**Regression Model**

Naixin Zhang

1. Null hypotheses test

According to the project objective, we want to investigate if health insurance status influences the hospital length of stay for patients with asthma conditions. We can do a statistical inference test ANOVA (F-TEST) to compare multiple groups at the same time.Suppose the length of stay in the hospital for four different insurance status is normally distributed with possibly means  **,but all with the equal variance .The study question is whether all four populations are the same. We formulate this question as to the test of hypotheses:**

**Ho:**

**H1: not all four population means are equal.**

**Ho is the null hypothesis. It assumes the four insurance status has the same mean of length of stay in hospital.**

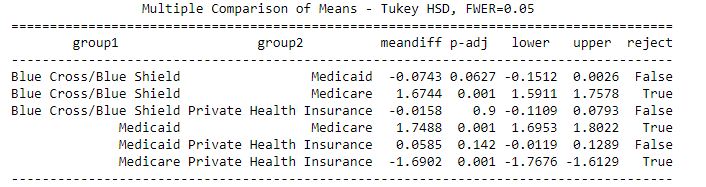
**H1 is the**[**alternative hypothesis**](javascript:void(0);)**, it assumes that not all four insurance status has the same mean of length of stay in hospital.**

**A crucial step in null hypothesis testing is finding the likelihood of the sample result if the null hypothesis were True. This probability is called the p value. A low p value means that the sample result would be unlikely if the null hypothesis were true and leads to the rejection of the null hypothesis. A high p-value means that the sample result would be likely if the null hypothesis were true and leads to the retention of the null hypothesis. Furthermore, we use the criterion called α to determine how low must the p-value be before the sample result is considered unlikely enough to reject the null hypothesis. We define . If there is less than a 5% chance of a result as extreme as the sample result if the null hypothesis were true, then the null hypothesis is rejected. When this happens, the result is said to be statistically significant.**

**After calculation by python, we got the p-value is 0.0, which is smaller than , It means we can reject the null hypothesis, and there is the significant difference among four different insurance status for the length of stay in hospital.**

**2. Post-hoc test:** Turkey's test

**Table 1** Turkey's test

****

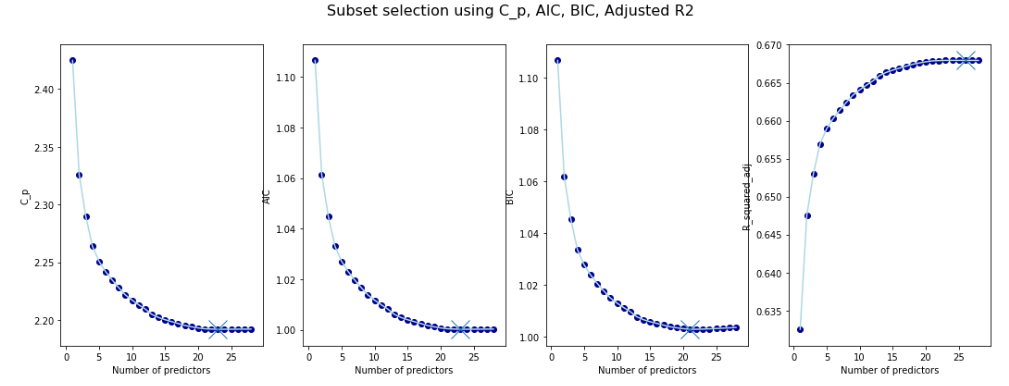
**After the** null hypotheses test, we conclude that the four different health insurance status has a significant influence on the length of stay in hospital. But we may wonder does the four insurance statuses are all significant differences between each other. we can use one of the most popular post-hoc test--- Turkey's test to investigate it. The purpose of Tukey’s test is to figure out which groups in your sample differ. It uses the "Honest Significant Difference", a number that represents the distance between groups, to compare every mean with every other mean.

According to table 1, we know the three groups: Blue Cross/Blue Shield and Medicaid, Blue Cross/Blue Shield and Private Health Insurance, Medicaid and Private Health Insurance have no statistical difference in length of stay in hospital. The other three groups: Blue Cross/Blue Shield, Medicaid and Medicare, Medicare and Private Health Insurance have a statistical difference in length of stay in hospital.

3. Multi-factor line regression.

3.1 variable selection

**Figure1 Subset selection**

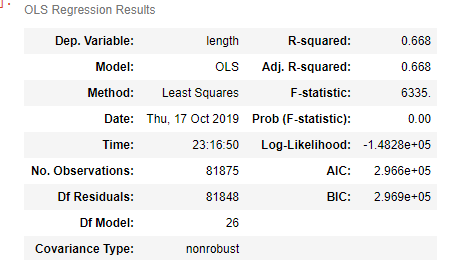


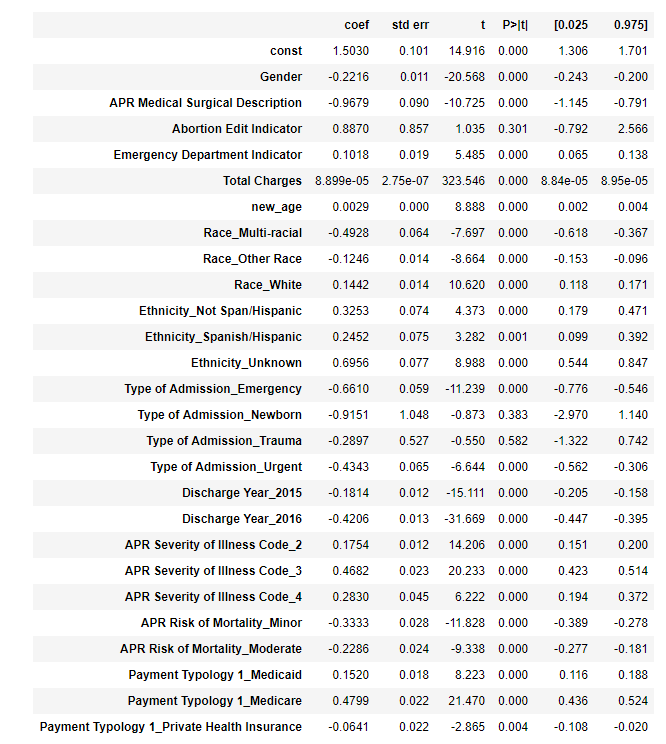
Since we have 37 variables, before running the regression, we'd better make the variable selection to simplify the model. Based on our large dataset, we choose the forward step wise selection. Forward Stepwise begins with a model containing no predictors, and then adds predictors to the model, one at the time. At each step, the variable that gives the greatest additional improvement to the fit is added to the model.

According to figure1, we select the 23 variables that have the lowest Mallow's Cp, Akaike's Information Criteria (AIC) and Bayesian Information Criteria(BIC), highest Adjusted R2.

3.2 Regression

**Table2 OLS regression result**





To further investigate how different insurance status influences the length of stay in the hospital, we can do the regression.

Since most of the independent variables are qualitative variables, we need to be careful of the multicollinearity issue. To solve this problem, in general, if a qualitative variable has k different values, we need to introduce k-1 0-1 variables. For example, for gender variable, k = 2, we can add a variable equal to 1 if the observation is male, 0 for female.

The regression model:

y=β0+β1x1+β2 x2 + β3D1+ β4 I1 + β5 A1 + β6 R1 +β7 R2+ β8R3 +β9 M1 + β10 M2 + β11 S1 +

β12S2 +β13S3 + β14 E1 + β15 E2 +β16 E3 + β17 Sur + β18 Y1+  β19 Y2 + β20 T1 + β21 T2 + β22 T3 + β23 p1+ β24 p2 + β25 p3

where

* β0 is the baseline, represent the length of stay for patient with Black/African American race , Multi-ethnic, Ethnicity, Elective type of admission, discharge year is 2014, minor severity of illness, extreme risk of mortality, Blue Cross/Blue Shield health care insurance.
* x1 is the Total Charges
* x2 is age
* D1 is gender, D1 = 1 represents male and D1 = 0 represents female,
* I1 is 'Emergency Department Indicator', I1 = 1 represents in Emergency department, I1 = 0 represents not in Emergency department.
* A1 is Abortion Edit Indicator, A1 = 1 represents has abortion history, A1 = 0 represents does not have abortion history
* R1 = 1 represents Race White
* R2 = 1 represents Other Race
* R3 = 1 represents Multi-racial
* M1 =1 represents Minor Risk of Mortality
* M2 = 1 represents the Moderate Risk of Mortality
* S1 is APR Severity of Illness Code\_3, represents major Severity of Illness
* S2 is APR Severity of Illness Code\_2, represents moderate Severity of Illness
* S3 is APR Severity of Illness Code\_4, represents extreme Severity of Illness
* E1 is Ethnicity Unknown
* E2 is Ethnicity Not Span/Hispanic
* E3 is Ethnicity Spanish/Hispanic
* Sur is APR Medical Surgical Description, S1 = 1 represents surgical, S1 = 0 represents Medical treatment
* Y1 is Discharge Year\_2015
* Y2 is Discharge Year\_2016
* T1 is a Type of Admission Emergency
* T2 is a Type of Admission Urgent
* T3 is a Type of Admission trauma
* T4 is a Type of Admission Newborn
* p1 is Payment Typology 1\_Medicaid
* p2 is Payment Typology 1\_Medicare
* p3 is Payment Typology 1\_Private Health Insurance

According to the table 2, the regression result is:

y= 1.5030+ 8.899e-05 X1+ 0.0029 X2 -0.2216 D1+ 0.1018 I1 + 0.8870 A1 + 0.1442 R1 -0.1246 R2 -0.4928 R3 -0.3333 M1 + -0.2286 M2 + 0.4682 S1 + 0.1754 S2 + 0.2830 S3 + 0.6956 E1 + 0.3253 E2 + 0.2452 E3 - 0.9679 Sur - 0.1814 Y1 -0.4206 Y2 -0.6610 T1 -0.4343 T2 -0.2897 T3- 0.9151 T4 + 0.1520 p1+ 0.4799 p2 -0.0641p3

General Comments:

Please refer to my slides talking about how to present your research in paper and in presentation. Clarity is the top priority! Define variable before using them (or concurrently); define terms immediately when you introduce them; explain table/figure first before show them;…

It seems that you did some pre-estimation test to motivate (?) your choice of regression model. You may need to reorganize the order of presentation, or at least strengthen the ties between the pre-test and the main model.

Overall, You need to show the flow of thoughts in constructing your regression model. Starting with your research questions, then data, then how that leads to your model selection and the final econometrics model that you will estimate. You may identify the econometric issues in the model construction (e.g. variable endogeneity) and how you will deal with it. You will (very likely) encounter other econometric issues in executing the regression (e.g. homoskedasticity test, weak instrument test…) and you need to figure out how to deal with those issues, which will be part of your next assignment.