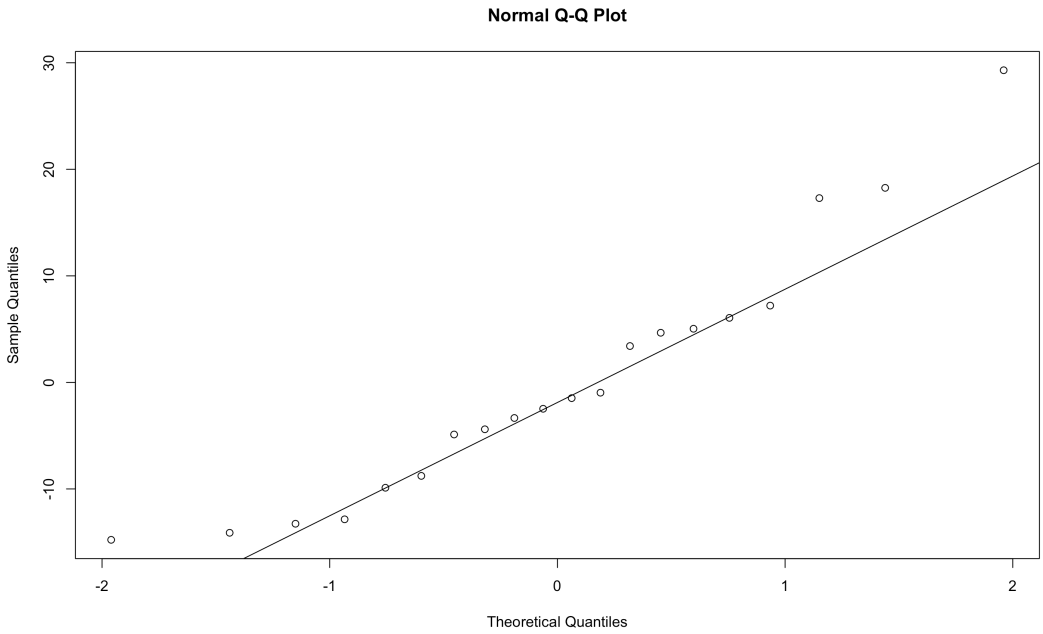


Можливо, дані тут розташовані хаотично. Тоді отримуємо діаграму прогноз – справжні значення відгуку:



Прогноз за формулою вловлює основну тенденцію спостережень, але розкиданість відносно прогнозу досить висока. Перевірка гауссовості розподілу похибок для наших даних не може бути достатньо обґрунтованою внаслідок малої кількості спостережень (20). Зокрема, побудова гістограми при такій кількості даних не доцільна.
Діаграма квантиль проти квантиля не показує відхилень від гауссовості розподілу похибок:

```
qqnorm(resLm$residuals)
qqline(resLm$residuals)
```



Але мені здається, що дані розташовані не зовсім хаотично, тому спробуємо покращити прогноз. На діаграмі трішки простежується нелінійна залежність, схожа на графік exp(x), який зміщується і повторюється три рази. При спаданні прогнозованих значень weight, зростає розкид залишків. Спробуємо застосувати мультиплікативну модель:

$WEIGHT_j = C \times HEIGHT^a_j \times \eta_j$,
де C і a - невідомі параметри, η - мультиплікативна похибка.

Параметр a в нашому випадку дуже схожий по ознакам на випадок залежності ваги і довжини оселедців. Оскільки людина росте більше в висоту, ніж в ширину, то можна сказати, що a тут приблизно має бути між 2 і 3, ближче до трійки.

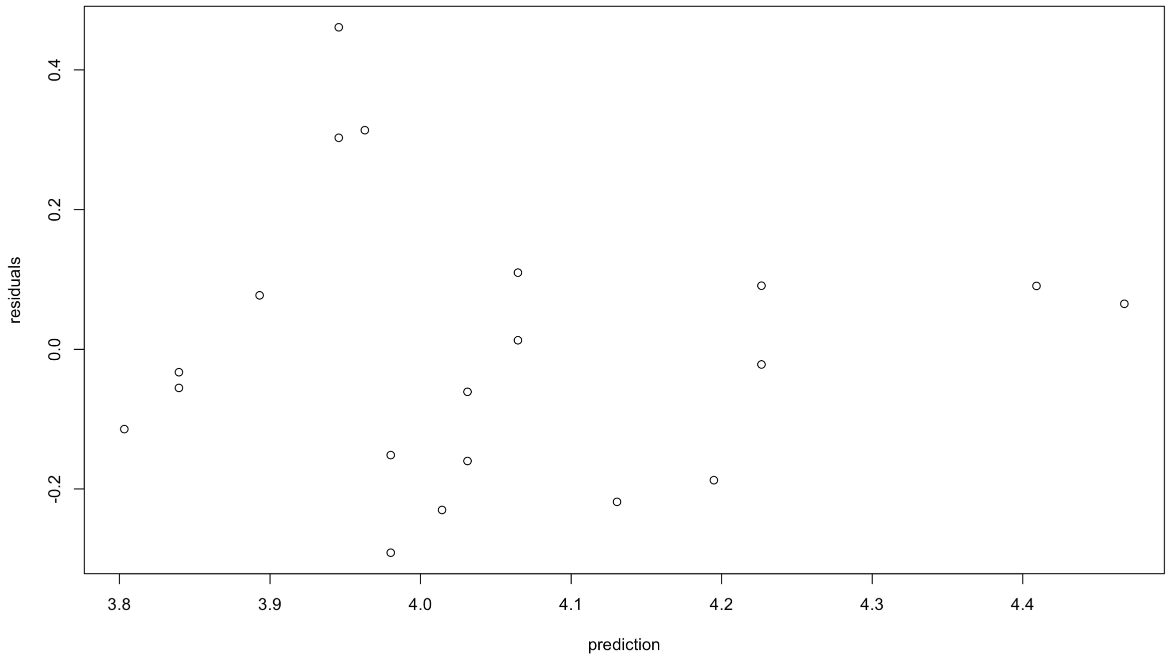
Для лінеаризації моделі перейдемо до змінних LW = log(WEIGHT), LL = log(HEIGHT). В результаті модель зводиться до лінійної з адитивною похибкою ε = log(n):
$LW_j = aLL_j + b + \varepsilon_j$,

де b = log(C). Проведемо підгонку цієї моделі та розглянемо діаграму прогноз-залишки:

```
resLog<-lm(log(WEIGHT)~log(HEIGHT),data=peo)
summary(resLog)
```

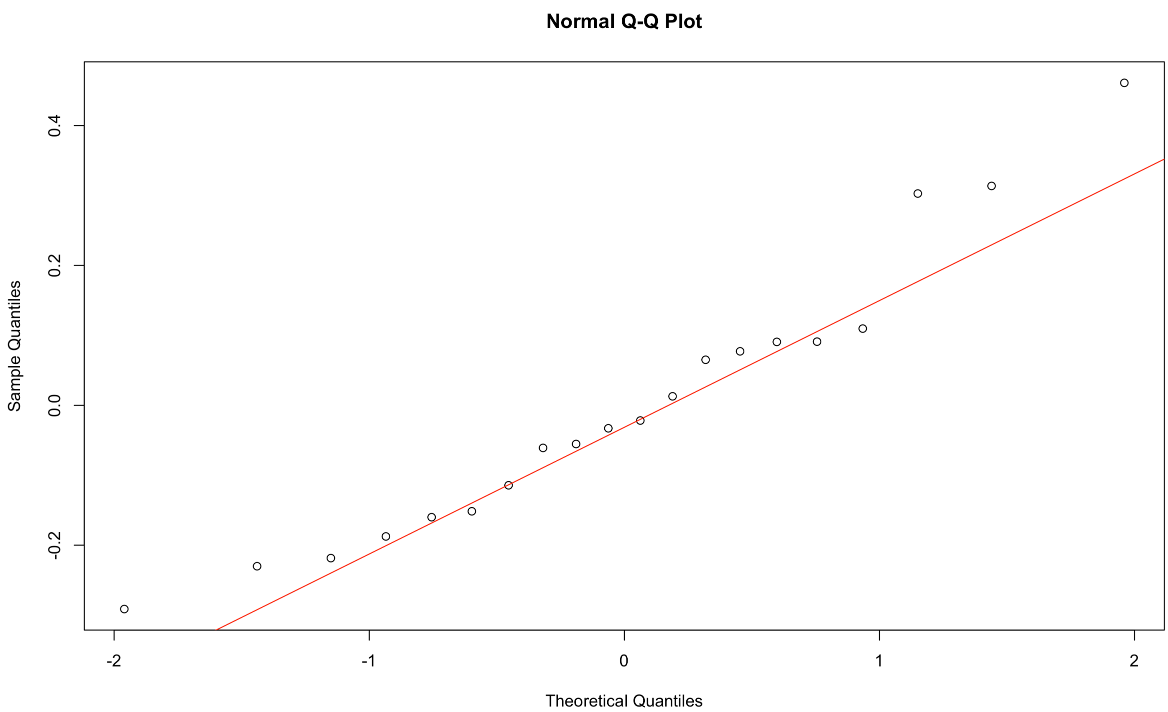
```
Call:
lm(formula = log(WEIGHT) ~ log(HEIGHT), data = peo)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.29131 -0.15366 -0.02731  0.09077  0.46105 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) -10.4712     3.7556  -2.788  0.01214 * 
log(HEIGHT)   2.8303     0.7318   3.868  0.00113 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2017 on 18 degrees of freedom
Multiple R-squared:  0.4538,	Adjusted R-squared:  0.4235 
F-statistic: 14.96 on 1 and 18 DF,  p-value: 0.001128
```

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



Ми отримали оцінку для коефіцієнта a = 2.8303 - це схоже на наше очікуване значення.
Отримані значення коефіцієнтів значущо відрізняються від 0.
Перевіримо нормальність похибок у отриманій лінеаризованій моделі, використовуючи QQ-діаграму:

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



Видно, що спостережуваний розподіл залишків добре описується нормальним розподілом. Отже, приймаємо наступну модель даних:

$$WEIGHT = 0.001128 \times HEIGHT^{2.8303}\ \eta j,$$

де η – мультиплікативна похибка з логнормальним розподілом.

Але, хоча ми і прийняли цю модель, вона майже не відрізняється від qq діаграми для resLm. Отже, залежності майже не виявлено, хоча і пробували поліпшити прогноз (при умові, що залежність все ж таки хоч і невелика, але є). Тому для кращої точності прогнозування, потрібно провести додатковий аналіз залежності ваги і зросту людини та можливо деяких інших факторів, що можуть впливати на залежність.