**1st Project: Logistic Regression Model for Hospital Readmission Within 30 Days After Discharge**

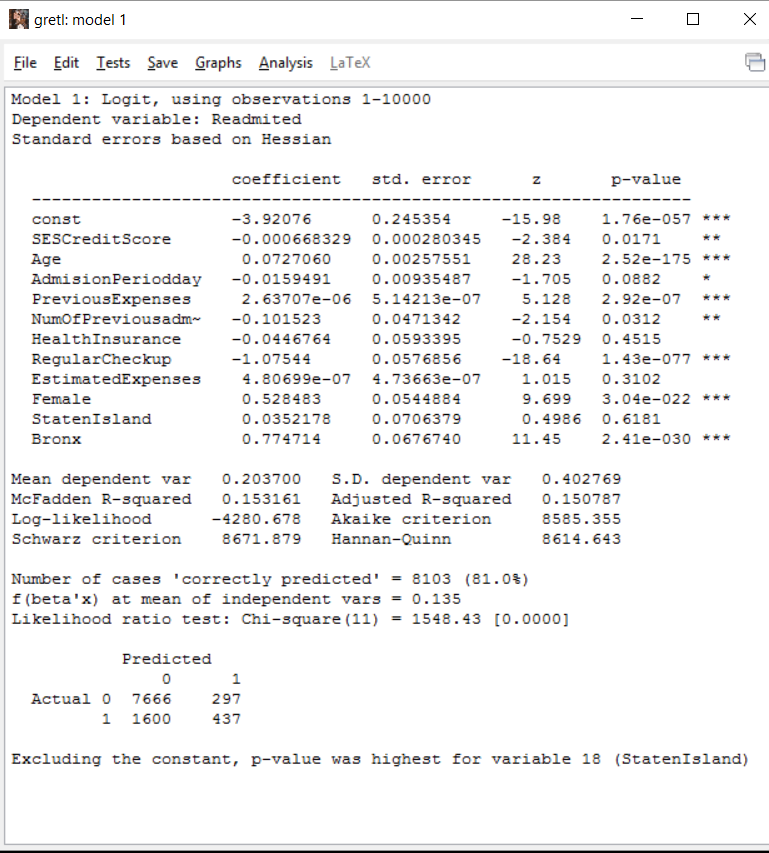
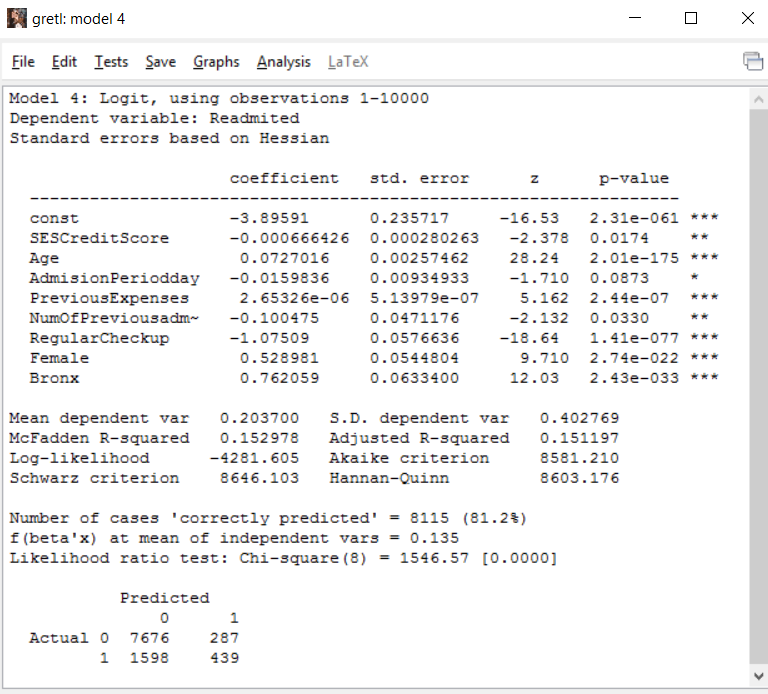
Kansagara and his colleagues conducted systematic review study to examine the regression models for predicting patients’ readmission to hospital. They reviewed 7843 citations. They selected 30 studies of 26 unique models. Most of them used 30-day readmission outcome. The details of the selected literatures were “only 1 model specifically addressed preventable readmissions. Fourteen models that relied on retrospective administrative data could be potentially used to risk-adjust readmission rates for hospital comparison; of these, 9 were tested in large US populations and had poor discriminative ability (c statistic range: 0.55-0.65). Seven models could potentially be used to identify high-risk patients for intervention early during a hospitalization (c statistic range: 0.56-0.72), and 5 could be used at hospital discharge (c statistic range: 0.68-0.83). Six studies compared different models in the same population and 2 of these found that functional and social variables improved model discrimination. Although most models incorporated variables for medical comorbidity and use of prior medical services, few examined variables associated with overall health and function, illness severity, or social determinants of health” (Kansagara, Englander, Salanitro, Kagen, Theobald, Freeman & Kripalani, 2011).

The conclusion is that developing regression models for hospital readmission is still important topic because there is no one universal model for all scenarios, and most of the developed models showed poor results.

1. **Information about the data set**:
2. It consists of 10,000 records of faked patients with 14 fields (columns).
3. The data set was imported in Tableau for predictive analyses. Add hoc A-B test was used to examine and visualize the effect of given variables on the outcome. The outcome is a binary outcome and it is presented by the last field (Readmitted).

Below I provided links for each A-B test visualization which located on my Tableau public website. Also, you can find all visualizations in one place on my GitHub website.

1. The 1st A-B test was for the effect of gender on readmission. <https://public.tableau.com/profile/omar.al.naimi#!/vizhome/HospitalReadmissionProject1/Gender>
2. The 2nd A-B test was the effect of hospital location on readmission. <https://public.tableau.com/profile/omar.al.naimi#!/vizhome/HospitalReadmissionProject2/HospitalLocation>
3. The 3rd A-B test was the effect of having health insurance on readmission. <https://public.tableau.com/profile/omar.al.naimi#!/vizhome/HospitalReadmissionProject3/HealthInsurance>
4. The 4th A-B test was the effect of at least one checkup after discharge on readmission. <https://public.tableau.com/profile/omar.al.naimi#!/vizhome/HospitalReadmissionProject4/RegularCheckup>
5. The 5th A-B test was the effect of the number of previous admissions on readmission. <https://public.tableau.com/profile/omar.al.naimi#!/vizhome/HospitalReadmissionProject5/PreviousAdmission>
6. The 6th A-B test was the effect of the period of previous admission (in days) on readmission. <https://public.tableau.com/profile/omar.al.naimi#!/vizhome/HospitalReadmissionProject6/AdmissionPeriodday>
7. The 7th A-B test was the effect of socio-economic status on readmission. <https://public.tableau.com/profile/omar.al.naimi#!/vizhome/HospitalReadmissionProject7/SES>
8. The 8th A-B test was the effect of age on readmission. <https://public.tableau.com/profile/omar.al.naimi#!/vizhome/HospitalReadmissionProject8/Age>
9. The last A-B test used the last digit of the patient insurance number to evaluate the data set as a sample. <https://public.tableau.com/profile/omar.al.naimi#!/vizhome/HospitalReadmissionProject9/EvaluatingDataSet>
10. **Statistics**: the visualizations those showed difference between the compared parameters have been checked by chi-squared test to acquire statistical significance (if there is a significance). The website <http://www.evanmiller.org/ab-testing/chi-squared.html> and <http://vassarstats.net/> were used to conduct chi-squared test. The results were the following:
11. The 1st A-B test showed that females have more readmission than males (p < 0.001).
12. The 2nd A-B test showed that there are difference between hospital locations (<.0001).
13. The 3rd A-B test showed that health insurance has no effect on readmission (p = 0.48).
14. The 4th A-B test showed that regular checkup has reduced the readmission (p < 0.001).
15. The 5th A-B test showed that 1 previous admission has effect on readmission more than 2 previous admissions (p < 0.001).
16. The 6th A-B test was the effect of the period of previous admission (in days) on readmission.
17. The 7th A-B test showed that socio-economic status has von effect on readmission (0.6714).
18. The 8th A-B test showed that age group has effect on readmission (<.0001).
19. **Model development**: logistic regression model will be used for this case with backward elimination process.
20. The software will be used for this model is Gretl.
21. The dependent variable of the model is “Readmited”.
22. The independent variables are all of them except RowNumber, PatientInsuranceNO, and Surname which obviously have no correlation with the dependent variable.
23. The first result is shown below. The variable that carry the highest p-value will be eliminated with each iteration. the first variable will be removed from the model is “StatenIsland” because it has the highest p-value (0.6181).
24. The valuable information on this table that need special attention are p-values, R-squared, and adjusted R-squared. Whenever you eliminate a variable, watch adjusted R-square. If it goes up then the model becoming better. Otherwise, you should stop eliminating more variable and consider that the given model is might be the best one.



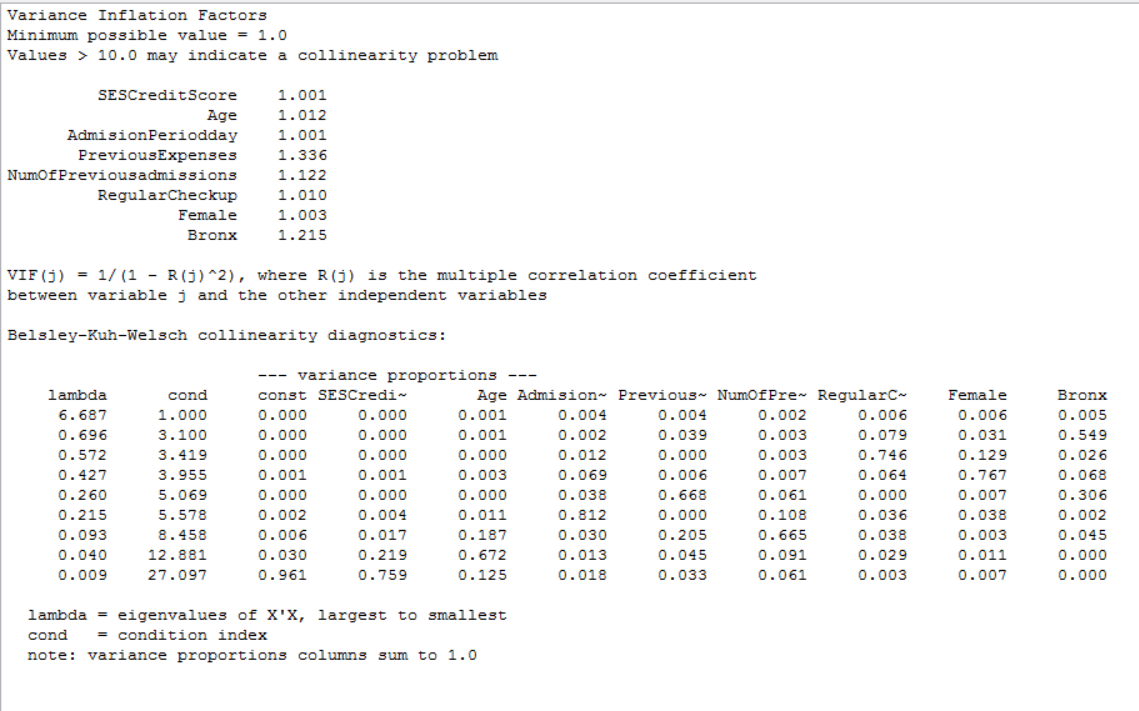
**Fig 1**: first model including all independent variables. **Fig 2**: the final best model.

1. **Model improvement**: repeat the process described in C with variable transformation. The transformation could be square root, X^2, and log of the variables. The chosen type of transformation depends on the business case and deep understanding of the data set.

With each transformation compare the p-value, adjusted R-squared, and accuracy of the new result with the final non-transformed result. Also, derived variables can be created, added to the model, and re-tested. The process takes many iterations and the final best result could be one or group of good models.

1. **Checking for collinearity:** Variance Inflation Factors (VIFs) and correlation matrix can be used to examine the multicollinearity.

Collinearity is risky on model goodness. Therefore, it is important to test it and exclude correlated independent variables.





**Fig 3**: VIFs shows all values about 1. No collinearity.

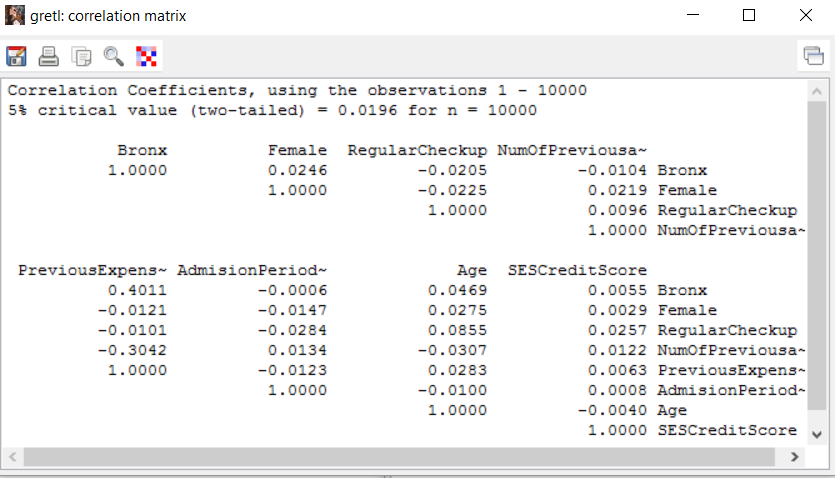




Fig 4: correlation matrix. The maximum value is 0.4 which is mild collinearity.

1. **Model assessment**: using Cumulative Accuracy Profile (CAP) to assess a model, compare it with others, and monitor its deterioration.
2. **Testing the model**: to avoid overfitting of the model, there is need for comparing the performance of training and test data sets. The line graph below (P-hat-TestData) can be compared with the line graph above. Both are very close and that means the model has not committed overfitting.

CAP curve can be used to compare it with future CAP curve from new collected data to evaluate the model stability (improvement, stay the same, or deterioration). If the comparison showed that there is deterioration in the model, then the process of model development should be start over for predicting better model.

The main reason behinds model deterioration is changing factors. The involved factors (independent variables) can be changed regarding their magnitude, interaction with other variables, disappearing, and appearing other variables. Model deterioration is common in biological, psychological, and sociological systems which are very liable to ecosystem disturbance. Therefore, the developed logistic regression model needs for re-evaluation every certain amount of time (depend on subject matter experts) to check its value.

Important note: maintain a model is a process over time. When there is a special event during data collection such as infectious disease epidemic, the collected data during this period cannot be used to re-evaluate the model because disruptive factor has happened.

1. **Logistic regression maintenance**: data scientist can maintain a model like the one above through three steps. The first one is re-assessment. In this case, the new collected sample can be used to evaluate the model based on the same independent variables those have been included in the original model. If the CAP shows stability of the model, then the original model is still valid.

The second step is re-training. The new collected data can be used to train and test the original model.

The third step is rebuilding the model. If the first and second steps showed deterioration in the model, then there is need for new model.

1. **Quantifying the effect of independent variables**: for this purpose, odd ratio is used to quantify the effect of independent variables. The report below shows the amount of effect of the involved independent variables in the logistic regression model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **p-value** | **std. error** | **z** | **coefficient** | **Odds-ratio** |
| const | 2.31E-61 | 0.235716662 | -16.52795571 | **-3.895914555** |  |
| SESCreditScore | 0.017413712 | 0.000280263 | -2.377854146 | **-0.000666426** | **0.9993** |
| Age | 2.01E-175 | 0.002574615 | 28.23784881 | **0.072701596** | **1.0754** |
| AdmisionPeriodday | 0.087340039 | 0.009349328 | -1.709599063 | **-0.015983602** | **0.966** |
| PreviousExpenses | 2.44E-07 | 5.14E-07 | 5.162189348 | **2.65E-06** | **1** |
| NumOfPreviousadmissions | 0.032971389 | 0.047117613 | -2.132431519 | **-0.100475082** | **0.825** |
| RegularCheckup | 1.41E-77 | 0.057663629 | -18.6441984 | **-1.075092143** | **0.3413** |
| Female | 2.74E-22 | 0.054480367 | 9.709578497 | **0.528981403** | **1.6972** |
| Bronx | 2.43E-33 | 0.063340024 | 12.03124151 | **0.76205913** | **2.1427** |

**Future projects**:

2nd project: regression model for predicting medical claim frauds.

3rd project: regression model for predicting people at risk of being obese or any chronic disease.