Capstone Project - High Cost Diseases Insurance

Rodrigo Silva, ASA, FCA

2023-11-24

Contents

1 Introduction

1.1 General Context

"Health High Cost" (HHC from now on) is a group health insurance product that can be offered by Insurance Companies (ICs from now on) in the framework required by the Social Security System (SSS from now on) in Colombia, therefore, it faces the challenges of meeting stringent regulatory requirements imposed by the regulation and oversight bodies (i.e. Government Agencies), as well as market forces, as any other insurance product.

HHC can be bought by some Health Services Providers (HSPs from now on, which are similar to Health Maintenance Organizations in the US [1]), these companies offer health services to its affiliated personnel and their nuclear family, all the people affiliated to the SSS must be covered by a single HSP, active workers and their nuclear family are affiliated to the SSS by the *"contributive"* regime and there is also a *"subsidized"* regime, these regimes differ in their financing scheme but offer the same coverage.

The HHC policy covers high cost events that may be faced by HSPs, with regard of an array of medical conditions such as nervous system surgery, joint replacement, chronic renal failure, HIV - AIDS, cardiac pathologies, among other established conditions. The product is designed to pay catastrophic events, therefore, the deductible and the limit for single events are usually high and of course, there is an aggregated cap for every policy. The IC issues a policy every year, and it is desirable that the policy is renewed year after year, this depends on renewal conditions and general experience in the previous year (i.e. client service, price, etc.), but these topics are outside the scope of the present analysis.

Once the IC receives a claim by the HSP, a reserve is opened and this initial opening amount is the objective of the predictive model studied in this project. If the reserve is too high, at the end of the day the IC faced an opportunity cost associated with the money reserved above the final claim amount and if the reserve is too low, the IC has to constitute the required reserve very quickly in order to pay the claim; a balance between opportunity cost and liquidity should be met.

After this initial reserve is opened, the IC triggers an internal process that includes a risk screening process, a medical auditory review for the case, and an administrative oversight that reviews proper documentation and there is a financial impact because a reserve is opened and subsequently adjusted, depending on the development of the case. If the IC does not receive the documents that must be provided by the HHC two years after the event took place, the process is dismissed and there is no payment; this regulatory requirement means that there are many claims with document filing just before this deadline.

HHC has to be profitable for the IC, and all the conditions required by the SSS must be offered. Every year, the IC negotiates the per capita price of HHC with every HSP, negotiation that takes into account the claim history of the HSP and the general market as well. If the HSP had no claims associated with a certain condition, it does not mean that the condition will not be faced in the future, it cannot be left outside policy coverage and the prime for this condition could take into account this history, but cannot be zero.

It is optional for ICs to offer this product in the market, but if offered, it must cover all the medical conditions as required by the Government.

The data used is the *"Claims File Book"* (CFB from now on) of an IC, data that has not been used before for a rigorous statistical analysis or predictive modeling building. The data in itself, has the following limitations:

- There are open claims, therefore, recent records could be truncated, fact that could introduce a bias towards lower results.
- There is not a huge amount of available data, the CFB contains about 330,000 claim records.

Given the data limitation, it could happen that the model does not have the desired predictive power, however, a data science approach is explored in order to predict the final paid amount per claim. This figure should be the amount used for future reserve openings.

The data was provided by an IC that offers this product, its usage is covered by *edX Privacy Policy* [2], *edX Honor Code* (Effective date: October 28, 2019) [3] and *edX Terms of Service* [4] of this online course [5].

Finally, the data was anonymized in order to not disclose sensitive individual information.

1.2 Model Outcome

The outcome that the model wants to predict is the final claim paid amount of claims made to the IC, this should be the proxy reserve the IC must constitute once the HPS informs the occurrence of an event.

2 Data Cleaning

In the original database, each row represents an event initiated by an HSP, it has the following possibilities:

- The event may result in a prescribed claim, in such a case, there is no payment and the HSP may begin a new claim process. These events are outside the scope of the present analysis.
- The event may trigger more than one claim (e.g. a patient may have required a bone marrow transplant and also a renal transplant), each individual claim is associated with a single medical condition or medical procedure (i.e. row in the original database).

In the first case the final paid amount is 0, because the medical auditory declares an event as not valid or the paperwork was not properly done, or presented beyond the two year limit.

A successful claim is the result of an auditing process and in order to have meaningful data, amounts arising from a single event are indexed to 30-sep-2023 using medical inflation and aggregated, resulting in accumulated amounts for every event (i.e. medical condition). This procedure takes into account that money has time value.

These variables in the final data set are underlined in this document, the following amounts are indexed from InvDate:

IAInv	IAPret	IAObj
IAGlos	IAAud	IAAccObj
IARaisObj		

The numerical variables are named using the following convention:

Convention	**Meaning**
I	Indexed value
A	Amount, monetary value
Acc	Accepted
Aud	Audited
Ded	Deducible
Inv	Invoice
Obj	Objection
Pret	Pretended
Rais	Raised

The final paid amount IAPaid is indexed from DatePayment, all of these variables are indexed amounts (the variable name begins with "I"), they are comparable among them and therefore, they are suitable for further analysis.

The original data was provided in Excel, the treatment was made in VBA, every record is read and there are two cases:

- The record is a prescribed process, therefore, the record is discarded.
- The record is the first of one or many claims, in this case, the described accumulation takes place.

This cleaning means that the original database with 328,794 records (contains data from 1-Mar-2001 to 30-sep-2023), result in 58,854 aggregated and non-prescribed records.

This processes is done on the original data (it is not shared in the deliverables of this project because it has confidential information), once the VBA is polished and a couple records are reviewed, the process runs

surprisingly fast. This process handles the database record by record, therefore, it does not require a lot of RAM and it is better handled by an Excel with 64 bits, it took a little less than 45 minutes to process the whole database.

The final database does not contain confidential information and the obtained final payments can be comparable, in other words, in this point the information is considered to be clean and therefore, exploration can take place.

3 Data Exploration

Every policy is issued by the IC for one year, a policy may face one or many events, each event is constituted by at least one claim. Once an event is informed, its amount is adjusted depending on the deductible, as a result of the auditing process and subsequent documentation, every single claim is thoroughly reviewed and the final amount can be accepted as informed or partially accepted, in this case, a portion of the initial claim is recognized. The recognized amount depends on the event, the medical auditor, the time lapse between the moment when the event occurred and the moment it was reported, etc. these variables are explored in this section.

3.1 Loading Data

Once that R code and data are polished, the whole database is read.

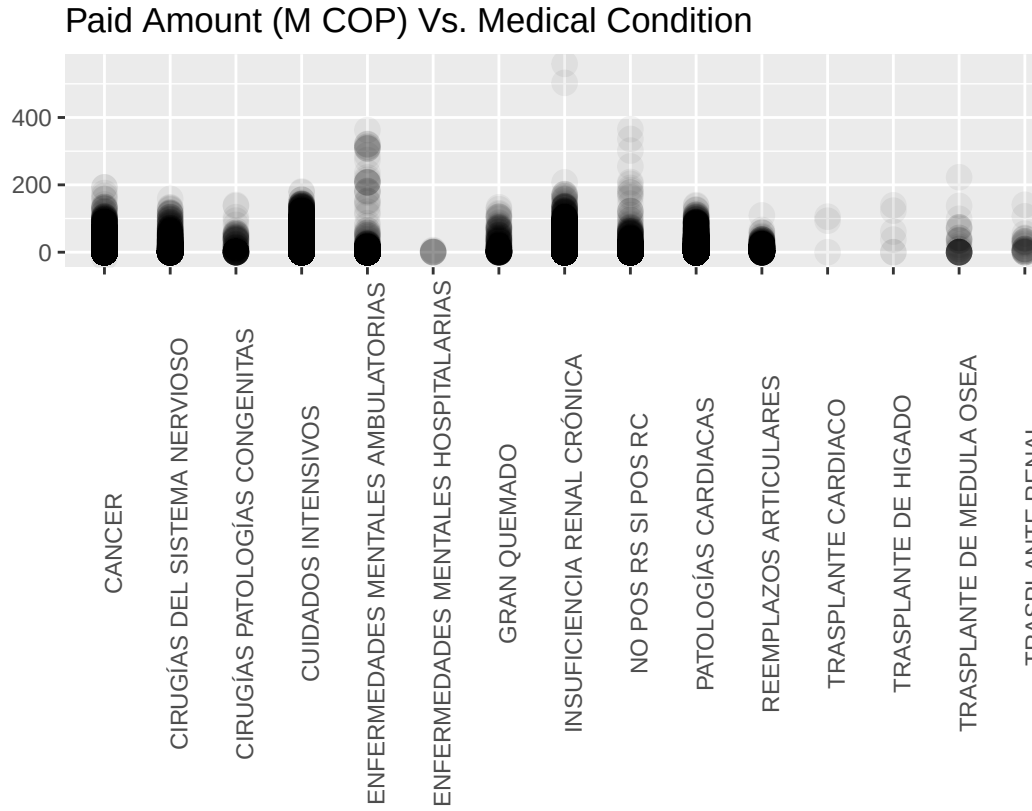
```
############################################################
# Loading Cleaned Data
############################################################
# Sets the working Directory to the path where files are located
setwd(dirname(rstudioapi::getSourceEditorContext()$path))

# Reads the data from XL File
HCData <- read_excel("Insurance.xlsx", sheet="DataV1")

# The first row (i.e. row 0) is the header, it contains appropriate variable names
# The second row is a dummy data row will all the fields with the correct
#   data type and non-empty values, it allows correct type identification
#   by R, and it it  deleted after loading data
HCData <- HCData[-1,]

# Convert MedConditions from char to factors
as.factor(HCData$Cond) -> MedConditions
HCData <- bind_cols(HCData, MedConditions)
```

```
## New names:
## * `` -> `...43`
```

```
HCData <- HCData[, -which(names(HCData) == "Cond")]
colnames(HCData)[42] <- "MedCond"
rm(MedConditions)

# Convert EventState from char to factors
as.factor(HCData$EvState) -> States
HCData <- bind_cols(HCData, States)
```

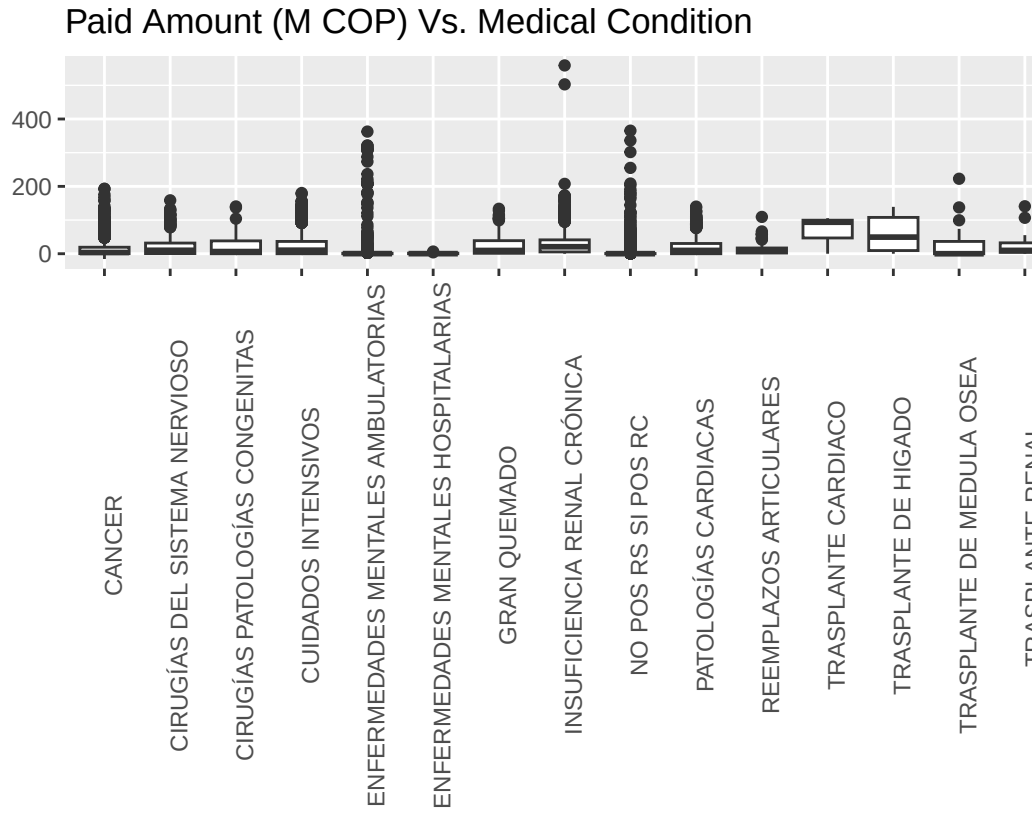
```
## New names:
## * `` -> `...43`
```

```r
HCData <- HCData[, -which(names(HCData) == "EvState")]
colnames(HCData)[42] <- "EvState"
rm(States)

# Generate Test and Train Data
set.seed(1, sample.kind="Rounding") # if using R 3.6 or later
# set.seed(1) # if using R 3.5 or earlier
test_index <- createDataPartition(y = HCData$IDEv, times = 1, p = 0.1, list = FALSE)

HC_Train <- HCData[-test_index,]
HC_Test <- HCData[test_index,]

# PointPlot
HC_Train %>%
  ggplot(aes(x=as.factor(MedCond), y=IAPaid/1000000)) +
  geom_point(alpha = 0.05, size = 4) +
  labs(x="", y="", title="Paid Amount (M COP) Vs. Medical Condition") +
  theme(axis.text.x=element_text (angle=90))
```



Cleaned Data (2th Chunk)-1.pdf

```r
# BoxPlot
HC_Train %>%
  ggplot(aes(x=factor(MedCond), y=IAPaid/1000000), fill=factor(MedCondition)) +
  geom_boxplot() +
  labs(x="", y="", title="Paid Amount (M COP) Vs. Medical Condition") +
  theme(axis.text.x=element_text (angle=90))
```



Cleaned Data (2th Chunk)-2.pdf

From the above graphs, the following conclusions can be obtained:

- The boxes in the boxplot are quite similar among all the conditions, therefore, the medical condition that triggered the claim is not a good descriptor.
- The amount paid is a heavy tailed variable, some payments are quite high (i.e. outliers).

Vioplot was initially obtained but is not a good visualization, all the violins are seen as a single line, it is desirable to have more data.

The payment has a heavy tailed distribution, with some outliers that could be the subject of future review with the IC. In fact, there are six negative payments, that could be data error, but without a criteria for discarding some records, the data as given, was used.

4 Results

According to the correlation matrix, the amount finally paid (IAPaid) is correlated (in decreasing order) with the following variables:

Variable	Correlation
Total Amount Audited	0.9461
Indexed Amount Audited	0.8620
Indexed Amount Pretended	0.5665
Indexed Amount Invoiced	0.4619
Indexed Amount Raised Objection	0.1225

```
# ALL Numerical Columns, columns that contain numbers (no dates), including IndexedAmountPaid (IAPaid)
NumCols <- c("AInsured", "ADed", "AInvTot", "ATAud", "ATObj", "IAInv", "IAPret", "IAObj", "IAGlos", "IA
```

```r
# High Cost Numerical Value, including IndexedAmountPaid (IAPaid)
HCNum_Train <- HCData[colnames(HC_Train) %in% NumCols]
HCNum_Test <- HCData[colnames(HC_Test) %in% NumCols]

cor(HCNum_Train)
```

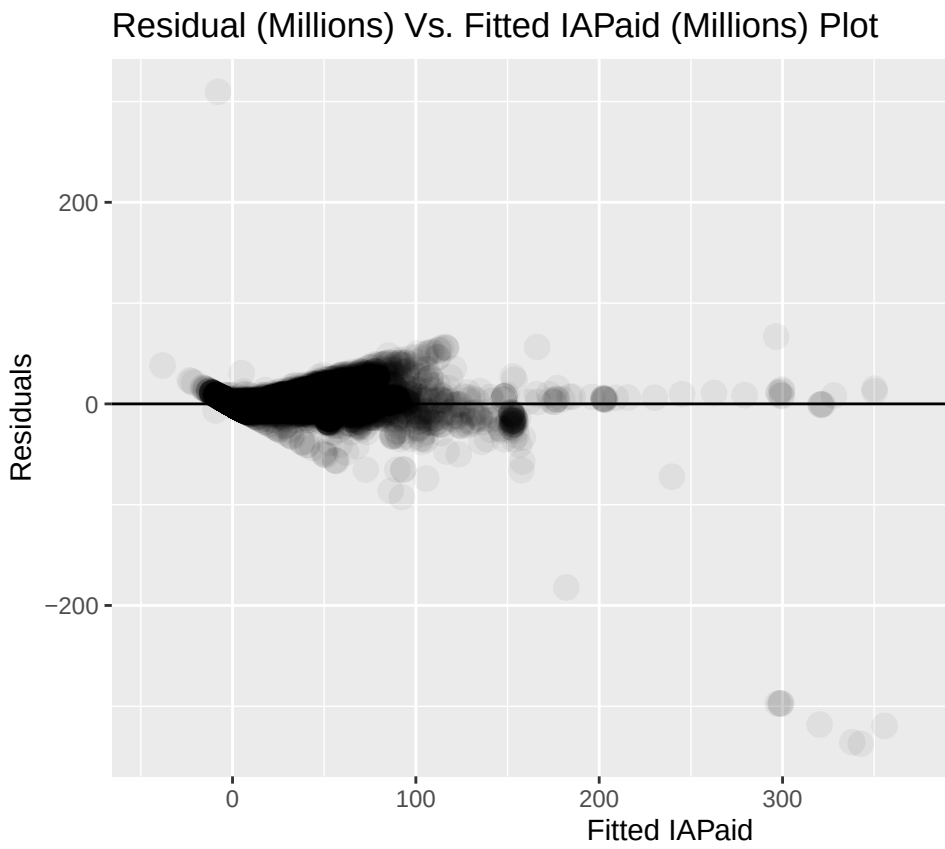
```
##              AInsured        ADed       ATAud       ATObj       IAInv
## AInsured   1.00000000 -0.47796091 -0.40002301 -0.18092084 -0.36933988
## ADed      -0.47796091  1.00000000  0.32134439  0.23844902  0.64576144
## ATAud     -0.40002301  0.32134439  1.00000000  0.05291001  0.50603020
## ATObj     -0.18092084  0.23844902  0.05291001  1.00000000  0.21784294
## IAInv     -0.36933988  0.64576144  0.50603020  0.21784294  1.00000000
## IAPret    -0.32712480  0.44768733  0.62616796  0.35120020  0.78373620
## IAObj     -0.15385446  0.23566608  0.03798483  0.95676200  0.23972902
## IAGlos    -0.05179375  0.08820113  0.01557720  0.04178933  0.09062748
## IAAud     -0.33794437  0.29854360  0.92395967  0.03144446  0.54145696
## IAPaid    -0.46088910  0.29006330  0.94610454  0.05938907  0.46195948
## IAAccObj  -0.02195909  0.02862697  0.05404283  0.04752187  0.05953151
## IARaisObj -0.06202092  0.03921078  0.12086328  0.01271021  0.07452342
##                IAPret       IAObj       IAGlos       IAAud       IAPaid
## AInsured  -0.32712480 -0.15385446 -0.051793746 -0.33794437 -0.460889105
## ADed       0.44768733  0.23566608  0.088201129  0.29854360  0.290063303
## ATAud      0.62616796  0.03798483  0.015577200  0.92395967  0.946104542
## ATObj      0.35120020  0.95676200  0.041789330  0.03144446  0.059389069
## IAInv      0.78373620  0.23972902  0.090627483  0.54145696  0.461959476
## IAPret     1.00000000  0.35880354  0.246635901  0.66378269  0.566471045
## IAObj      0.35880354  1.00000000  0.040230121  0.04889726  0.034848163
## IAGlos     0.24663590  0.04023012  1.000000000  0.01353257  0.007599269
## IAAud      0.66378269  0.04889726  0.013532565  1.00000000  0.862011459
## IAPaid     0.56647105  0.03484816  0.007599269  0.86201146  1.000000000
## IAAccObj   0.22585431  0.04341406  0.322278850  0.04655751  0.039869489
## IARaisObj  0.06415822  0.01808779  0.001997064  0.14677655  0.122516690
##              IAAccObj    IARaisObj
## AInsured  -0.02195909 -0.062020916
## ADed       0.02862697  0.039210778
## ATAud      0.05404283  0.120863275
## ATObj      0.04752187  0.012710211
## IAInv      0.05953151  0.074523420
## IAPret     0.22585431  0.064158216
## IAObj      0.04341406  0.018087795
## IAGlos     0.32227885  0.001997064
## IAAud      0.04655751  0.146776550
## IAPaid     0.03986949  0.122516690
## IAAccObj   1.00000000  0.015863331
## IARaisObj  0.01586333  1.000000000
```

It is quite strange that the total amount audited is more correlated with its total value than with the indexed value, the last figure is the result of the indexation and accumulation process, it has more information.

4.1 Linear Regression

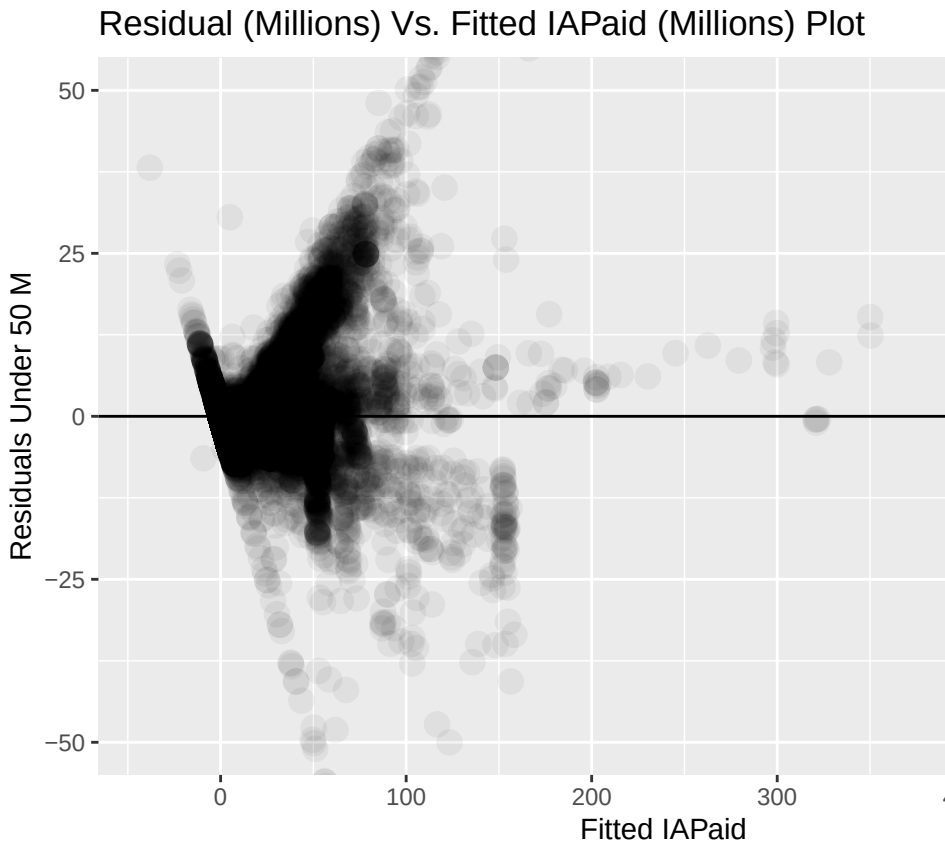
Having discarded the medical condition as a predictive element, a linear regression on the numerical values is performed.

```r
# Numerical Columns, columns that contain numbers (no dates), excluding IndexedAmountPaid (IAPaid)
# IAGlos and IAAccObj are non significant
Model1 <- lm(IAPaid ~ AInsured + ADed + ATAud + ATObj + IAInv + IAPret + IAObj + IAAud + IARaisObj, data

# Residuals Plot (in COP Million)
Model1 %>%
  ggplot(aes(x=.fitted/1000000, y=.resid/1000000)) +
  geom_point(alpha = 0.05, size = 4) +
  geom_hline(yintercept = 0) +
  labs(title = "Residual (Millions) Vs. Fitted IAPaid (Millions) Plot", x="Fitted IAPaid", y = "Residual
```



Regression - First Model (4th Chunk)-1.pdf

```r
Model1 %>%
  ggplot(aes(x=.fitted/1000000, y=.resid/1000000)) +
  geom_point(alpha = 0.05, size = 4) +
  coord_cartesian(ylim=c(-50, 50)) +
  geom_hline(yintercept = 0) +
  labs(title = "Residual (Millions) Vs. Fitted IAPaid (Millions) Plot", x="Fitted IAPaid", y = "Residual
```



Regression - First Model (4th Chunk)-2.pdf

```
summary(Model1)
```

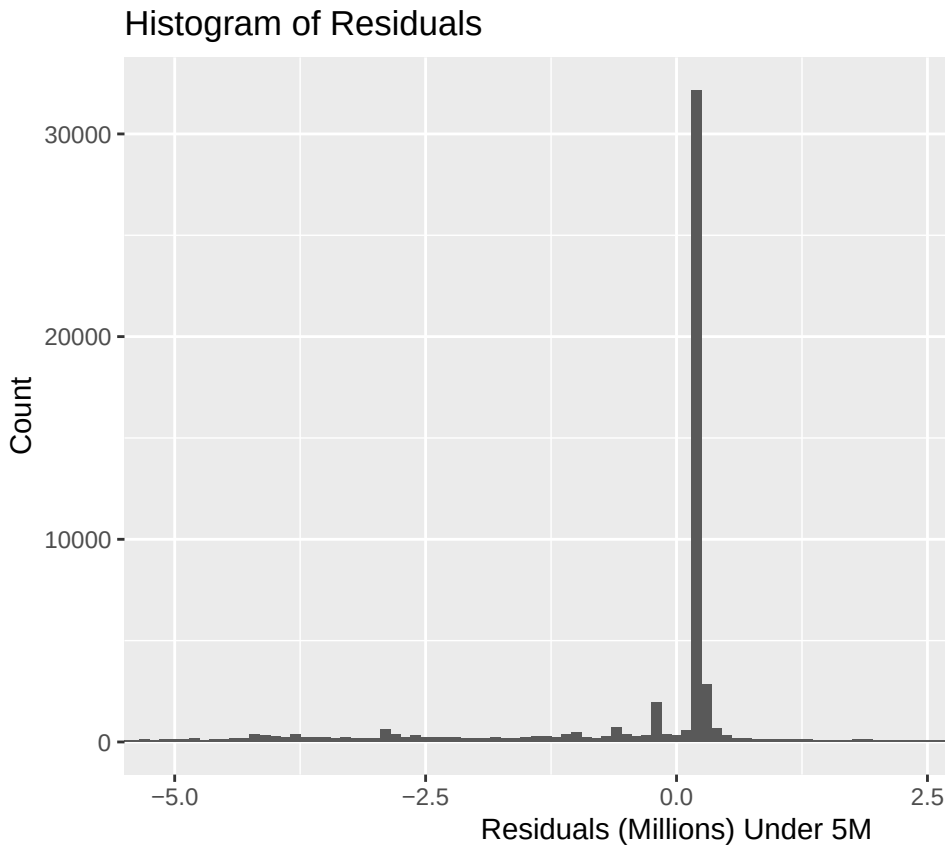
```
##
## Call:
## lm(formula = IAPaid ~ AInsured + ADed + ATAud + ATObj + IAInv +
##     IAPret + IAObj + IAAud + IARaisObj, data = HCNum_Train)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -337194929    -190395     205716     217906  309674253
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.258e+06  8.808e+04  82.398  < 2e-16 ***
## AInsured    -3.731e-02  4.712e-04 -79.188  < 2e-16 ***
## ADed        -8.107e-02  2.191e-03 -37.002  < 2e-16 ***
## ATAud        1.843e+00  6.733e-03 273.730  < 2e-16 ***
## ATObj        2.683e-01  1.379e-02  19.450  < 2e-16 ***
## IAInv        8.992e-03  7.226e-04  12.444  < 2e-16 ***
## IAPret      -3.443e-02  1.552e-03 -22.190  < 2e-16 ***
## IAObj       -1.390e-01  8.333e-03 -16.682  < 2e-16 ***
## IAAud       -2.449e-02  3.779e-03  -6.482 9.14e-11 ***
## IARaisObj    5.250e-02  9.649e-03   5.441 5.33e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 5705000 on 58844 degrees of freedom
## Multiple R-squared:  0.9081, Adjusted R-squared:  0.908
## F-statistic: 6.457e+04 on 9 and 58844 DF,  p-value: < 2.2e-16
```

```
# Extract residuals in COP Millions
residuals <- residuals(Model1)/1e6

summary(residuals)
```

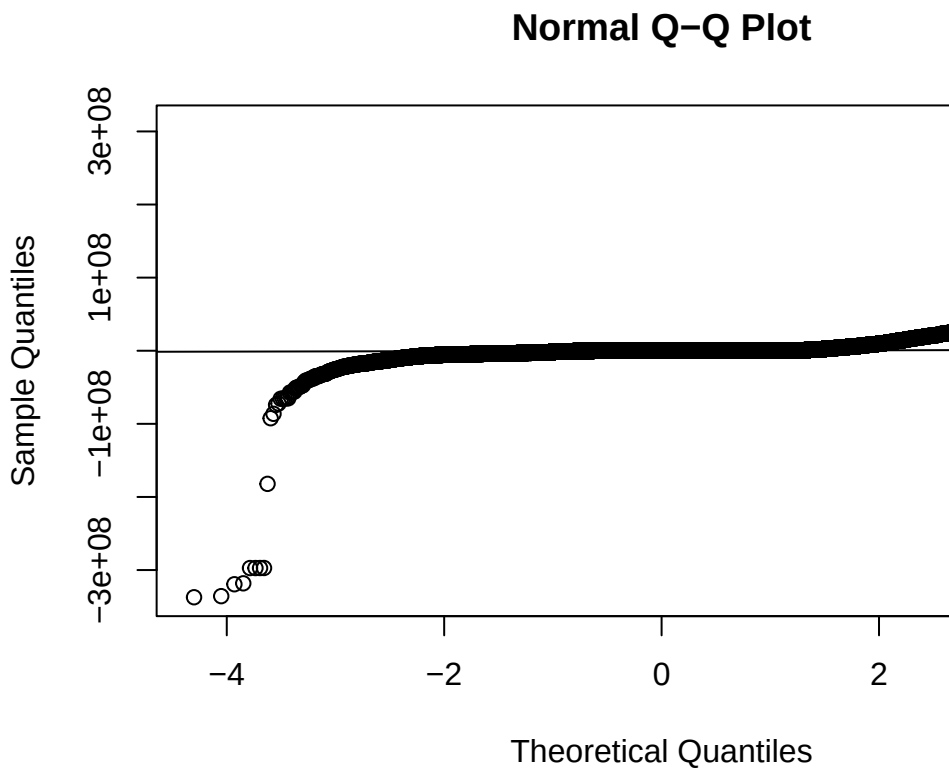
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -337.1949   -0.1904    0.2057    0.0000    0.2179  309.6743
```

```
# Create a dataframe with residuals
residual_df <- data.frame(Residuals = residuals)

residual_df %>%
  ggplot(aes(x = Residuals)) +
  geom_histogram(binwidth = 0.1) +
  coord_cartesian(xlim=c(-5, 5))  +
  labs(title = "Histogram of Residuals", x = "Residuals (Millions) Under 5M", y = "Count")
```



Regression - First Model (4th Chunk)-3.pdf

```
qqnorm(Model1$residuals)
qqline(Model1$residuals)
```



Regression - First Model (4th Chunk)-4.pdf

4.2 Anova Test

Another linear regression with less variables is generated, in order to see if a linear model with less variables (i.e. discarding the variables with lower correlated) is a better fit.

```
# IAAud and IARaisObj have low coefficients
Model2 <- lm(IAPaid ~ AInsured + ADed + ATAud + ATObj + IAInv + IAPret + IAObj, data = HCNum_Train)

# Not a better model, cannot reject null hypothesis
anova(Model1, Model2)
```

```
## Analysis of Variance Table
##
## Model 1: IAPaid ~ AInsured + ADed + ATAud + ATObj + IAInv + IAPret + IAObj +
##     IAAud + IARaisObj
## Model 2: IAPaid ~ AInsured + ADed + ATAud + ATObj + IAInv + IAPret + IAObj
##   Res.Df        RSS Df   Sum of Sq     F    Pr(>F)
## 1  58844 1.9155e+18
## 2  58846 1.9176e+18 -2 -2.1204e+15 32.57 7.293e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

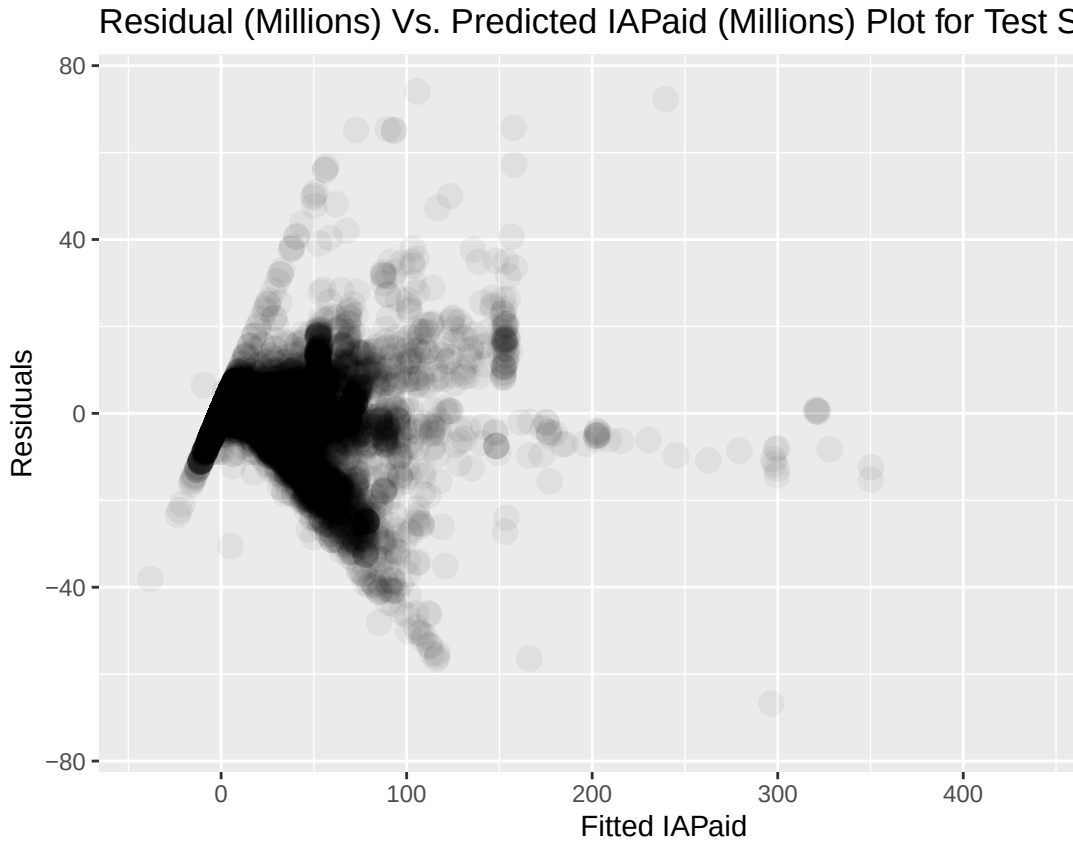
This reduced model is essentially the same than the first, the first model is chosen to predict IAPaid.

4.3 Predicting With the Linear Model

The linear model is used to predict the values in the test set. There are clearly some outliers, that can be seen as dots with lower opacity. There are limits in the graphs, in order to not show outliers.

The fitted values are quite similar to the original values, and there are more dispersed for higher amounts, where there is less data available.

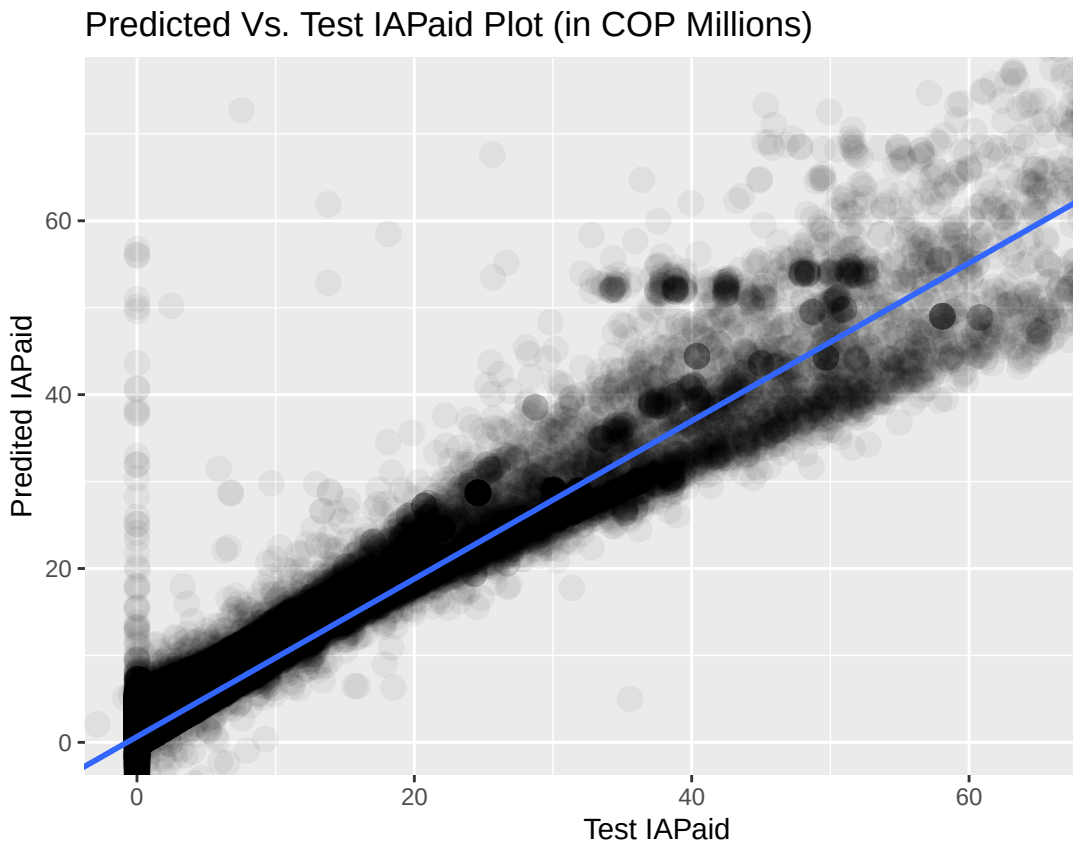
```
HCNum_Test %>%
  mutate(IAPaid_hat = predict(Model1, HCNum_Test)) %>%
  mutate(Error = IAPaid_hat - IAPaid) %>%
  ggplot(aes(x=IAPaid_hat/1000000, y=Error/1000000)) +
    geom_point(alpha = 0.05, size = 4) +
    coord_cartesian(ylim=c(-75, 75)) +
    labs(title = "Residual (Millions) Vs. Predicted IAPaid (Millions) Plot for Test Set", x="Fitted IAPa
```



with Model 1 (6th Chunk)-1.pdf

```
# Use predict to get predictions and intervals
predictions <- predict(Model1, newdata = HCNum_Test, interval = "prediction")

# Create a dataframe with HCNum_Test, fitted Value and Intervals
results <- cbind(HCNum_Test, predictions)

results %>%
  ggplot(aes(IAPaid/1e6, fit/1e6)) +
  geom_point(alpha = 0.05, size = 4) +
  coord_cartesian(xlim=c(0, 75), ylim=c(0, 75)) +
  stat_smooth(method = lm) +
  labs(title = "Predicted Vs. Test IAPaid Plot (in COP Millions)", x="Test IAPaid", y = "Predited IAPaid
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



with Model 1 (6th Chunk)-2.pdf

5 Conclusion

Many steps were taken over the course of this project to design and implement an indexed amount paid by the IC. Real data (i.e. as is) was used, no modification was made on the data (except for the aggregation and indexation process) and there could be data outliers and errors that should be fixed in future iterations of this model, for example, negative paid amounts (there are 6 negative payments) that could be the result of objections made after a payment was done.

The model is simple (just a linear regression), but given the data characteristics, it gives an appropriate fit for the Indexed Amount Paid. The error is negligible in statistical terms, the mean of the residuals is zero (section 4.1, Linear Regression with the training data), and the histogram is almost a big bar around zero; not totally normal, product of the heavy tailed distribution of IAPaid.

This prediction could be useful for the IC in order to have an opening estimate of this amount, and it can also be used to estimate other reserves associated with this event.

The approach of using data for building a statistical model will be the subject of interesting conversations with the IC, data gathering practices, data polishing and final review will help the IC to implement better practices regarding the CFB. The outcome of the model could be used for business decisions, the first obvious one is the reserve amount of informed events, but it could also be the base of other reserves that the IC must constitute. This approach will be more important under IFRS 17, given the fact that the spirit of this accounting standard is to reflect business reality and uncertainty as best as possible.

A Acronyms

The acronyms used in this paper, are alphabetically listed in this section.

- CFB: Claims File Book. Database generated by the insurance company, for every transaction that the insurance product has along the time.
- HHC: Health High Cost. Insurance policy that pays certain health events.
- HSP: Health Services Providers. Companies that offer health services, as required by the Government.
- IC: Insurance Company. Company that may issue HHC insurance policies, in the framework stated by the Government.
- SSS: Social Security System. General Government regulatory framework that oversights the health of the citizens of a country.

B Knit Troubleshoot

The .rmd file was used in different platforms (Windows and Linux), and the obtained graphs are not correctly displayed (they are rendered to the right). These graphs are included in the submission for inspection, they were generated running the code in the Console.

These graphs could not be rendered in a more appropriate way.

C Bibliography

[1] W. F. B. E. Al., *Group insurance, fourth edition.* Actex Publications, 2003.

[2] "edX privacy policy." https://www.edx.org/edx-privacy-policy.

[3] "edX honor code." https://www.edx.org/edx-terms-service#honor-code.

[4] "edX terms of service." https://www.edx.org/edx-terms-service.

[5] "Hardvard university, data science: capstone." https://www.edx.org/course/data-science-capstone.