
# รายงานประมวลความรู้รายวิชา DSI204

เรื่องการทำนาย

**นำเสนอ**

ผศ.ดร.บุญฤทธิ์ ชูประดิษฐ์

**สมาชิก**

| | |
|---|---|
| นายธนารักษ์ ลีนานนท์ | 6524650030 |
| นายศิรภพ จุลละภมร | 6524650089 |
| นายวัชรนันท์ พันมูล | 6524650071 |
| นายวสันต์ อารัมภ์สกุล | 6524651400 |
| นายณิชพน รัถยาบัณฑิต | 6524651244 |

รายงานนี้เป็นส่วนหนึ่งของการศึกษารายวิชา DSI204 Probability Thinking

ตามหลักสูตรวิทยาศาสตร์บัณฑิต สขาวิทยาศาสตร์และนวัตกรรมข้อมูล

วิทยาลัยสหวิทยาการ มหาวิทยาลัยธรรมศาสตร์

# การพยากรณ์ปริมาณการปลดปล่อยก๊าซ Co2 ในประเทศแคนาดา

1.ลักษณะของกลุ่มข้อมูล (Meta Data)

Data set ที่ทางกลุ่มคณะผู้จัดทำได้นำมาศึกษาและทำการทดลองคือ $Co_2$ Emission_Canada หรือปริมาณการปลดปล่อย $Co_2$ โดยรถยนต์สันดาปประเภทต่างๆในประเทศแคนาดา ซึ่งมีเนื้อหาข้อมูลเกี่ยวกับรายการองค์ประกอบของเครื่องยนต์, ประเภทเชื้อเพลิง, รุ่นของรถยนต์และองค์ประกอบอื่นๆ เป็นต้น โดยกลุ่มคณะผู้จัดทำได้เลือกใช้โปรแกรม R-studio ในการทำการทดลองและวิเคราะห์หาประเด็นสำคัญต่างๆในชุดข้อมูล

นำชุดข้อมูลมาจาก

https://www.kaggle.com/datasets/debajyotipodder/co2-emission-by-vehicles

โดย Data set ที่นำมาชื่อว่า $Co_2$ Emission_Canada ซึ่งมีรายละเอียดดังนี้

- Make = ชื่อบริษัทที่เป็นผู้ผลิตของรถยนต์
- Model = ชื่อ Model ของรถประจำบริษัทที่ผลิตนั้นๆ ซึ่งประกอบไปด้วย
    - 4WD/4X4 = Four-wheel Drive
    - AWD = All-wheel drive
    - FFV = Flexible-fuel vehicle
    - SWB = Short wheelbase
    - LWB = Long wheelbase
    - EWB = Extended wheelbase
- Vehicle Class = ประเภทของยานพาหนะซึ่งอ้างอิงตามประโยชน์ใช้สอย, ความจุและน้ำหนัก
- Engine size = ขนาดของเครื่องยนต์โดยใช้หน่วยเป็นลิตร
- Cylinders = จำนวนลูกสูบ
- Transmission = ประเภทเกียร์และจำนวน
    - A = Automatic
    - AM = Automated manual
    - AS = Automatic with select shift
    - AV = Continuously variable

- M = Manual
- 3 - 10 = Number of gears

- Fuel Type = ประเภทเชื้อเพลิง
  - X = Regular gasoline
  - Z = Premium gasoline
  - D = Diesel
  - E = Ethanol (E85)
  - N = Natural gas
- Fuel consumption in city roads (L/100 Km) = อัตราสิ้นเปลืองเชื้อเพลิงบนถนนในเมือง (ลิตร/100 กม.)
- Fuel consumption in highways (L/100 km) = อัตราสิ้นเปลืองเชื้อเพลิงบนทางหลวง (ลิตร/100 กม.)
- Fuel Consumption Comb (mpg) = อัตราสิ้นเปลืองเชื้อเพลิงแบบผสม (ในเมือง 55% ทางหลวง 45%) แสดงเป็น L/100 กม
- $Co_2$_Emission = ปริมาณการปลดปล่อยก๊าซ $Co_2$

## การใช้สถิติพรรณนา (Descriptive statistics) เพื่อหาประเด็นสำคัญต่างๆ

ทางกลุ่มคณะผู้จัดทำได้มีการใช้สถิติพรรณนาในเรื่องของการวัดตำแหน่งข้อมูลโดยกลุ่มคณะผู้จัดทำรายงานได้เลือกใช้ Quartile และ Inter Quartile Range (IQR) มาเป็นหลักการในการกำจัดค่านอกเกณฑ์ (Outlier) ของ Feature data ที่ชื่อว่า Engine size และตรวจสอบลักษณะของข้อมูลลว่ามีการแจกแจงแบบปกติมาตรฐานหรือไม่โดยใช้กราฟฮิสโทแกรม

$$Q_r = \frac{r}{4} \times (n+1)$$

$$IQR = Q_3 - Q_1$$

$$Outlier = (Q1 - 1.5IQR) \cup (Q3 + 1.5IQR)$$

ตัวอย่าง code ภาษา R

```r
library(dplyr)
library(ggplot2)
library(tidyr)
library(forcats)
library(caret)
library(corrplot)
library(patchwork)


df <- read.csv("C:\\Users\\PEAM\\Documents\\PEAM\\Lecture\\DSI204\\dsi204_proje

names(df)

# Cleaning Outlier
q <- quantile(df$Engine.Size.L., probs=c(0.25,0.75))

iqr <- q[2] - q[1]

df <- df[(df$Engine.Size.L. >= q[1] - 1.5*iqr) &
           (df$Engine.Size.L. <= q[2] + 1.5*iqr),]
```

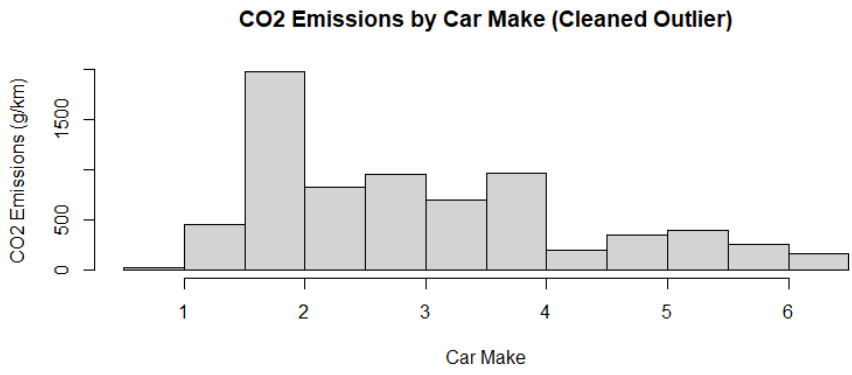
```
> q
25% 75%
2.0 3.7
> iqr
75%
1.7
>
```

ทำให้เราทราบว่าใน Feature Engine size มี Q1 = 2.0 และ Q3 = 3.7 และ IQR = 1.7

ทำการกำจัด Outlier จะพบว่าชุดข้อมูลจะเหลือ 7248 records จาก 7385 records

หลังจากทำการ cleaned outlier นำมา plot histogram เพื่อดูการกระจายตัวที่เกิดขึ้น

```r
hist(df$Engine.Size.L.,
     data = df,
     xlab = "Car Make",
     ylab = "CO2 Emissions (g/km)",
     main = "CO2 Emissions by Car Make (Cleaned Outlier)")
```



**3.Data preparation**

จากกราฟฮิสโทแกรมที่เกิดขึ้นทำให้ทางกลุ่มคณะผู้จัดทำทราบว่าชุดข้อมูลนี้ไม่มีการกระจายตัวเป็นปกติ มาตรฐานเพื่อให้สามารถใช้เทคนิคการถดถอย(regression) ได้ทางคณะผู้จัดทำจึงต้องมีการสร้าง data frame ชุดใหม่ขึ้นมาโดยเรียกว่า X และ Y ซึ่งมีรายละเอียดดังนี้

X คือ data frame ที่ประกอบไปด้วย feature ดังนี้ fuel.type, engine.size.L, fuel consumption ทั้ง 3 รูปแบบ

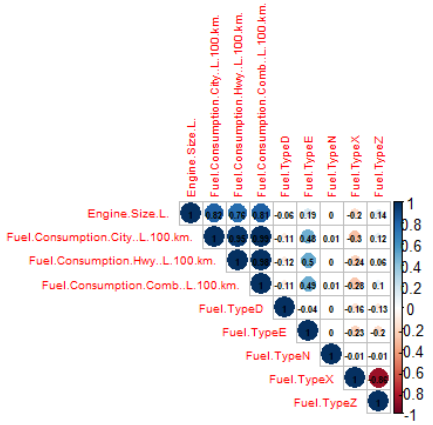
Y คือ data frame ที่เป็น Labeled data ที่มีชื่อว่า Co2.Emission.g.km

ทำการแปลง X ที่ไม่ใช่ Numeric feature ให้เป็น Dummy variable

```
X <- df[,
        (names(df) %in%
           c("Fuel.Type" ,
             "Engine.Size.L.",
             "Fuel.Consumption.City..L.100.km.",
             "Fuel.Consumption.Hwy..L.100.km.",
             "Fuel.Consumption.Comb..L.100.km."  ))]

y <- df[, names(df) == "CO2.Emissions.g.km."]



transformer <- dummyVars(~  Fuel.Type,
                         data= X)
new_X <- data.frame(predict(transformer,
                            newdata = X))

X <- X[, names(X) != "Fuel.Type"]
X <- cbind(X,new_X)
```

ทำการPlot เพื่อดูสหสัมพันธ์(Correlation) ระหว่าง X, Y

```
corrplot(round(cor(X),2),
         method = "circle",
         type = "upper",
         tl.cex = 0.65,
         addCoef.col = "black",
         number.cex = 0.5,
         digits = 2)
```



## 4.Feature selection

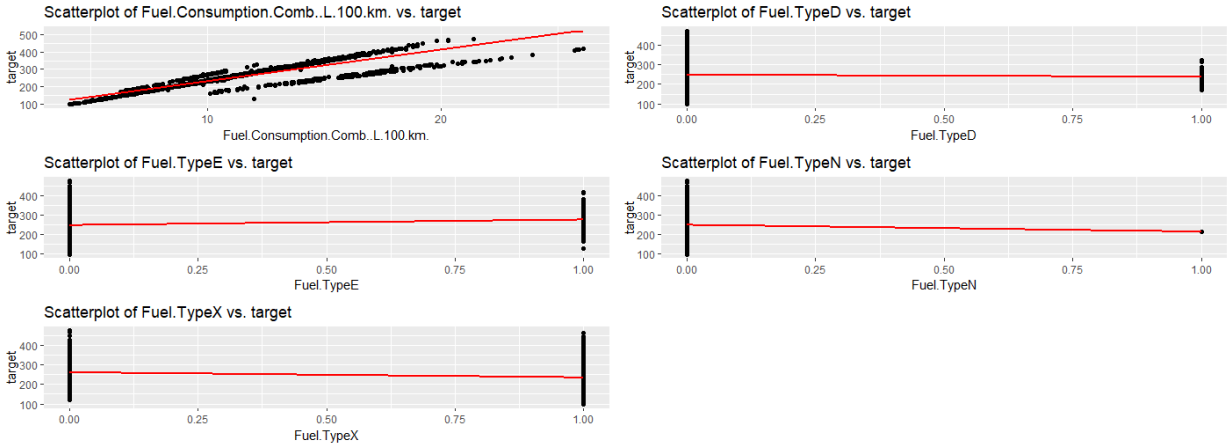
จาก correlation ที่เกิดขึ้นทำให้เราสามารถเลือก X ที่ส่งผลกับ Y ได้โดยอิงจากค่า correlation ที่เกิดขึ้น

ตรวจสอบค่าความสัมพันธ์ที่เกิดขึ้นหากตัด feature fuel.type, engine.size.L, fuel consumption ทั้ง 3 รูปแบบ ออกและสร้างเป็นคอลัมน์ใหม่ที่ทับค่าคอลัมน์เดิม

```
X <- X[,
       (!names(X) %in%
          c("Engine.Size.L.",
            "Fuel.Consumption.City..L.100.km.",
            "Fuel.Consumption.Hwy..L.100.km.",
            "Fuel.TypeZ"))]

corrplot(round(cor(X),2),
         method = "circle",
         type = "upper",
         tl.cex = 0.65,
         addCoef.col = "black",
         number.cex = 0.5,
         digits = 2)
```

จาก correlation ที่เกิดขึ้นทำให้เราสามารถสร้าง data frame ชุดใหม่ที่สามารถนำไปใช้ต่อในเทคนิค regression ได้และหากนำไปตรวจสอบความ linearity จะได้ดังนี้

```
new_df <- data.frame(cbind(X,y))

plots <- list()
feature_vars <- names(X)
for (var in feature_vars)
  {
  p <- ggplot(new_df,
              aes(x = .data[[var]],
                  y = y)) + geom_point()+

    stat_smooth(method = "lm",
                se = FALSE,
                color = "red") +

    ggtitle(paste("Scatterplot of",
                  var,
                  "vs. target")) +
    xlab(var) +
    ylab("target")
  plots[[var]] <- p
}

wrap_plots(plots, ncol = 2)
```




## 5.Regression

ใช้เทคนิค linear regression โดยอิงข้อมูลจาก new_df ในการแบ่งส่วน train-test แบบ 70:30

```
set.seed(204)
trainIndex <- createDataPartition(new_df$y,
                                  p = 0.7,
                                  list = FALSE)

train <- new_df[trainIndex, ]
test <- new_df[-trainIndex, ]
```

ทำการสร้าง model linear regression จะได้ว่า

```
Call:
lm(formula = y ~ ., data = new_df)

Residuals:
    Min      1Q  Median      3Q     Max
-67.221  -2.685   0.000   2.236  44.594

Coefficients:
                                Estimate Std. Error  t value Pr(>|t|)
(Intercept)                      6.12994    0.32271   18.995  < 2e-16 ***
Fuel.Consumption.Comb..L.100.km. 22.74454    0.02738  830.632  < 2e-16 ***
Fuel.TypeD                       30.46087    0.43334   70.292  < 2e-16 ***
Fuel.TypeE                     -114.54174    0.34057 -336.328  < 2e-16 ***
Fuel.TypeN                      -81.98560    5.51360  -14.870  < 2e-16 ***
Fuel.TypeX                       -0.37966    0.13865   -2.738  0.00619 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.513 on 7242 degrees of freedom
Multiple R-squared:  0.9903,    Adjusted R-squared:  0.9903
F-statistic: 1.474e+05 on 5 and 7242 DF,  p-value: < 2.2e-16
```

สมการรูปทั่วไปของ Linear regression คือ

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n$$

ซึ่งสมการที่เกิดขึ้นของ data set ชุดข้อมูล train เป็นดังนี้

```
Call:
lm(formula = y ~ ., data = new_df)

Residuals:
    Min      1Q  Median      3Q     Max
-67.221  -2.685   0.000   2.236  44.594

Coefficients:
                                Estimate Std. Error  t value Pr(>|t|)
(Intercept)                      6.12994    0.32271   18.995  < 2e-16 ***
Fuel.Consumption.Comb..L.100.km. 22.74454    0.02738  830.632  < 2e-16 ***
Fuel.TypeD                       30.46087    0.43334   70.292  < 2e-16 ***
Fuel.TypeE                     -114.54174    0.34057 -336.328  < 2e-16 ***
Fuel.TypeN                      -81.98560    5.51360  -14.870  < 2e-16 ***
Fuel.TypeX                       -0.37966    0.13865   -2.738  0.00619 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.513 on 7242 degrees of freedom
Multiple R-squared:  0.9903,    Adjusted R-squared:  0.9903
F-statistic: 1.474e+05 on 5 and 7242 DF,  p-value: < 2.2e-16
```

กำหนดให้

Fuel.Consumption.Comb..L.100.Km = $X_1$ Fuel.TypeD = $X_2$ Fuel.TypeE = $X_3$ Fuel.TypeN = $X_4$ Fuel.TypeX = $X_5$

เราจึงสามารถเขียนสมการรูปทั่วไปของโมเดลได้ดังนี้

$$\hat{y} = 6.12994 + 22.74454x_1 + 30.46087x_2 - 114.54171x_3 - 81.98560x_4 - 0.37966x_5$$

ซึ่งสามารถตีความผลได้ดังนี้

*b0* = เมื่อให้ตัวแปรอิสระ $x_1$-$x_5$ มีค่าเป็น 0 จะส่งผลให้ *y* มีค่าเพิ่มขึ้น 6.12994

*b1* = เมื่อ Fuel.Consumption.Comb..L.100.Km มีค่าเพิ่มขึ้น 1 หน่วยจะส่งผลให้ *y* มีค่าเพิ่มขึ้น 22.74454

*b2* = เมื่อ Fuel.TypeD มีค่าเพิ่มขึ้น 1 หน่วยจะส่งผลให้ *y* มีค่าเพิ่มขึ้น 30.46087

*b3* = เมื่อ Fuel.TypeE มีค่าเพิ่มขึ้น 1 หน่วยจะส่งผลให้ *y* มีค่าลดลง 114.54171

*b4* = เมื่อ Fuel.TypeN มีค่าเพิ่มขึ้น 1 หน่วยจะส่งผลให้ *y* มีค่าลดลง 81.98560

*b5* = เมื่อ Fuel.TypeX มีค่าเพิ่มขึ้น 1 หน่วยจะส่งผลให้ *y* มีค่าลดลง 0.37966

หากใช้ชุดข้อมูล test (model ไม่เคยเห็นชุดข้อมูลนี้มาก่อน) จะได้ผลดังนี้

| Metrics | value |
|---|---|
| MSE | 30.55181 |
| RMSE | 5.527369 |
| R-squared | 0.9902685 |

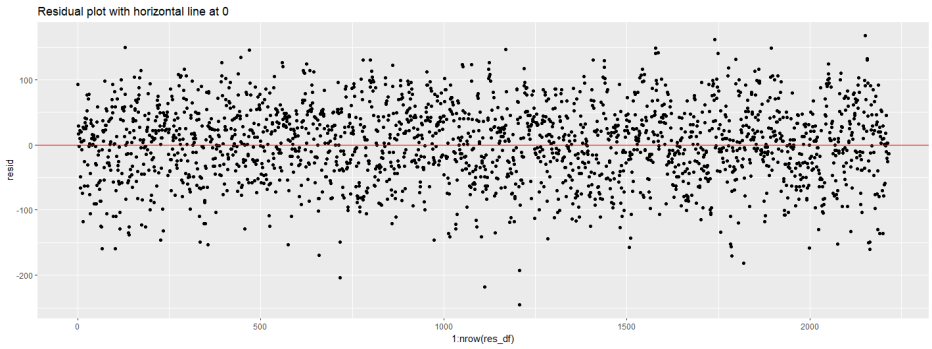
1.R-squared หมายถึง ตัวแปรอิสระX1,X2,X3,X4,X5 สามารถอธิบาย y ได้อย่างถูกต้องร้อยละ99.02% ส่วนที่เหลือสามารถอธิบายได้ด้วยปัจจัยอื่นๆ

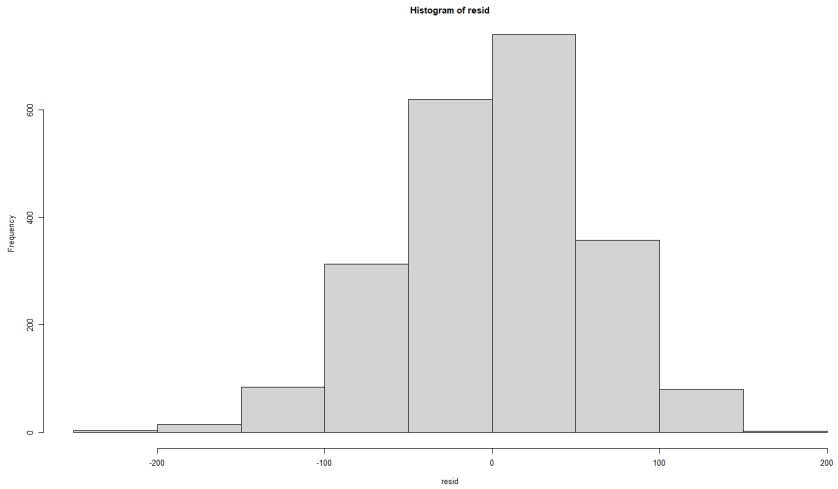
2.MSE หมายถึง ผลรวมค่าเฉลี่ยสัมบูรณ์ของ Residuals มีค่าอยู่ที่ 30.55181หรือประมาณตามหลักนัยสำคัญคือ 31

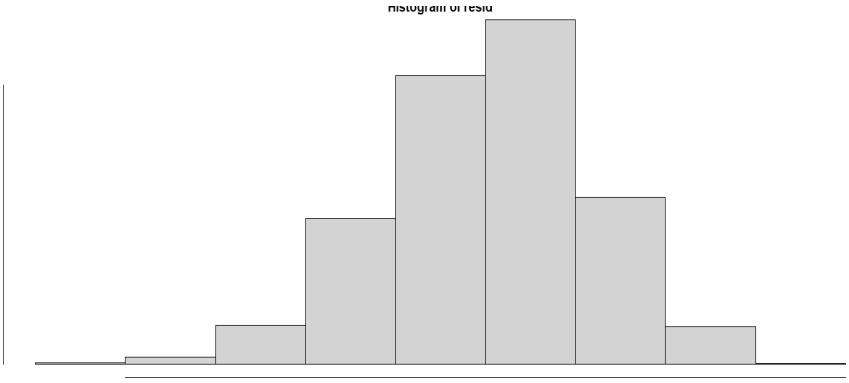
3.RMSE หมายถึง ผลรวมค่าเฉลี่ยสัมบูรณ์ที่ไม่ถูกยกกำลังสองของ ของ Residuals มีค่าอยู่ที่ 5.527369หรือประมาณตามหลักนัยสำคัญคือ 6

เพื่อพิสูจน์ Consumption ทางคณะผู้จัดทำรายงานจึงได้ทำการใช้ Shapiro-wilk test และ plot residual ซึ่งได้ผลลัพธ์ดังนี้

```r
# Plot the residuals
resid <- y_pred - test$y

res_df <- data.frame(resid)

# plot residuals with horizontal line at 0
ggplot(res_df, aes(x = 1:nrow(res_df), y = resid)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red") +
  ggtitle("Residual plot with horizontal line at 0")

shapiro.test(resid)

hist(resid)
qqnorm(model$residuals)
qqline(model$residuals)
```



## 6.สรุปผล

จะพบว่าโมเดลของเราสามารถใช้ได้เนื่องจาก residuals มีการกระจายตัวอย่างเป็นปกติมาตรฐาน(Normal Distribution) ซึ่งแปลว่าโมเดลนี้ผ่าน Assumption ในเบื้องต้น และจากสมการรูปทั่วไปทำให้เราทราบว่าปัจจัยที่มีผลต่อการปลดปล่อยก๊าซ Co2 มากที่สุดคือ Fuel type D หรือ *Diesel*