

# Predict Patients' Revisit Tendency by Multiple Linear Regression, Stepwise Selection and Clustering

By Cai Huihan

# Data Description

- **1.1 Datasets**

- 1. bill\_id (3 columns, 13600 entries)
- 2. bill\_amount (2 columns, 13600 entries)
- 3. demographics (5 columns, 3000 entries)
- 4. clinical\_data (26 columns, 3400 entries)

# Data Description

## ○ 1.2 Data Pre-processing

- Merge 4 datasets with patient\_id being the key.
- Reserve latest entry of each patient in clinical\_data
- Calculate total number of visits for each patient from 2012 to 2015
- Remove NAs
- **Final dataset contains 2555 complete data entries.**
- **2012-2014: training dataset (1979 entries)**
- **2015: testing dataset (576 entries).**

# Data Description

## ○ 1.3 Variable Construction

- 1. visit (dependent variable)
- 2. average\_pay
- 3. age at the date\_of\_admission
- 4. bmi:  $\text{weight}/(\text{height}/100)^2$
- 5. Dummies for gender, race, resident\_status, bmi
- **The final dataset contains 43 variables, we selected 32 variables for model constructions.**

# Model Specification

## 2.1 Multiple Linear Regression

```
Call:
lm(formula = visit ~ ., data = data_model)

Residuals:
    Min       1Q   Median       3Q      Max
-0.50585 -0.17396 -0.13058 -0.03426  1.74256

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.382e+00  1.352e-01  10.218 < 2e-16 ***
medical_history_1 1.127e-01  2.010e-02   5.606 2.36e-08 ***
medical_history_2 3.494e-02  1.583e-02   2.206 0.027478 *
medical_history_3 4.384e-02  2.125e-02   2.063 0.039214 *
medical_history_4 5.188e-03  3.206e-02   0.162 0.871476
medical_history_5 3.845e-03  3.025e-02   0.127 0.898863
medical_history_6 7.805e-02  1.703e-02   4.582 4.89e-06 ***
medical_history_7 -5.611e-03  1.689e-02  -0.332 0.739774
preop_medication_1 -1.911e-02  1.483e-02  -1.289 0.197545
preop_medication_2 2.019e-02  1.502e-02   1.344 0.179190
preop_medication_3 7.286e-03  1.976e-02   0.369 0.712362
preop_medication_4 8.342e-03  1.490e-02   0.560 0.575718
preop_medication_5 4.664e-03  1.958e-02   0.238 0.811733
preop_medication_6 1.984e-02  1.705e-02   1.163 0.244856
```

```
symptom_1      4.099e-02  1.542e-02   2.657 0.007938 **
symptom_2      7.686e-02  1.594e-02   4.822 1.53e-06 ***
symptom_3      1.008e-01  1.512e-02   6.667 3.40e-11 ***
symptom_4      4.673e-02  1.682e-02   2.778 0.005524 **
symptom_5      1.760e-01  1.662e-02  10.591 < 2e-16 ***
lab_result_1   -1.848e-03  4.207e-03  -0.439 0.660513
lab_result_2   -2.048e-03  2.972e-03  -0.689 0.490820
lab_result_3   -6.336e-04  4.813e-04  -1.317 0.188132
average_pay    -1.748e-05  7.728e-07 -22.624 < 2e-16 ***
age            3.646e-03  5.339e-04   6.829 1.14e-11 ***
Male           9.367e-03  1.485e-02   0.631 0.528160
Chinese        3.110e-02  3.442e-02   0.904 0.366312
Malay          2.039e-01  3.752e-02   5.434 6.22e-08 ***
Indian         1.351e-01  4.067e-02   3.322 0.000909 ***
Singaporean    -4.429e-01  4.002e-02 -11.068 < 2e-16 ***
PR             -3.882e-01  4.218e-02  -9.203 < 2e-16 ***
Overweight     6.208e-02  2.026e-02   3.065 0.002209 **
Underweight    -1.086e-01  2.339e-01  -0.464 0.642508
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3271 on 1947 degrees of freedom
Multiple R-squared:  0.2225,    Adjusted R-squared:  0.2101
F-statistic: 17.98 on 31 and 1947 DF,  p-value: < 2.2e-16
```



# Model Specification

## 2.2 Stepwise Selection

```
Call:
lm(formula = visit ~ medical_history_1 + medical_history_2 +
    medical_history_3 + medical_history_6 + symptom_1 + symptom_2 +
    symptom_3 + symptom_4 + symptom_5 + lab_result_3 + average_pay +
    age + Malay + Indian + Singaporean + PR + Overweight, data = data_model)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.47394	-0.17436	-0.13080	-0.04211	1.74513

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.369e+00	7.621e-02	17.969	< 2e-16 ***
medical_history_1	1.111e-01	1.999e-02	5.559	3.08e-08 ***
medical_history_2	3.454e-02	1.578e-02	2.189	0.02868 *
medical_history_3	4.556e-02	2.116e-02	2.153	0.03144 *
medical_history_6	7.882e-02	1.694e-02	4.654	3.47e-06 ***
symptom_1	4.173e-02	1.536e-02	2.717	0.00664 **
symptom_2	7.686e-02	1.587e-02	4.842	1.39e-06 ***
symptom_3	1.025e-01	1.505e-02	6.810	1.30e-11 ***
symptom_4	4.668e-02	1.671e-02	2.794	0.00525 **
symptom_5	1.768e-01	1.656e-02	10.676	< 2e-16 ***

lab_result_3	-7.056e-04	4.778e-04	-1.477	0.13988
average_pay	-1.751e-05	7.671e-07	-22.822	< 2e-16 ***
age	3.610e-03	5.310e-04	6.800	1.39e-11 ***
Malay	1.759e-01	2.045e-02	8.604	< 2e-16 ***
Indian	1.064e-01	2.526e-02	4.214	2.62e-05 ***
Singaporean	-4.451e-01	3.981e-02	-11.180	< 2e-16 ***
PR	-3.900e-01	4.196e-02	-9.296	< 2e-16 ***
Overweight	6.371e-02	2.009e-02	3.171	0.00154 **

---

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

Residual standard error: 0.3265 on 1961 degrees of freedom

Multiple R-squared: 0.2194, Adjusted R-squared: 0.2127

F-statistic: 32.43 on 17 and 1961 DF, p-value: < 2.2e-16

# Model Specification

## ○ 2.3 Clustering

	patient_id	X3	X4	...	X20
1	00225710a878eff524a1d13be817e8e2	3	3	...	8
2	0029d90eb654699c18001c17efb0f129	2	2	...	19
3	0040333abd68527ecb53e1db9073f52e	3	3	...	14

# Model Specification

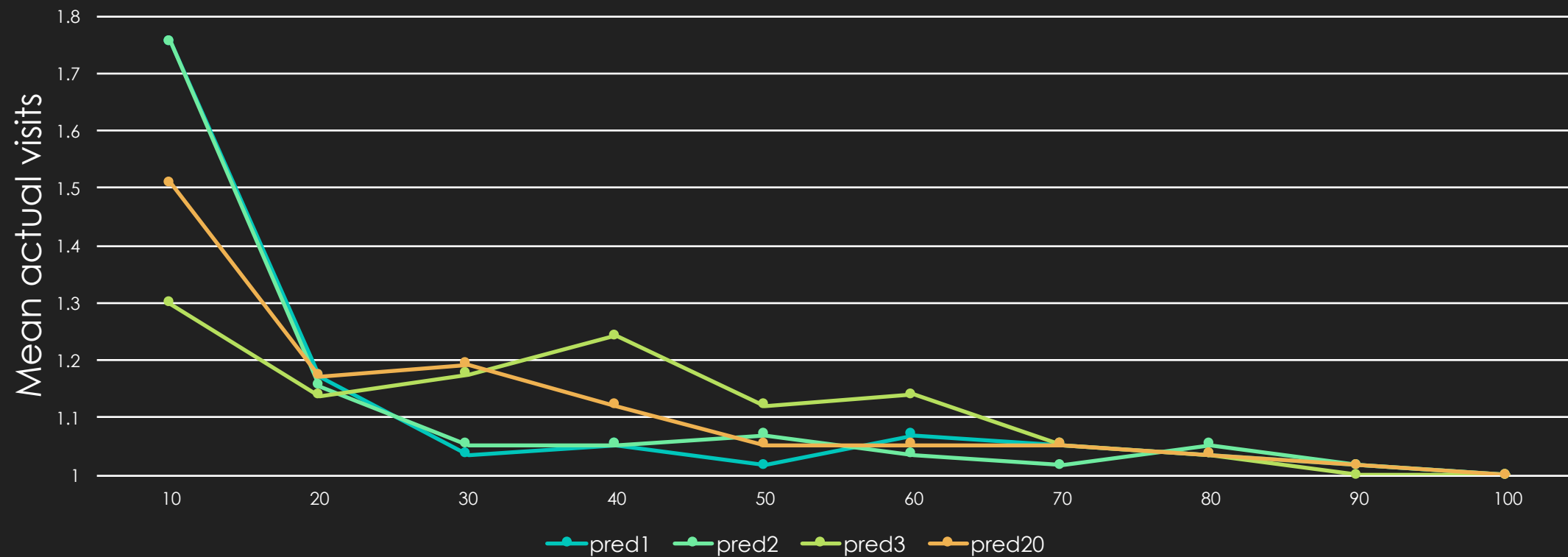
## ○ 2.3 Clustering

Number of clusters/ Cluster number	Tendency Score computed for each cluster			
	3	4	...	20
1	1	1	...	1
2	1.020028612	1.040207523	...	1.117241379
3	1.222896791	1.280046674	...	1.030487805
4	NA	1	...	1
...	...	...	...	...
20	NA	NA	...	1.024193548



# Model Evaluation

Top n percentile of patients ranked by visit tendency scores



# Conclusion

- By identifying groups of customers with higher tendency to be re-admitted to hospital, the hospital could provide consultancy services for the targeted group in advance to
- **Suggest the target patients to do health checks regularly**
- **Supervise on patients' health condition to arrest the growth of disease**