

Лабораторна робота №2

А. Виконати чистку dataset

Що було зроблено:

1. Перевірка на пропущені дані. (Їх не було)
2. Зміна типу даних деяких колонок на factor, бо там можливі тільки два значення.
3. Змінили назви деяких колонок на більш зручні.

Було побудовано модель, яка показувала залежність ціни на будинок в Індії з п'ятьма різними факторами. А також ми перевірили правильність проведених розрахунків за допомогою cbind.

Коефіцієнти зійшлись, тому все було побудовано правильно.

```
x <- cbind(1, x1,x2,x3,x4,x5)
beta <- solve(t(X) %*% X) %*% t(X) %*% Y
beta
modAll
```

```
data <- read.csv("train.csv")
```

```
#### (A) ####
```

```
# Перевіряємо чи є пропущені дані. Їх немає.
sum(is.na(data))
```

```
summary(data)
```

```
# Змінюємо тип у деяких даних на factor, оскільки вони мають тільки 2 можливих значення
```

```
data$UNDER_CONSTRUCTION <- as.factor(data$UNDER_CONSTRUCTION)
summary(data$UNDER_CONSTRUCTION)
```

```
data$BHK_OR_RK <- as.factor(data$BHK_OR_RK)
summary(data$BHK_OR_RK)
```

```
data$READY_TO_MOVE <- as.factor(data$READY_TO_MOVE)
summary(data$READY_TO_MOVE)
```

```
data$RESALE <- as.factor(data$RESALE)
summary(data$RESALE)
```

```
# Змінюємо назви колонок для зручності
```

```
colnames(data)[colnames(data)=="TARGET_PRICE_IN_LACS."] <- "PRICE"
```

```
colnames(data)[colnames(data)=="BHK_NO."] <- "BHK_NO"
```

```
# Залежна змінна PRICE
```

```
|
```

```
Y <- data$PRICE
```

```
# Незалежні
```

```
x1 <- data$RERA
```

```
x2 <- data$BHK_NO
```

```
x3 <- data$SQUARE_FT
```

```
x4 <- data$LONGITUDE
```

```
x5 <- data$LATITUDE
```

```
# Будуємо модель з 5-ма незалежними змінними
```

```
modAll <- lm(Y ~ x1 + x2 + x3 + x4+ x5)
```

```
summary(modAll)
```

В. Побудуйте однофакторні лінійні моделі

Було побудовано п'ять однофакторних моделей.

(В)

Будуємо однофакторні лінійні моделі за 5-ма факторами

```
mod1 <- lm(Y~x1)
summary(mod1)
```

```
mod2 <- lm(Y~x2)
summary(mod2)
```

```
mod3 <- lm(Y~x3)
summary(mod3)
```

```
mod4 <- lm(Y~x4)
summary(mod4)
```

```
mod5 <- lm(Y~x5)
summary(mod5)
```

С. Визначте кращу модель за R^2

Оцінено отримані моделі. Судячи з наведених нижче результатів, виявилось, що найкращою моделлю є модель №3 через найбільше значення R^2

№1

```
> mod1 <- lm(Y~x1)
> summary(mod1)
```

Call:

```
lm(formula = Y ~ x1)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|--------|-------|--------|-------|---------|
| -207.0 | -96.6 | -67.6 | -27.6 | 29887.4 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|------------|
| (Intercept) | 112.567 | 4.624 | 24.34 | <2e-16 *** |
| x1 | 95.408 | 8.201 | 11.63 | <2e-16 *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 655.4 on 29449 degrees of freedom

Multiple R-squared: 0.004575, Adjusted R-squared: 0.004541

F-statistic: 135.3 on 1 and 29449 DF, p-value: < 2.2e-16

№2

```

> mod2 <- lm(Y~x2)
> summary(mod2)

Call:
lm(formula = Y ~ x2)

Residuals:
    Min       1Q   Median       3Q      Max
-1605.0   -98.9   -65.0   -10.0  29806.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -57.815     11.028   -5.243 1.59e-07 ***
x2             83.901       4.327   19.391 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 652.7 on 29449 degrees of freedom
Multiple R-squared:  0.01261,    Adjusted R-squared:  0.01257
F-statistic:  376 on 1 and 29449 DF,  p-value: < 2.2e-16

```

№3

```

> mod3 <- lm(Y~x3)
> summary(mod3)

Call:
lm(formula = Y ~ x3)

Residuals:
    Min       1Q   Median       3Q      Max
-7552.8  -102.3   -78.4   -40.2  13640.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.401e+02  3.504e+00   40.0   <2e-16 ***
x3           1.391e-04  1.843e-06   75.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 601.3 on 29449 degrees of freedom
Multiple R-squared:  0.1622,    Adjusted R-squared:  0.1621
F-statistic:  5700 on 1 and 29449 DF,  p-value: < 2.2e-16

```

№4

```

> mod4 <- lm(Y~x4)
> summary(mod4)

Call:
lm(formula = Y ~ x4)

Residuals:
    Min       1Q   Median       3Q      Max
-269.3  -105.3   -78.3   -33.4  29829.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 213.0504    13.6789   15.575 < 2e-16 ***
x4          -3.2935     0.6166   -5.342 9.28e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 656.6 on 29449 degrees of freedom
Multiple R-squared:  0.000968, Adjusted R-squared:  0.000934
F-statistic: 28.53 on 1 and 29449 DF, p-value: 9.279e-08

```

№5

```

> mod5 <- lm(Y~x5)
> summary(mod5)

Call:
lm(formula = Y ~ x5)

Residuals:
    Min       1Q   Median       3Q      Max
-345.8  -104.0   -80.3   -37.2  29857.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 225.3832    28.1158    8.016 1.13e-15 ***
x5          -1.0735     0.3625   -2.961 0.00307 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 656.8 on 29449 degrees of freedom
Multiple R-squared:  0.0002977, Adjusted R-squared:  0.0002637
F-statistic: 8.769 on 1 and 29449 DF, p-value: 0.003066

```

< |

D. Зробіть прогноз на наступний період. Опишіть для яких змінних і які значення беруться для прогнозу

Для прогнозування ціни на житло ми використали отримані коефіцієнти та отримали формулу для знаходження прогнозу.

Щодо даних, які ми взяли:

BHK_NO – кількість

SQUARE_FT – площа житла.

LONGITUDE - довжина

LATITUDE – ширина

Що ж до X, ми взяли приблизно середні значення з таблиці.

```
#### (D) ####
# Формула знаходження прогнозу
#  $Y = 101.8 * X1 + 85 * X2 + 0.0002 * X3 - 4.8 * X4 - 1.5 * X5$ 

# Для прогнозу візьмемо такі значення:
# X1 (RERA)      = 1
# X2 (BHK_NO)    = 3
# X3 (SQUARE_FT) = 1300
# X4 (LONGITUDE) = 25
# X5 (LATITUDE)  = 77

Y_predict <- 101.8 * 1 + 85 * 3 + 0.0002 * 1300 - 4.8 * 25 - 1.5 * 77
Y_predict
# Y_predict = 121.56
```

Е. Опишіть ваші дії, припущення та висновки.

Підсумовуючи, ми:

- Редагували нашу таблицю, зробивши її більш зручною як для читання, так і для побудови моделей.
 - Зробили моделі для того, щоб зробити висновки щодо залежності ціни на житло від п'яти числових змінних з нашої таблиці.
 - Спрогнозували ціну на житло з заданими нами характеристиками.
- При виконанні роботи ми дійшли таких висновків: від площі житла залежить не так багато, однак від кількості кімнат та розташування житлового приміщення сильно залежить кінцева ціна.