

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)

## 1.2 Data Pre-processing

The provided datasets contain bills, demographics and clinical information for 3000 identical patients from 2012 to 2015. The datasets are merged to facilitate predictive model construction with patient\_id being the key. For clinical\_data, only the latest entry of each patient is reserved, while total number of visits for each patient from 2012 to 2015 is calculated as the dependent variable.

By removing NAs, the dataset contains 2555 complete data entries. We select data entries before 2015 as training dataset (1979 entries) and data entries in 2015 as testing dataset (576 entries).

## 1.3 Variable Construction

The following variables are added to make the data more meaningful.

1. visit: total number of visits of each patient (dependent variable)

2. average\_pay: total bill amount of each patient divided by visit

3. age: age of patient at the date\_of\_admission

4. bmi: weight/(height/100)^2

5. Dummies for gender (Male), race (Chinese, Malay, Indian), resident\_status (Singaporean, PR), bmi(Overweight, Underweight)

The final dataset contains 43 variables, we selected 32 variables for model constructions.

[1] "medical\_history\_1" "medical\_history\_2" "medical\_history\_3" "medical\_history\_4"

[5] "medical\_history\_5" "medical\_history\_6" "medical\_history\_7" "preop\_medication\_1"

[9] "preop\_medication\_2" "preop\_medication\_3" "preop\_medication\_4" "preop\_medication\_5"

[13] "preop\_medication\_6" "symptom\_1" "symptom\_2" "symptom\_3"

[17] "symptom\_4" "symptom\_5" "lab\_result\_1" "lab\_result\_2"

[21] "lab\_result\_3" "visit" "average\_pay" "age"

[25] "Male" "Chinese" "Malay" "Indian"

[29] "Singaporean" "PR" "Overweight" "Underweight"

The following variables are arbitrarily deleted to avoid information overlap.

[1] "patient\_id" "total\_bill" "gender" "race" "resident\_status"

[6] "date\_of\_birth" "date\_of\_admission" "date\_of\_discharge" "weight" "height"

[11] "bmi"

To further improve on variable construction, dummies for lab results should be created. For example, lab\_1\_high, lab\_1\_low, etc. However, due to limit information, we can only use the numerical lab results which may result in inaccurate model interpretations.

# 2. Model Specification

We aim to identify patients with higher tendency to revisit the hospital. With this information, we could target this group of patients to provide certain services. Thus, we construct 3 general types of models:

1. Model based on multiple linear regression
2. Model based on stepwise regression
3. Model based on clustering algorithm

For each model, regardless of the machine learning technique used, we aim to obtain a revisit tendency score. Then we plot a graph of Mean Actual Visits against Top N Percentile of Patients Ranked by Revisit Tendency Scores to evaluate the effectiveness of each model.

## 2.1 Multiple Linear Regression

All 31 variables are used to construct model1. Significant variables are marked with “\*” s.

> summary(model1)

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

From the summary, 22.25% percent of variance can be explained by the model.

It is interesting to observe that

1. medical\_history\_1 is the most significant variable with greatest coefficient in medical histories, followed by medical\_history\_6, indicating that patients with these medical histories are more likely to revisit.

2. preop\_medication\_1\_to\_6 do not significantly affect the number of visits.

3. symptom\_1\_to\_5 all have positive and significant impact on number of visits.

4. average\_pay has a negative and significant impact on visit, implying that the more you pay for medical bills each time, you have less tendency to revisit the hospital.

5. Elderlies tend to visit the hospital more frequently.

6. Malay and Indian visit hospital more frequently than patients from other races.

7. Singaporean and PR visit hospital less frequently than foreigners.

8. People who are overweight have a higher chance to be readmitted into the hospital.

## 2.2 Stepwise Regression

The final model contains 17 variables which are all significant except for lab\_result\_3.

> summary(model2)

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

From the summary, 21.94% percent of variance can be explained by the model.

We can obtain interpretations similar to model1.

By dropping 15 independent variables, the R squared is only dropped by 0.31%. Hence, model2 is considered more efficient than model1 in terms of complexity.

## 2.3 Clustering

With the hypothesis that patients with similar demographic and clinical characteristics would have similar tendency to revisit, patients are clustered based on their demographic and clinical information. Revisit tendency scores are calculated for patients under each cluster based on past visiting behaviors of all patients under this cluster. Potential patients with higher tendency scores are more likely to revisit in the future. Hence, it is viable for hospitals to build connections with these patients.

### 2.3.1 Model Construction

Step1. Cluster all patients based on demographic and clinical records.

* K-means clustering using Euclidean distance is applied to divide zip codes into k clusters where k ranges from 3 to 20.
* The following table displays the resulting clusters for each patient\_id when k = 3 to 20.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | patient\_id | X3 | X4 | … | X20 |
| 1 | 00225710a878eff524a1d13be817e8e2 | 3 | 3 | … | 8 |
| 2 | 0029d90eb654699c18001c17efb0f129 | 2 | 2 | … | 19 |
| 3 | 0040333abd68527ecb53e1db9073f52e | 3 | 3 | … | 14 |

Step2. For each cluster, find the mean number of visits for all patient\_id under the cluster. We are using data from first 2012 to 2014 for training.

* The mean number of visits for each cluster will be used as a tendency score indicating the revisit tendency for each cluster.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |

# 3. Model Evaluation

The 2015 data is used to test the effectiveness of each model.

## 3.1 Multiple Linear Regression

Predict the number of visits for each patient in testing dataset. Rank the patients based on the prediction. Then plot a graph of Mean Actual Visits against Top N Percentile of Patients Ranked by the Prediction. (pred1)

## 3.2 Stepwise Regression

Same method as 3.1(pred2)

## 3.3 Clustering

For each patient\_id, we map it with cluster scores obtained previously. By expectation, the higher the score, the more likely the patient would revisit. Then plot a graph of Mean Actual Visits against Top N Percentile of Patients Ranked by the Tendency Score.(pred3 for k=3, pred20 for k=20)

* To better visualize the model performance, patients are ranked by their tendency scores. The patients are then divided into 10 deciles and the mean actual number of visits of all patients in each decile is calculated.
* For clustering, 18 graphs are plotted for number of clusters from 3 to 20, and 20-means clustering model displays best result.
* It is observed that all curves are downward sloping, which proves our hypothesis that the higher the tendency score, the more frequently the patients would visit the hospital. Compare “pred3” to “pred20”, “pred20” gives better performance for first 3 deciles, hence it is a better predictive model for actual number of visits compare to “pred3”.
* Overall, models based on multiple linear regression renders better performance than clustering models as patients in first 3 deciles have significantly greater number of visits than the other deciles.

# 4. Conclusion

We recommend stepwise linear regression to predict the patients’ revisit tendency.