Natthanan Bhukan 61070507206

King Mongkut's University of Technology Thonburi
Faculty of Engineering, Department of Computer Engineering

**CPE 213 Data Model, 2/2020**

LAB Lecture 9: Linear Regression

Assign Date: 12 April 2021 Due Date: 22 April 2021

---

Experiment and create the best regression model for predicting daily

Answer

| Target..Total.orders. |
|---|
| 0.301805864 |
| -0.481510166 |
| 0.907253271 |
| 0.624550912 |
| 0.602445645 |
| 0.892010948 |
| 0.698149726 |
| -0.008783591 |
| 0.079290905 |
| 0.573524259 |
| 0.734971135 |
| 0.119061627 |
| 1.000000000 |

**รูปที่ 1 Correlation ของข้อมูลที่สัมพันธ์กับ Target**

เริ่มแรกทำการดู Correlation ของข้อมูล เพื่อดูว่ามีตัวแปรไหนบางที่สัมพันธ์กันในข้อมูล Target ของเรา เพื่อใช้สำหรับในการเลือก Feature สำหรับมาทำ Model จากรูปที่ 1

| | target |
|---|---|
| | <dbl> |
| Non.urgent.order | 0.9345178 |
| Urgent.order | 0.7297847 |
| Order.type.B | 0.9006729 |
| Order.type.C | 0.8049838 |
| Banking.orders..1. | 0.6303650 |
| Banking.orders..2. | 0.7984470 |
| Target..Total.orders. | 1.0000000 |

**รูปที่ 2 Correlation ของข้อมูลที่สัมพันธ์กับ Target ที่ทำการเลือก**

หลักจากนั้นทำการกรองเอาค่า Correlations ที่มากกว่า 0.6 และ น้อยกว่า -0.6 จะได้ดังรูปที่2 เพื่อใช้เป็นตัวเลือกสำหรับในการสร้าง Model

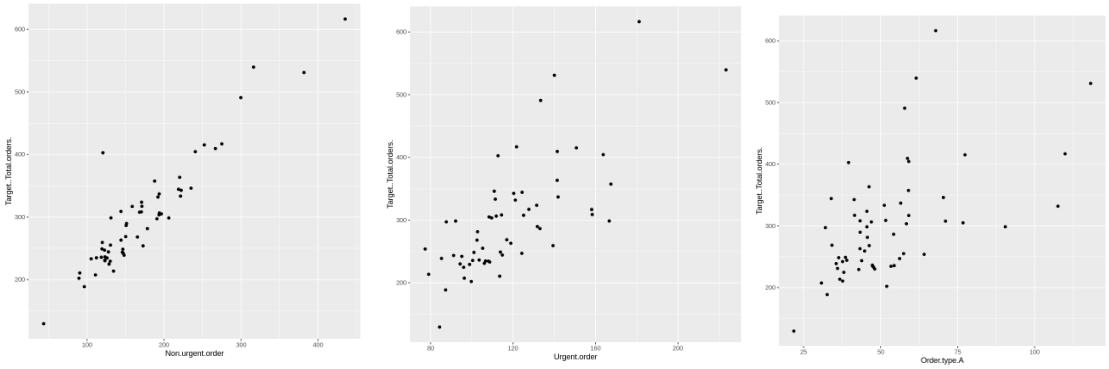
King Mongkut's University of Technology Thonburi
Faculty of Engineering, Department of Computer Engineering

**CPE 213 Data Model, 2/2020**

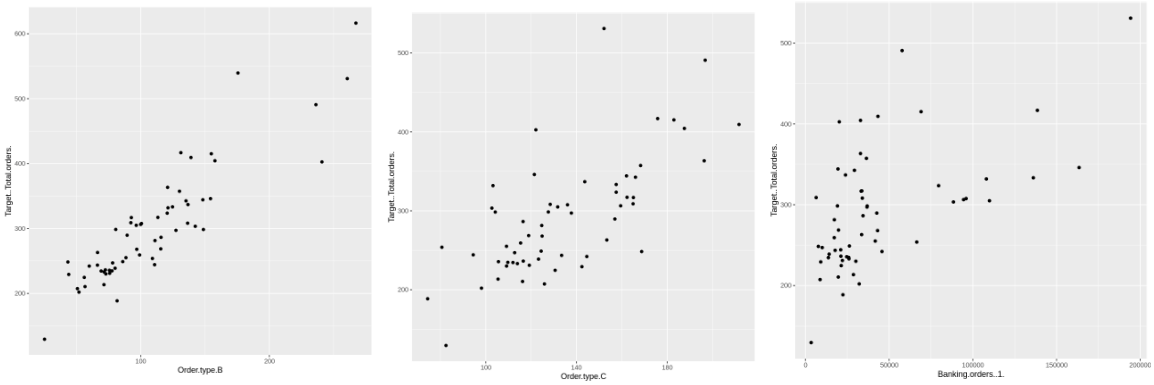
LAB Lecture 9: Linear Regression

Assign Date: 12 April 2021 Due Date: 22 April 2021

---




**รูปที่ 3 แสดง Scatter plot ของ Non-urgent-order กับ target (ซ้าย) , Urgent-order กับ target (ตรงกลาง) และ Order-type-A กับ target (ขวา)**




**รูปที่ 4 แสดง Scatter plot ของ Order-type-B กับ target (ซ้าย) , Order-type-C กับ target (ตรงกลาง) และ Banking-order-1 กับ target (ขวา)**

โดยจากการทดลอง จะเห็นได้ว่าบางความสัมพันธ์จะมี Outlier อยู่มากเลยทำการ Outlier ออกด้วย IQR จากการทดลองได้ทำการเอา Outlier Urgent-order ออกอยู่ค่าเดียวเพราะส่งผลต่อ Correlations ในทางที่ดีขึ้น จากรูปที่ 3

King Mongkut's University of Technology Thonburi
Faculty of Engineering, Department of Computer Engineering
**CPE 213 Data Model, 2/2020**

LAB Lecture 9: Linear Regression
Assign Date: 12 April 2021 Due Date: 22 April 2021

---

| | target |
|---|---|
| | <dbl> |
| **Non.urgent.order** | 0.9072533 |
| **Urgent.order** | 0.6245509 |
| **Order.type.A** | 0.6024456 |
| **Order.type.B** | 0.8920109 |
| **Order.type.C** | 0.6981497 |
| **Banking.orders..2.** | 0.7349711 |
| **Target..Total.orders.** | 1.0000000 |

**รูปที่ 5 Correlation ของข้อมูลที่สัมพันธ์กับ Target ที่ทำการเลือก หลังจากเอา Outlier ออก**

โดยเมื่อนำค่า Outlier ออกทำให้ Correlation บางค่ามีค่าลดลงแต่ ค่าส่วนมากจะเพิ่มขึ้น ดังนั้นเลยได้ใช้ข้อมูลนี้ต่อไปในการ Linear Regression model ต่อไป จากรูปที่ 5

King Mongkut's University of Technology Thonburi
Faculty of Engineering, Department of Computer Engineering

**CPE 213 Data Model, 2/2020**

LAB Lecture 9: Linear Regression

Assign Date: 12 April 2021 Due Date: 22 April 2021

---

```
[16] model <- lm(target ~ nonUrgentOrder + orderTypeB + bankingOrder2, df)
     summary(model)


     Call:
     lm(formula = target ~ nonUrgentOrder + orderTypeB + bankingOrder2,
         data = df)

     Residuals:
         Min      1Q  Median      3Q     Max
     -57.059 -16.555   1.022  15.986  47.306

     Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
     (Intercept)    9.824e+01  9.055e+00  10.849 3.53e-15 ***
     nonUrgentOrder 6.440e-01  9.342e-02   6.893 6.14e-09 ***
     orderTypeB     7.313e-01  1.063e-01   6.882 6.39e-09 ***
     bankingOrder2  1.242e-04  1.251e-04   0.993    0.325
     ---
     Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Residual standard error: 22.65 on 54 degrees of freedom
     Multiple R-squared:  0.9111,    Adjusted R-squared:  0.9062
     F-statistic: 184.5 on 3 and 54 DF,  p-value: < 2.2e-16

[17] target_pred = predict(model,df)

[18] sum((target_pred - df$target)^2)

     27711.8422317505
```

**รูปที่ 6 model ที่ 1**

โดยสมการของ model ที่ 1 จากรูปที่ 6 จะได้เป็น target = 0.644nonUrgentOrder + 0.7313orderTypeB + 0.0001242bankingOrder2 + 0.9824 ซึ่งจะได้ค่า S หรือ Sum error square เท่ากับ 27711.84

```
[19] model <- lm(target ~ nonUrgentOrder + orderTypeB, df)
     summary(model)


     Call:
     lm(formula = target ~ nonUrgentOrder + orderTypeB, data = df)

     Residuals:
         Min      1Q  Median      3Q     Max
     -53.459 -14.716  -0.678  15.860  49.657

     Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
     (Intercept)    98.18791    9.05360  10.845 2.82e-15 ***
     nonUrgentOrder  0.68725    0.08263   8.317 2.66e-11 ***
     orderTypeB      0.75321    0.10394   7.246 1.49e-09 ***
     ---
     Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Residual standard error: 22.65 on 55 degrees of freedom
     Multiple R-squared:  0.9095,    Adjusted R-squared:  0.9062
     F-statistic: 276.4 on 2 and 55 DF,  p-value: < 2.2e-16

[20] target_pred = predict(model,df)
     sum((target_pred - df$target)^2)

     28217.8195948243
```

**รูปที่ 7 model ที่ 2**

โดยสมการของ model ที่ 2 จากรูปที่ 7 จะได้เป็น target = 0.648725nonUrgentOrder + 0.75321orderTypeB + 98.18791 ซึ่งจะได้ค่า S หรือ Sum error square เท่ากับ 28217.819 จะสามารถสังเกตุได้ว่า ค่า Adjusted R-squared ไม่ได้เพิ่มขึ้นและ S ก็เยอะมากขึ้นด้วย ซึ่งอาจจะหมายความว่า อาจจะไม่ความจำเป็นที่ต้องบวกสมการ

King Mongkut's University of Technology Thonburi
Faculty of Engineering, Department of Computer Engineering

**CPE 213 Data Model, 2/2020**

LAB Lecture 9: Linear Regression

Assign Date: 12 April 2021 Due Date: 22 April 2021

---

```
[21] model <- lm(target ~ nonUrgentOrder, df)
     summary(model)

     Call:
     lm(formula = target ~ nonUrgentOrder, data = df)

     Residuals:
         Min      1Q  Median      3Q     Max
     -45.826 -21.164  -3.519  11.918 163.158

     Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
     (Intercept)    100.15420   12.53876   7.988  8.1e-11 ***
     nonUrgentOrder   1.15474    0.07153  16.142  < 2e-16 ***
     ---
     Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Residual standard error: 31.38 on 56 degrees of freedom
     Multiple R-squared:  0.8231,    Adjusted R-squared:  0.8199
     F-statistic: 260.6 on 1 and 56 DF,  p-value: < 2.2e-16

[22] target_pred = predict(model,df)
     sum((target_pred - df$target)^2)

     55157.594051875
```

**รูปที่ 8 model ที่ 3**

โดยสมการของ model ที่ 3 จากรูปที่ 8 จะได้เป็น target = 1.15474nonUrgentOrder + 100.15420 ซึ่งจะได้ค่า S หรือ Sum error square เท่ากับ 55157.594 โดยสามารถสังเกตได้ว่า 1 ตัวแปรไม่เพียงพอต่อการสร้าง model ถึงแม้ ตัวแปรนี้จะสัมพันธ์กับ Target ของเรามากที่สุด

```
[23] model <- lm(target ~ nonUrgentOrder*bankingOrder2, df)
     summary(model)

     Call:
     lm(formula = target ~ nonUrgentOrder * bankingOrder2, data = df

     Residuals:
         Min      1Q  Median      3Q     Max
     -48.750 -18.596  -4.412  12.071 160.687

     Coefficients:
                                  Estimate Std. Error t value Pr(>|
     (Intercept)                 1.134e+02  2.763e+01   4.103 0.000
     nonUrgentOrder              9.349e-01  1.843e-01   5.073 4.95e
     bankingOrder2               1.393e-04  3.478e-04   0.401 0.690
     nonUrgentOrder:bankingOrder2 8.972e-07  1.677e-06   0.535 0.594
     ---
     Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Residual standard error: 30.96 on 54 degrees of freedom
     Multiple R-squared:  0.8341,    Adjusted R-squared:  0.8248
     F-statistic: 90.47 on 3 and 54 DF,  p-value: < 2.2e-16

[24] target_pred = predict(model,df)
     sum((target_pred - df$target)^2)

     51745.8325507857
```

**รูปที่ 9 model ที่ 4**

โดยสมการของ model ที่ 4 จากรูปที่ 9 จะได้เป็น target = 8.972e-07nonUrgentOrder + 1.393e-04bankingOrder2 + 8.972e-07bankingOrder2*nonUrgentOrder + 1.134e+02 ซึ่งจะได้ค่า S หรือ Sum error square เท่ากับ 51745.832

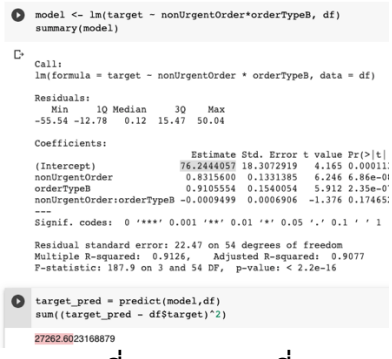
King Mongkut's University of Technology Thonburi
Faculty of Engineering, Department of Computer Engineering

**CPE 213 Data Model, 2/2020**

LAB Lecture 9: Linear Regression

Assign Date: 12 April 2021 Due Date: 22 April 2021

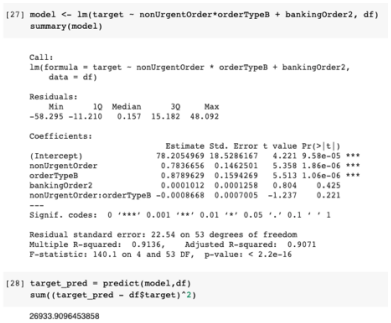
---



```
model <- lm(target ~ nonUrgentOrder*orderTypeB, df)
summary(model)

Call:
lm(formula = target ~ nonUrgentOrder * orderTypeB, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-55.54 -12.78    0.12   15.47   50.04

Coefficients:
                             Estimate Std. Error t value Pr(>|t|
(Intercept)                76.2444057 18.3072919   4.165 0.00011
nonUrgentOrder              0.8315600  0.1331385   6.246 6.86e-0
orderTypeB                  0.9105554  0.1540054   5.912 2.35e-0
nonUrgentOrder:orderTypeB -0.0009499  0.0006906  -1.376 0.17465
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.47 on 54 degrees of freedom
Multiple R-squared:  0.9126,    Adjusted R-squared:  0.9077
F-statistic: 187.9 on 3 and 54 DF,  p-value: < 2.2e-16

target_pred = predict(model,df)
sum((target_pred - df$target)^2)

27262.6023168879
```

**รูปที่ 10 model ที่ 5**

โดยสมการของ model ที่ 5 จากรูปที่ 10 จะได้เป็น target = 0.83156nonUrgentOrder + 0.9105554orderTypeB - 0.0009499nonUrgentOrder*orderTypeB + 76.2444057
ซึ่งจะได้ค่า S หรือ Sum error square เท่ากับ 27262.60 ซึ่งจะสามารถสังเกตุเทียบจาก model ที่ 4 ได้ เมื่อทำการเปลี่ยนตัวแปรที่มี ค่า Correlations สูงขึ้นแล้วนำมา Interaction กัน เราจะได้ ค่า Adjusted R-squared ที่สูงกว่าด้วย ดีกว่าการเอาไป additional กัน



```
[27] model <- lm(target ~ nonUrgentOrder*orderTypeB + bankingOrder2, df)
     summary(model)

     Call:
     lm(formula = target ~ nonUrgentOrder * orderTypeB + bankingOrder2,
         data = df)

     Residuals:
         Min       1Q   Median       3Q      Max
     -58.295 -11.210   0.157  15.182  48.092

     Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
     (Intercept)                78.2054969 18.5286167   4.221 9.58e-05 ***
     nonUrgentOrder              0.7836656  0.1462501   5.358 1.86e-06 ***
     orderTypeB                  0.8789629  0.1594269   5.513 1.06e-06 ***
     bankingOrder2               0.0001012  0.0001258   0.804    0.425
     nonUrgentOrder:orderTypeB -0.0008668  0.0007005  -1.237    0.221
     ---
     Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Residual standard error: 22.54 on 53 degrees of freedom
     Multiple R-squared:  0.9136,    Adjusted R-squared:  0.9071
     F-statistic: 140.1 on 4 and 53 DF,  p-value: < 2.2e-16

[28] target_pred = predict(model,df)
     sum((target_pred - df$target)^2)

     26933.9096453858
```

**รูปที่11 model ที่ 6**

โดยสมการของ model ที่ 6 จากรูปที่ 11 จะได้เป็น target = 0.7836656 nonUrgentOrder + 0.8789629orderTypeB + 0.0001012bankingOrder2 - 0.0008668nonUrgentOrder*orderTypeB + 78.2054969 ซึ่งจะได้ค่า S หรือ Sum error square เท่ากับ 26933.91 ซึ่งเมื่อทดลองนำ ค่าที่ Correlation สูง ๆ มาบวกเพิ่ม ก็ไม่ได้ให้ผลที่ต่างกันมากจากเดิม เมื่อเทียบกับ model ที่ 5

King Mongkut's University of Technology Thonburi
Faculty of Engineering, Department of Computer Engineering

**CPE 213 Data Model, 2/2020**

LAB Lecture 9: Linear Regression

Assign Date: 12 April 2021 Due Date: 22 April 2021

---

```
model <- lm(target ~ nonUrgentOrder*orderTypeB + orderTypeA, df)
summary(model)
```

```
Call:
lm(formula = target ~ nonUrgentOrder * orderTypeB + orderTypeA,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-53.581 -12.529   0.153  16.774  39.158

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                61.4909820 18.1403032   3.390  0.00133 **
nonUrgentOrder              0.7519976  0.1292424   5.819 3.52e-07 ***
orderTypeB                  0.9657336  0.1470257   6.568 2.23e-08 ***
orderTypeA                  0.5080484  0.1867807   2.720  0.00881 **
nonUrgentOrder:orderTypeB -0.0011624  0.0006577  -1.768  0.08290 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.25 on 53 degrees of freedom
Multiple R-squared:  0.9233,    Adjusted R-squared:  0.9175
F-statistic: 159.5 on 4 and 53 DF,  p-value: < 2.2e-16
```

```
[30] target_pred = predict(model,df)
     sum((target_pred - df$target)^2)
```

```
23923.0579573635
```

**รูปที่ 10 model ที่ 7**

โดยสมการของ model ที่ 7 จากรูปที่ 10 จะได้เป็น target = 0.7519976 nonUrgentOrder + 0.9657336orderTypeB + 0.5080484orderTypeA - 0.0011624orderTypeB*nonUrgentOrder + 61.490982 ซึ่งจะได้ค่า S หรือ Sum error square เท่ากับ 23923.05 ซึ่งจะสามารถสังเกตุเทียบจาก model ที่ 6 ได้ เมื่อทำการเปลี่ยนตัวแปรที่มี ค่า Correlations ต่ำกว่า แต่ว่ากลับจะได้ S ที่มากขึ้น นั้นแปลว่าตัวแปรนี้อาจเป็นตัวแปรที่มีความสำคัญแต่ไม่สามารถแสดงในรูปของความสัมพันธ์ได้ ถ้านำมา interaction อาจจะให้ผลที่ดีขึ้น

King Mongkut's University of Technology Thonburi
Faculty of Engineering, Department of Computer Engineering

**CPE 213 Data Model, 2/2020**

LAB Lecture 9: Linear Regression

Assign Date: 12 April 2021 Due Date: 22 April 2021

---

```
model <- lm(target ~ nonUrgentOrder*orderTypeB + orderTypeA*orderTypeC, df)
summary(model)
```

```
Warning message in summary.lm(model):
"essentially perfect fit: summary may be unreliable"

Call:
lm(formula = target ~ nonUrgentOrder * orderTypeB + orderTypeA *
    orderTypeC, data = df)

Residuals:
       Min         1Q     Median         3Q        Max
-4.363e-14 -1.458e-14 -3.309e-15  7.638e-15  1.391e-13

Coefficients:
                            Estimate Std. Error    t value Pr(>|t|)
(Intercept)               -1.129e-13  5.249e-14 -2.151e+00   0.0362 *
nonUrgentOrder            -1.296e-16  2.292e-16 -5.650e-01   0.5743
orderTypeB                 1.000e+00  2.053e-16  4.870e+15   <2e-16 ***
orderTypeA                 1.000e+00  1.026e-15  9.751e+14   <2e-16 ***
orderTypeC                 1.000e+00  4.390e-16  2.278e+15   <2e-16 ***
nonUrgentOrder:orderTypeB -1.326e-18  9.823e-19 -1.350e+00   0.1829
orderTypeA:orderTypeC      2.125e-19  7.360e-18  2.900e-02   0.9771
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.857e-14 on 51 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 6.369e+31 on 6 and 51 DF,  p-value: < 2.2e-16
```

```
[36] target_pred = predict(model,df)
     sum((target_pred - df$target)^2)
```

```
1.567119519875585e-25
```

**รูปที่11 model ที่ 8**

โดยสมการของ model ที่ 8 จากรูปที่ 11 จะได้เป็น target = - 1.296e-16 nonUrgentOrder + orderTypeB + orderTypeA + orderTypeC - 1.326e-18 nonUrgentOrder*orderTypeB + nonUrgentOrder*orderTypeC ซึ่งจะได้ค่า S หรือ Sum error square เท่ากับ 1.57e-25 โดยเมื่อลองนำ ตัวแปลที่ Correlation ไม่สูงมากมา Interaction กัน เราจะได้ model ที่มีค่า Adjusted R-squared เท่ากับ 1 ซึ่งเป็น model ที่ fit กับข้อมูลมาก ๆ จน Overfit โดยสามารถสังเกตุได้อีกย่าง คือ ตัว Feature orderTypeB orderTypeA orderTypeC มีค่าความชันเป็น 1 ซึ่งมันอาจหมายความว่าตัวแปรเหล่านี้เป็นตัวแปรที่กำหนดสมการหลัก ในแปรผันตาม Target ของเรา

King Mongkut's University of Technology Thonburi
Faculty of Engineering, Department of Computer Engineering

**CPE 213 Data Model, 2/2020**

LAB Lecture 9: Linear Regression

Assign Date: 12 April 2021 Due Date: 22 April 2021

---

```
## 75% of the sample size
smp_size <- floor(0.75 * nrow(df))

## set the seed to make your partition reproducible
set.seed(123)
train_ind <- sample(seq_len(nrow(df)), size = smp_size)

train <- df[train_ind, ]
test <- df[-train_ind, ]
```

```
model <- lm(target ~ nonUrgentOrder*orderTypeB + orderTypeA*orderTypeC, train)
summary(model)
```

```
Warning message in summary.lm(model):
"essentially perfect fit: summary may be unreliable"

Call:
lm(formula = target ~ nonUrgentOrder * orderTypeB + orderTypeA *
    orderTypeC, data = train)

Residuals:
       Min         1Q     Median         3Q        Max
-1.748e-13 -9.339e-15  6.000e-15  1.116e-14  5.829e-14

Coefficients:
                              Estimate Std. Error    t value Pr(>|t|)
(Intercept)                  -1.734e-14  7.737e-14 -2.240e-01    0.824
nonUrgentOrder                0.000e+00  3.581e-16  0.000e+00    1.000
orderTypeB                    1.000e+00  2.799e-16  3.573e+15   <2e-16 ***
orderTypeA                    1.000e+00  1.766e-15  5.663e+14   <2e-16 ***
orderTypeC                    1.000e+00  7.465e-16  1.340e+15   <2e-16 ***
nonUrgentOrder:orderTypeB    -1.882e-19  1.545e-18 -1.220e-01    0.904
orderTypeA:orderTypeC         8.072e-18  1.407e-17  5.740e-01    0.570
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.498e-14 on 36 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 3.328e+31 on 6 and 36 DF,  p-value: < 2.2e-16
```

```
target_pred = predict(model,test)
sum((target_pred - test$target)^2)
```

```
2.66571877092284e-26
```

**รูปที่12 ทำ Train test split เพื่อทดสอบ model ที่ 8**

โดยเมื่อลองทำ Train test split ก็ได้ผลลัพธ์ที่ดีมาก จากการทดลอง รูปที่12 ซึ่งจากการทดลองนี้ข้อมูลอาจจะน้อยเกินไป จนอาจจะไม่ reflect ถึง data ที่แท้จริงได้ ทำให้ model มีค่าที่ Fit ขนาดนี้

**Why is sometime adding predictors do not help prediction?**

**Answer** จากการทดลองเราจะเห็นได้ว่า บางครั้งการที่เราเพิ่มตัวแปรที่มี Correlations สูงก็อาจจะไม่ได้ทำให้ model มีประสิทธิภาพสูงเสมอไป แต่เราต้องหาตัวแปรที่เป็น insight ของข้อมูลของเรา จะช่วยเพิ่มประสิทธิภาพของ model ได้ดีมากกว่า ที่จะช่วยในการ predict จริง ๆ