Statistical Analysis of Fraud Detection in the Healthcare System

Nagham Saade Abou Jaoude

MSBA399

April 14, 2020

**Contents**

Introduction......................................................................................................................... 3

Data Collection ................................................................................................................... 3

Goals and Variables..................................................................................................... 4

Data Analysis...................................................................................................................... 5

Descriptive Statistics................................................................................................... 5

Correlation .................................................................................................................. 9

t-test at 0.05 significance level.................................................................................. 16

ANOVA..................................................................................................................... 18

Linear Regression ..................................................................................................... 22

References......................................................................................................................... 27

Conclusion........................................................................................................................ 26

**Introduction**

Fraud and abuse are two of the biggest problems in the insurance industry. The total expense of any insurance company tends to increase largely due to fraudulent claims planned by different healthcare providers and the insured. Insurance companies usually develop expert systems to detect fraud that are done by providers, physicians, laboratories, pharmacies and the insured. In these expert systems, data scientists try to identify ambiguous diagnoses used for the most expensive procedures and drugs. Due to that, insurance companies increase their insurance premiums and as a result healthcare is becoming more and more expensive.

Some of the most common types of frauds by providers are:

• Billing for services that were not provided.

• Duplicate submission of a claim for the same service.

• Misrepresenting the service provided.

• Charging for a more complex or expensive service than was actually provided.

• Billing for a covered service when the service actually provided was not covered.

The goal of this study is to predict the potential providers and beneficiaries committing fraud and abuse transactions based on some demographical data and some insurance information.

In the insurance industry, experts try to differentiate between in-hospital claims and out of hospital claims. In this study, the data will be combined and merged with data related to the beneficiary. The data is retrieved from Kaggle; it will be used to analyze the various factors that affect fraud. A statistical analysis is performed such as descriptive analysis, multiple regression, ANOVA, correlation and t-test.

**Data Collection**

As mentioned in the introduction, the data is collected from Kaggle.com. The focus will be on the following factors: state, race, medical procedure, healthcare provider, diagnosis and attending physician. Two data sets will be used : The first one is related to the insurance claims and the second one will be demographic data related to the beneficiary. The two files will be merged based on the beneficiary ID. The analysis will focus on the distribution of each variable to find a correlation between the variables, to compare two groups with t-test and to find the final equation of fraudulent claims using multiple regressions. Below is a table showing different variables in the data files:

**Table1: Table description**

|  |  |
| --- | --- |
| Field | Description |
| BeneID | Beneficairy ID |
| ClaimID | Claim ID |
| ClaimStartDt | Claim Start Date |
| ClaimEndDt | Claim End Date |
| Provider | Healthcare Provider |
| InscClaimAmtReimbursed | Insurance Claim Reimbursed Amount |
| AttendingPhysician | Attending Physician |
| OperatingPhysician | Operating Physician |
| OtherPhysician | Other Physician |
| AdmissionDt | Admission Date |
| ClmAdmitDiagnosisCode | Claim Admission diagnosis code (ICD) |
| DeductibleAmtPaid | Deductible Amount Paid |
| DischargeDt | Discharge Date |
| DiagnosisGroupCode | Diagnosis Group Code |
| ClmDiagnosisCode\_1 | Claim Diagnosis Code |
| ClaimType | Claim Type (In or Out) |
| PotentialFraud | Potential Fraudulent Claim |
| Race | Beneficiary Race |
| State | State |
| DOB | Beneficiary date of birth |

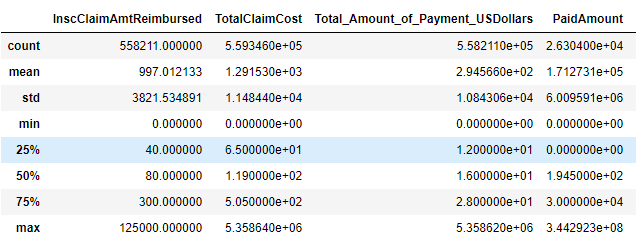
**Data Analysis**

**Descriptive analysis**

1. **Summary statistics**

In the claims data file, there are 558,211 records and 17 fields . Here’s an overall summary statistic:

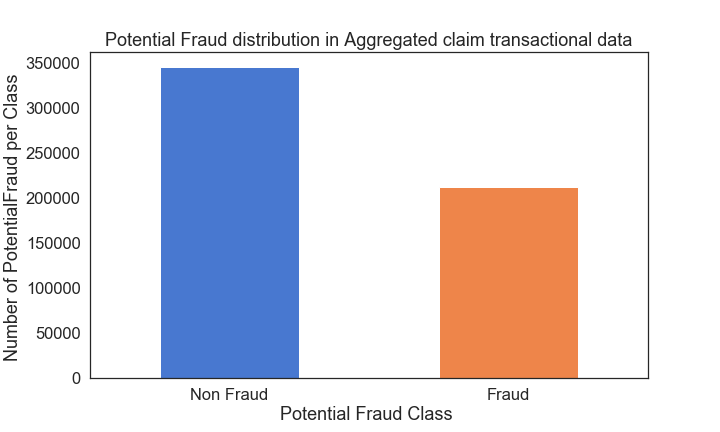
Table3: Claims summary statistics



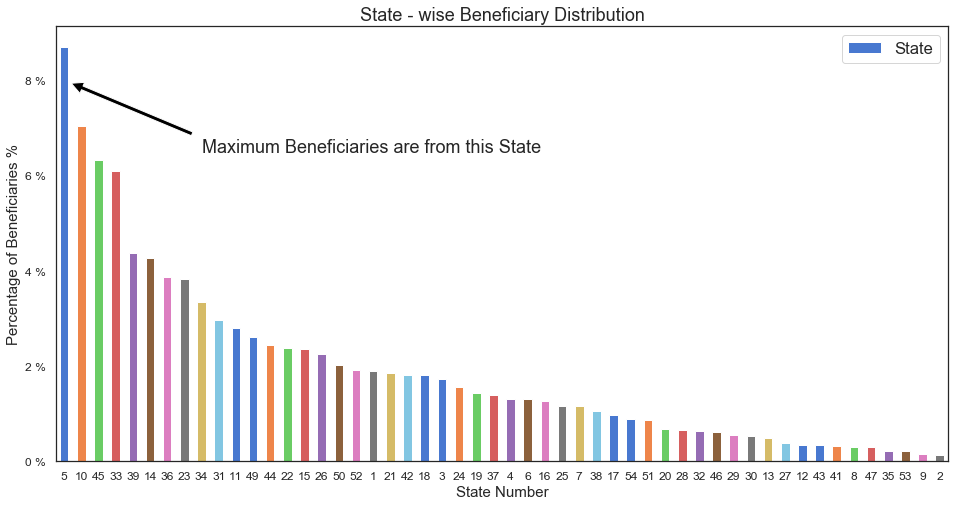
*Figure1*: Summary statistics

On average the number of reimbursed claims is $997.01 with a standard deviation of 3821.53, a minimum value of 0 and a maximum value of $125,000. The average deductible amount paid is $1068. Note that the insurance claim amount reimbursed describes the payment that the healthcare providers receive for giving a medical service from the insurance company. This amount is a good indicator for fraud detection because its opposite amount is related to the rejected claims. Claim rejection reason is usually related to fraud and/or abuse.

1. **Histogram for different variables**

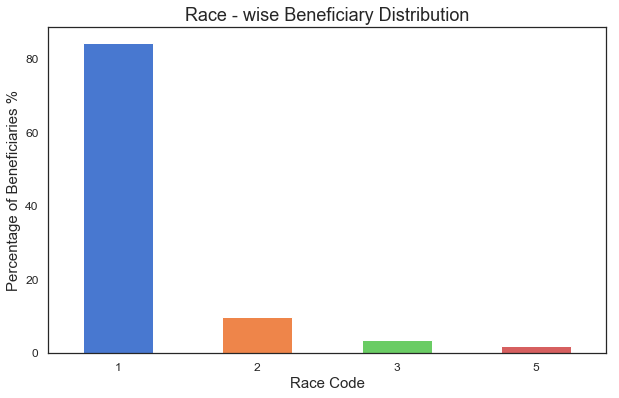
*****Figure 2:*Potential Fraud distribution in aggregated claims

It is clear from the above histogram that more than 50% of the claims are of type potential fraud. These claims need further investigation and will be used to try to understand the behavior of the related parties (providers and beneficiaries).

****

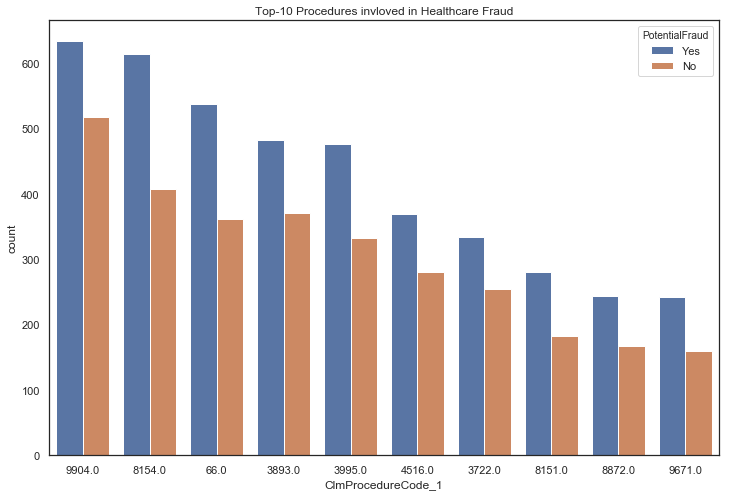
*Figure 3*: State – Wise Beneficiary Distribution

State # 5 shows the maximum number of beneficiaries with a total of 8%. State # 10 is the second one with a total of 7%. State #2 is the least state as in beneficiaries with a total of 2%.

****

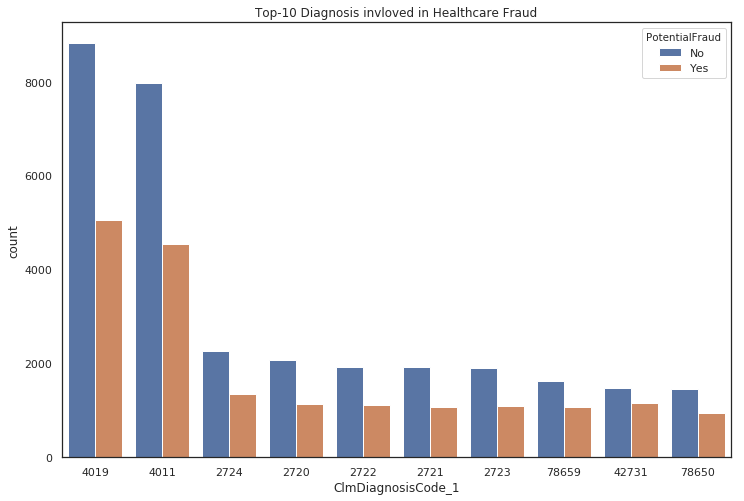
*Figure 4:* Race Beneficiary Distribution

If we take a look at the distribution by race, we can clearly see that more than 80% of the beneficiaries are from race #1. This means that the maximum number of populations in the dataset originated from the same race.



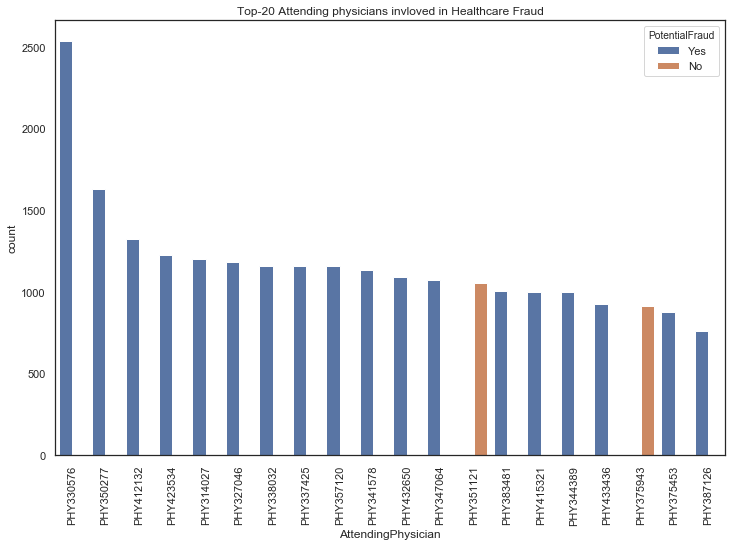
*Figure 5*: Top 10 procedures involved in healthcare fraud

Procedures 9904,8154 and 66 are top procedures involving fraud. In terms of amount in dollars and number of claims. These procedures are the most suspicious as fraudulent claims.



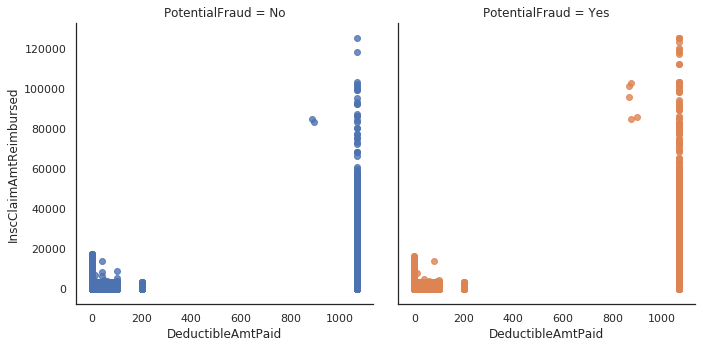
*Figure 6*: Top 10 Diagnosis involved in healthcare fraud.

4019 and 4011 are the top diagnoses that involve suspicious fraudulent claims. Almost 50% of each category is considered to be fraudulent.



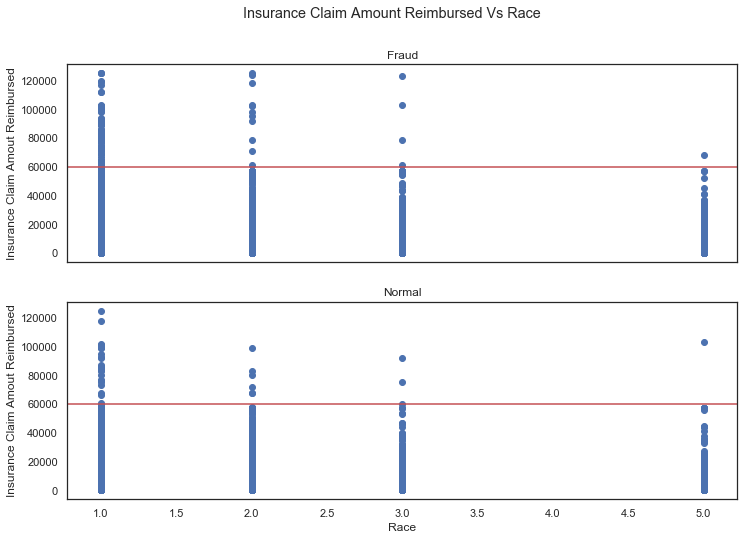
*Figure 7*: Top 20 attending physicians involved in healthcare fraud.

The above chart shows physicians that are involved in a high number of fraudulent claims. PHY330576 is involved in almost 2500 suspicious claims; he is in rank 1. The second physician is PHY350277 with almost 1600 suspicious claims.



*Figure 8:* Deductible Amount Paid Vs Insurance Claim Amount Reimbursed.

From the above 2 graphs it is difficult to differentiate between fraud and non-fraud based on the deductible amount paid and insurance claims reimbursed. The study will show in another graph a combination between a certain beneficiary demographic criterion and insurance reimbursement.

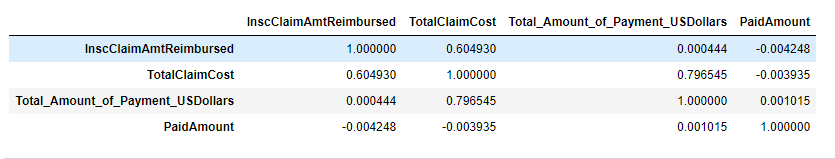


*Figure 9*: Reimbursed insurance amount and beneficiary race.

Race #2 looks like the most suspicious when it comes to fraudulent claims. Note that race is a good factor that the study can apply to make decisions on the background of people with fraudulent behaviors.

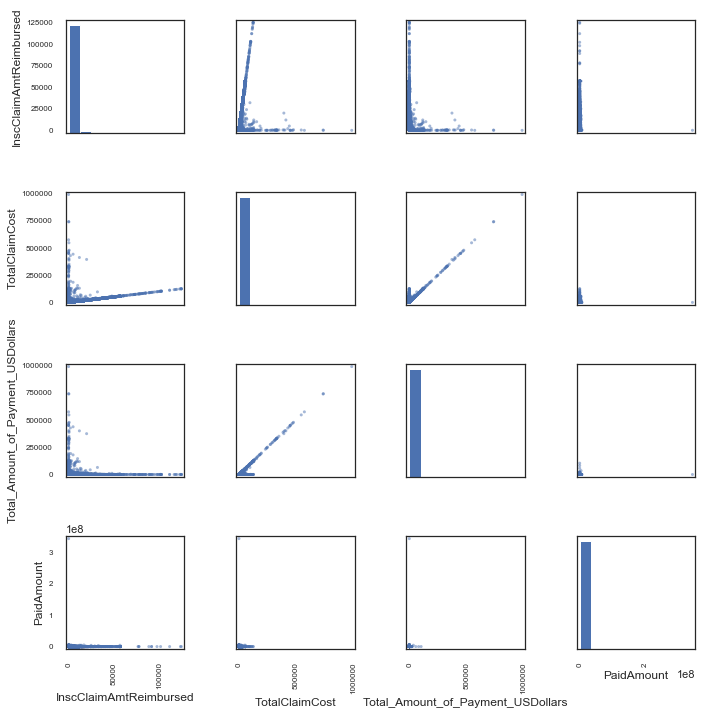
**Correlation**

In this section the study will show a correlation analysis between the measures that are most meaningful in the insurance business for detecting fraud. The data set in filtering only the records that are of type potential fraud. Below is the list of measures to be studied: Reimbursed claim amount, total claim amount and total amount paid. The difference between total amount and the total amount paid could be an indication of fraud. On the other hand, the total amount reimbursed is related to the claims that were stopped from payment then reimbursed to the beneficiary

*Figure 10***:** Correlation Matrix

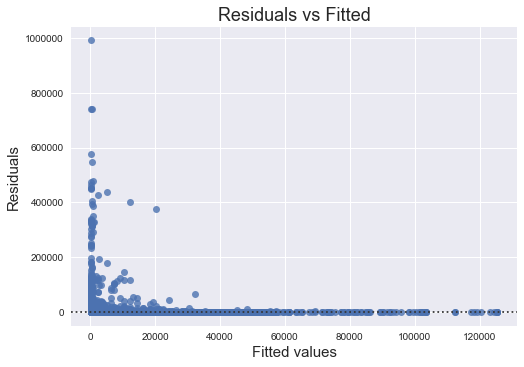
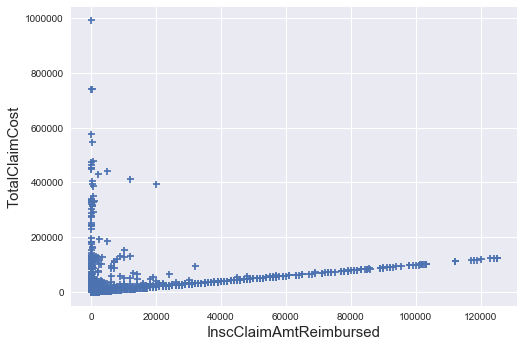
From the above matrix the study shows that the correlation is significant and positive only between claim cost and claim reimbursement (0.604) and between claim cost and claim paid (0.796545).

Below is a matrix plot showing the four measures.



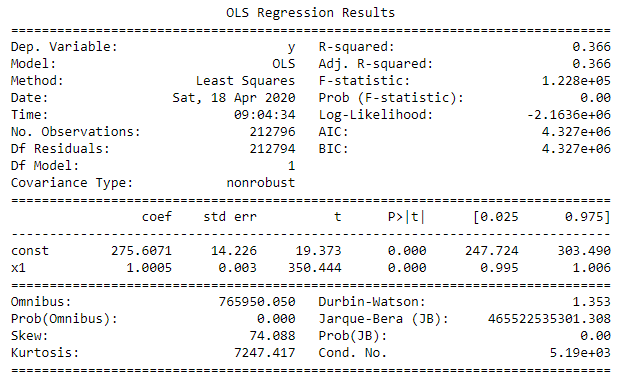
*Figure 11*: Correlation plot

Analysis of claim reimbursement and claim cost: **'InscClaimAmtReimbursed ~ TotalClaimCost':**

****

*Figure 12*: Residual vs. fitted

The relationship between reimbursement and claim cost is not linear. Residual looks random. The correlation is 0.609 and is well explained by reimbursement towards other variables.

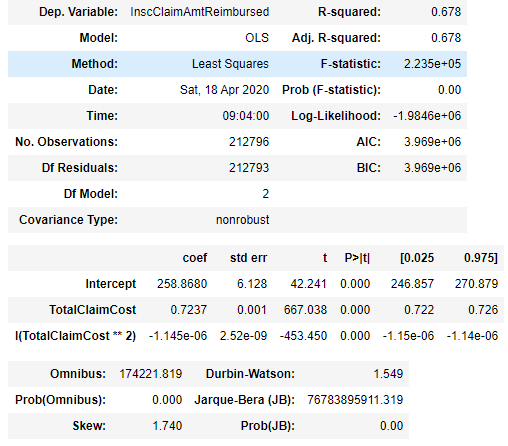


*Table 1*: Regression results

The p-value is very low. The F-statistics is very high with very low p-value suggest that land value is a significance predictor. On the other hand, adjusted R-squared is 0.366 indicates that only 33.6% of variance in reimbursed claims is explained by claim cost.

The regression equation is: Claim cost = 212,794+ 1.321 \* Reimbursement

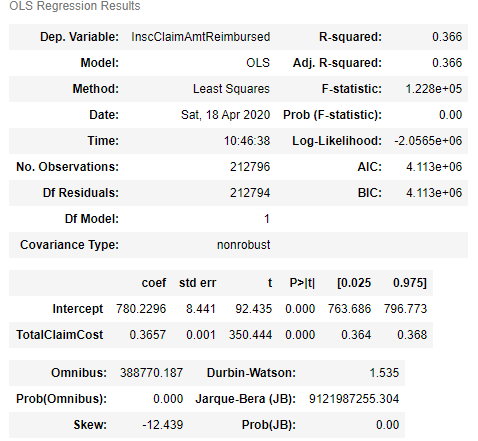
Building the quadratic model: InscClaimAmtReimbursed ~ TotalClaimCost + I(TotalClaimCost\*\*2



*Table 2*: Quadratic model

The intercept is 258.867997 The quadratic regression: Reimbursement amount= 292.055 - 14.368 \* Claim cost + 0.132 \* Claim cost^2. R square is .678. Which means 67.8 %% of the observed data can be explained by the quadratic model. Note: This indicated that the quadratic model is better than the linear model.

Building the Cubic model: Reimbursement amount ~ ClaimCost + I(ClaimCost \*\*2) + I(Claim Cost \*\*3)



The cubic regression: Reimbursement amount = 285.64 + 0.959 \* claim cost - 0.033 \* claim cost ^2 + 0.000176 \* claim cost ^3. R square is 0,366

which means 36.6% of the observed data can be explained by the cubic model. Note: This indicates that model 2 is still better than the others.

**ANOVA to compare between the three models:**

df\_resid ssr df\_diff ss\_diff F Pr(>F)

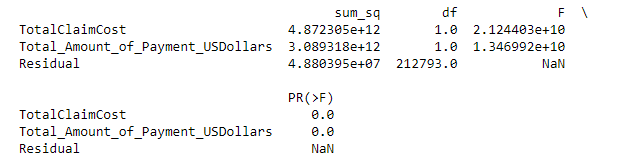
212794.0 3.089366e+12 0.0 NaN NaN NaN

212793.0 1.571176e+12 1.0 1.518191e+12 3.866008e+05 0.0

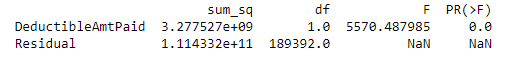
212793.0 8.356433e+11 -0.0 7.355323e+11 -inf NaN

If we try to combine between different measures and we look into the anova results of these combination, the following results will be shown:

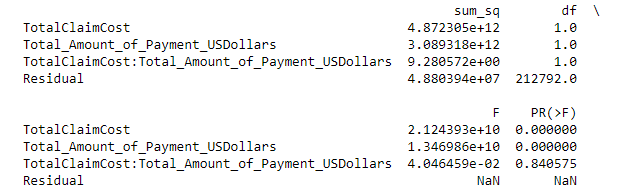
**'InscClaimAmtReimbursed ~ TotalClaimCost + Total\_Amount\_of\_Payment\_USDollars'**

****

**'InscClaimAmtReimbursed ~ DeductibleAmtPaid'**

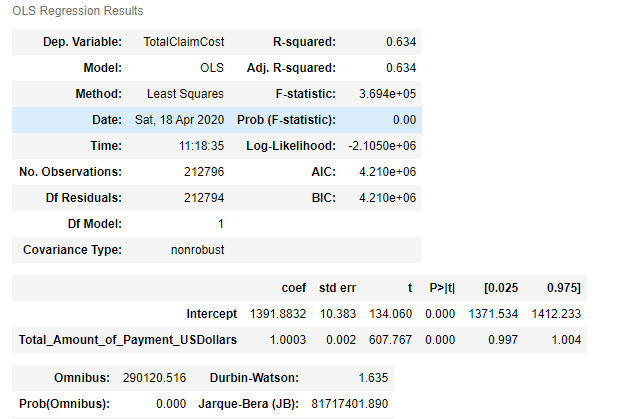
****

**'InscClaimAmtReimbursed ~TotalClaimCost \* Total\_Amount\_of\_Payment\_USDollars'**

****

The study will show the regression between two new variables to see if there’s any analysis we can conclude from it.

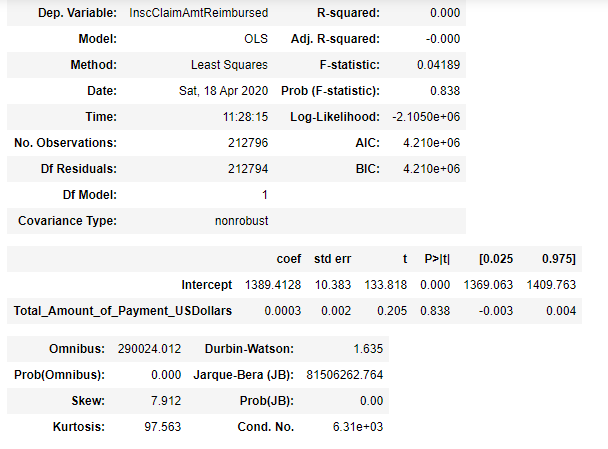
**Claim Cost and Paid Claim shows the below results**

****

Intercept 1391.883213

Total\_Amount\_of\_Payment\_USDollars 1.000333

**Reimbursed amount and paid amount**

****

Intercept 1389.412798

Total\_Amount\_of\_Payment\_USDollars 0.000337

**Conclusion**

The statistical analysis shows that there are different measures that should be considered when trying to investigate fraudulent claims in the healthcare industry. The reimbursement amount is a good factor to look at in combination with the total claims amount. The study also showed the most suspicious characteristics of a fraudulent claims are States 5, 10 and 45 and Race 1 and procedure 9904, 8154 and 66 and Diagnosis 4019 and 4011 and physicians PHY33057, PHY350277 and PH412132. This network of parties that are most probably involved in fraudulent claims are identified with a combination of two measures reimbursement amount and claim cost.

**References**

Bresnick, J. (2019). AHIMA: Data integrity is key for analytics, ICD-10 data maps. Retrieved from: https://healthitanalytics.com/news/ahima-data-integrity-is-key-for-analytics-icd-10-data-maps

Business Intelligence in Healthcare. (2020). Retrieved from https://www.villanovau.com/resources/bi/business-intelligence-in-healthcare/

Jesus, A. (2019). Business Intelligence in Healthcare – Current Applications. Retrieved from https://emerj.com/ai-sector-overviews/business-intelligence-healthcare-current-applications/

Thomas, T. (2020). Top ten common insurance terms. Retrieved from https://www.infinityauto.com/knowledge-center/understanding-insurance/insurance-terms