Київський національний університет
імені Т.Шевченка

Звіт

до лабораторної роботи №2
на тему:

«Побудова математичної моделі методами аналізу даних»

Студента третього курсу
Групи ДО-3
Факультету комп'ютерних наук
та кібернетики

Київ
2021

Вхідні данні

Для виконання лабораторної роботи було використано набір даних з ресурсу https://www.kaggle.com/tombutton/body-measurements. Даний набір містить наступні статистичні дані про заміри людського тіла (250 спостережень): вік, висота, вага, окружність грудей, окружність талії, відсоток жиру в організмі

Хід роботи

Побудова математичної моделі методами аналізу даних

Постановка задачі: провести побудову математичної моделі методами аналізу даних для обраних скалярних змінних. Для цього, максимально використовуючи результати отримані у лабораторній роботі №1:

- визначитися з класом апроксимуючих параметричних функцій для правої частини моделі,
- обчислити точечні та множинні оцінки невідомих параметрів моделі та їх характеристики, які вважаєте потрібними,
- уточнити структуру Вашої математичної моделі, з обгрунтуванням,
- з'ясувати якість отриманої математичної моделі, чисельний та графічний супровід рекомендується,
- по отриманій математичній моделі сформулювати: висновки, вказати на виявлені її недоліки та шляхи її покращення

P.S. Усі графічні зображення повинні мати усі відповідні надписи.

Виконання:

На жаль, для побудови математичної модулі не вдасться використати проаналізований раніше набір даних (через відсутність лінійної залежності), тому було обрано та додатково досліджено новий набір даних:

```
# Correlation between Body fat and {Weight, Chest, Abdominal}
xFat <- lm(fat~weight+chest+abdom)
cor.test(xFat$model$fat, xFat$fitted.values)
```

```
##
##  Pearson's product-moment correlation
##
## data:  xFat$model$fat and xFat$fitted.values
## t = 25.455, df = 248, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8120222 0.8814779
## sample estimates:
##       cor
## 0.8504141
```

Можемо спостерігати залежність між відсотком жиру в організмі та вагою, окружністю грудей та окружністю талії. Отже, за залежну змінну візьмемо

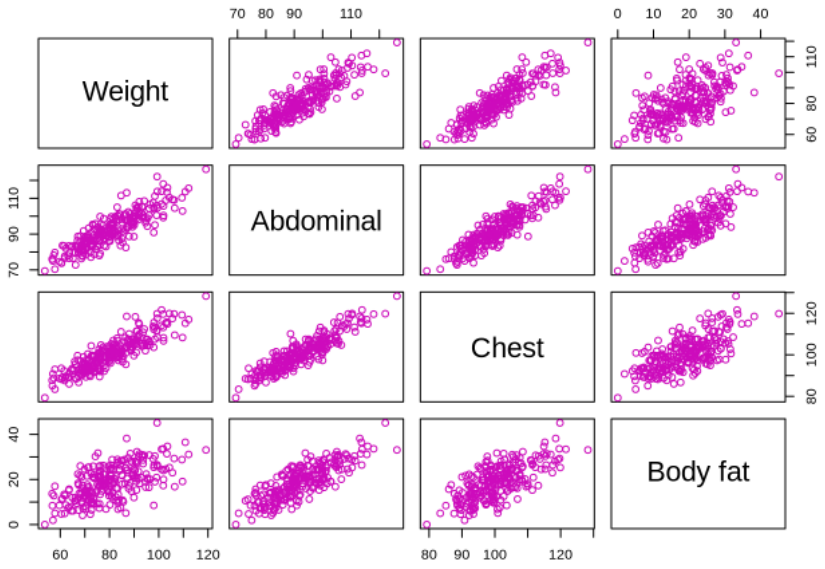
«Відсоток жиру в організмі», а за незалежні – «Вагу», «Окружність грудей» та «Окружність талії» відповідно. Оскільки маємо на меті побудувати математичну модель істотних зв'язків між виключно кількісними змінними, використаємо для цього регресійний аналіз. Математичну модель шукатимемо у вигляді $\eta = f(\bar{\xi}) + \varepsilon$, де $f(\bar{\xi})$ – клас функцій, лінійний по вектору невідомих параметрів $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_p)^T$, а ε – залишкова похибка апроксимації. Тобто можемо переписати шукану модель у наступному матричному вигляді: $y = X \cdot \alpha + e$.

Для початку побудуємо матричну діаграму даних:

```r
# dataset source: https://www.kaggle.com/tombutton/body-measurements
dataset <- read.csv("/home/maxym_ko/cyb/cyb_r/lab_2/measures.csv", header=TRUE)

weight <- dataset$weight * 0.453592 # convert from lb to kg
chest <- dataset$chest
abdom <- dataset$abdom
fat <- dataset$brozek

pairs(list(weight, abdom, chest, fat), col = 6,
      labels = list("Weight", "Abdominal", "Chest", "Body fat"))
```



Можемо переконатися, що залежність між обраною залежною змінною та множиною незалежних змінних справді є, при чому лінійна. В такому разі можна розпочати будувати модель та аналізувати її:

```
model <- lm(fat ~ weight + abdom + chest)
summary(model)
```

```
##
## Call:
## lm(formula = fat ~ weight + abdom + chest)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.793 -2.877 -0.058  3.001  9.975
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -40.70243    4.09834  -9.931  < 2e-16 ***
## weight       -0.25509    0.04914  -5.190 4.39e-07 ***
## abdom         0.92683    0.06460  14.348  < 2e-16 ***
## chest        -0.05363    0.08674  -0.618    0.537
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.066 on 246 degrees of freedom
## Multiple R-squared:  0.7232, Adjusted R-squared:  0.7198
## F-statistic: 214.2 on 3 and 246 DF,  p-value: < 2.2e-16
```

Отже, отримали наступні результати:

- залишки в межах номи, однак слід провести додаткові обстеження
- отримали оцінку вектора невідомих параметрів, таким чином отримали наступну модель: Відсоток жиру в організмі $= \mathbf{-40,70243} - \mathbf{0,25509} \cdot$ Вага $+ \mathbf{0,92683} \cdot$ Оружність грудей $- \mathbf{0,05363} \cdot$ Окружність талії
- коефіцієнт детермінації та скоригований коефіцієнт детермінації (R^2 та $Adjusted\ R^2$ відповідно) доволі значний – 0,7232 та 0,7198 відповідно
- при перевірці гіпотези, $\alpha = 0$ за критерієм Фішера досягнутий рівень значущості $p = 2{,}2 \cdot 10^{-6}$, отже гіпотеза відкидається; аналогічно відкидаємо гіпотези $\alpha_2 = 0$ та $\alpha_3 = 0$; в свою чергу гіпотеза $\alpha_4 = 0$ має високий рівень значущості ($p = 0{,}537$), тому цю гіпотезу приймаємо, а отже змінна «Окружність грудей» є незначущою та її можна викинути

Перебудуємо модель, відкинувши змінну «Окружність грудей»:

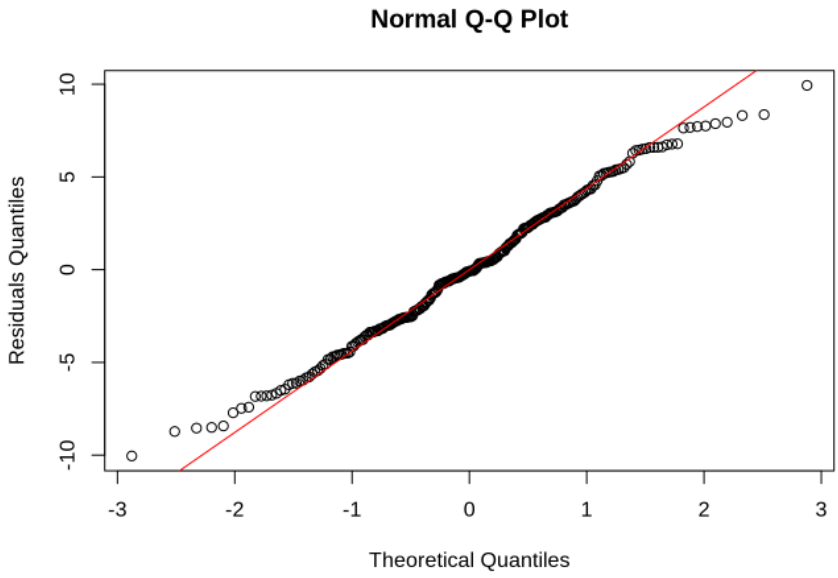
```
model <- lm(fat ~ weight + abdom)
summary(model)
```

```
##
## Call:
## lm(formula = fat ~ weight + abdom)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0453  -2.9594  -0.0841   2.9597   9.9377
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -42.73623    2.44164 -17.503  < 2e-16 ***
## weight       -0.26957    0.04314  -6.248  1.8e-09 ***
## abdom         0.90305    0.05183  17.423  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.061 on 247 degrees of freedom
## Multiple R-squared:  0.7228, Adjusted R-squared:  0.7205
## F-statistic:   322 on 2 and 247 DF,  p-value: < 2.2e-16
```

Бачимо, що результат не зазнав значних змін, однак було спрощено модель, відкиданням непотрібну змінну: Відсоток жиру в організмі $= \mathbf{-42,73623 - 0,26957} \cdot$ Вага $- \mathbf{0,90305} \cdot$ Окружність талії

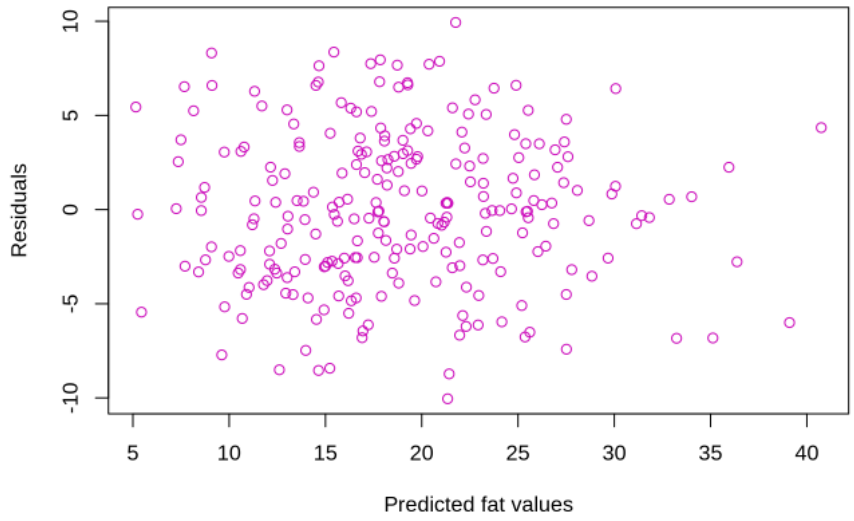
Залишилося переконатися, що залишки мають нормальний розпоідл та є незалежними відносно передбачених регресією значень. Для перевірки першго побудуємо Q-Q графік:

```
# check residuals for normality distribution
qqnorm(model$residuals, ylab = "Residuals Quantiles")
qqline(model$residuals, col = 'red')
```



Аналізуючи графік, переконуємся, що надлишки мають нормальний розподіл. Для перевірки другого побудуємо графік залишків – залежність між залишками та передбаченими регресією значеннями:

```r
# check residuals and predicted values for independence
plot(model$fitted.values, model$residuals, col=6,
     xlab = "Predicted fat values", ylab = "Residuals")
```



Аналізуючи графік, можна прийти до висновку, що залежність відсутня, що добре характеризує нашу модель.

Висновок

В результаті виконання лабораторної роботи було побудовано математичну модель істотних зв’язків між залежною змінною «Відсоток жиру в організмі» та незалежними змінними «Вага», «Окружність грудей», «Окружність талії», яка має вигляд $\text{Відсоток жиру в організмі} = \mathbf{-42,73623 - 0,26957} \cdot \text{Вага} \mathbf{- 0,90305} \cdot \text{Окружність талії}$. В процесі аналізу було побудовано дві моделі – початкову та спрощену (відкинули зайву змінну, а саме «Окружність грудей»). Попри невисокий коефіцієнт детермінації, значних недоліків отриманої моделі не було виявлено. Для програмної реалізації було освоєно та використано мову програмування R та середовище розробки RStudio

Список джерел

- https://www.wikipedia.org
- https://rdocumentation.org
- https://cran.r-project.org
- http://www.r-tutor.com