BZAN540 project

Ali Benek, Ritwik Bhriguvanshi, Chiran Chappagai , Vu Hoang

Executive Report

This study was undertaken to study the relationship between Hospital-acquired conditions (HACs) and patient outcome measures. The study focuses on two measures, length of stay, and total charges incurred for corresponding index hospitalization. The scope of the study was limited to patients with Heart Failure, Heart Attack, and Pneumonia. The investigation employed exploratory data analysis to identify any relationship between the subject of interest and other variables included in the study. The study identified that HACs is prevalent in people with higher age. The subject of study, length of study, and total charges are significantly higher in people with HACs on average. A critical finding from exploratory data analysis was that the number of visits wasn't significantly different in people with or without HACs. We observed that HAC was highly prevalent in heart failure patients and the least prevalent in heart attack patients. Another important finding was that the HACs were mostly observed in hospitals that are relatively smaller in size, with fewer beds for patients' hospitalizations.

Regression was employed to identify whether HAC plays a significant role in explaining the length of stay and total charges. In all the regression models, we observed that the HACs indeed played a significant role in explaining the length of stay as well as total charges. Further, the HACs were also found to be central in explaining the length of stay as well as total charges even when the study attempted to identify variables that might be redundant in explaining our subject of interest. The results of the study from regression analysis can be reasonably interpreted as HACs indeed exhibit a significant relationship with the length of study and total charges.

The study studied the causal relationship between our subject of interest and HACs. The study employed regression adjustments and found that a causal relationship exists between the length of stay as well as total charges. The hospital-acquired conditions increased the length of stay by 18.8% to 19.08% at a 95% confidence interval. However, the imputed values for treatment are much less than the observed values. In contrast, the imputed value of the control case is slightly higher than the observed values indicating that the causal effect might be a conservative estimate. With the inclusion of total charges as one of the variables affecting the length of stay, we observed a downward revision of the causal effect of 11.47%-11.61%, which was also a conservative estimate from the observed values. One important conclusion that could be drawn is that HACs do have a causal relationship with the length of stay.

Further, the study found a causal relationship between total charges and HACs, with HACs, having a positive causal effect of around 19% when the length of stay was excluded in the study. This effect of HAC on total charges was downward revised to 9.11 % when the variable length of stay was included in the study. Either of the cases was a conservative estimate of the causal effect when compared with observed values of control cases. These estimates of causality between HACs and total charges were less conservative than the causality between HACs and length of stay. We can safely conclude that HAC has a causal relationship with total charges as well as the length of stay, and the impact of HAC on length of stay is somewhere between 12.2% to 20.86% and the impact of HACs on total charges is between 9.11% and 19%.

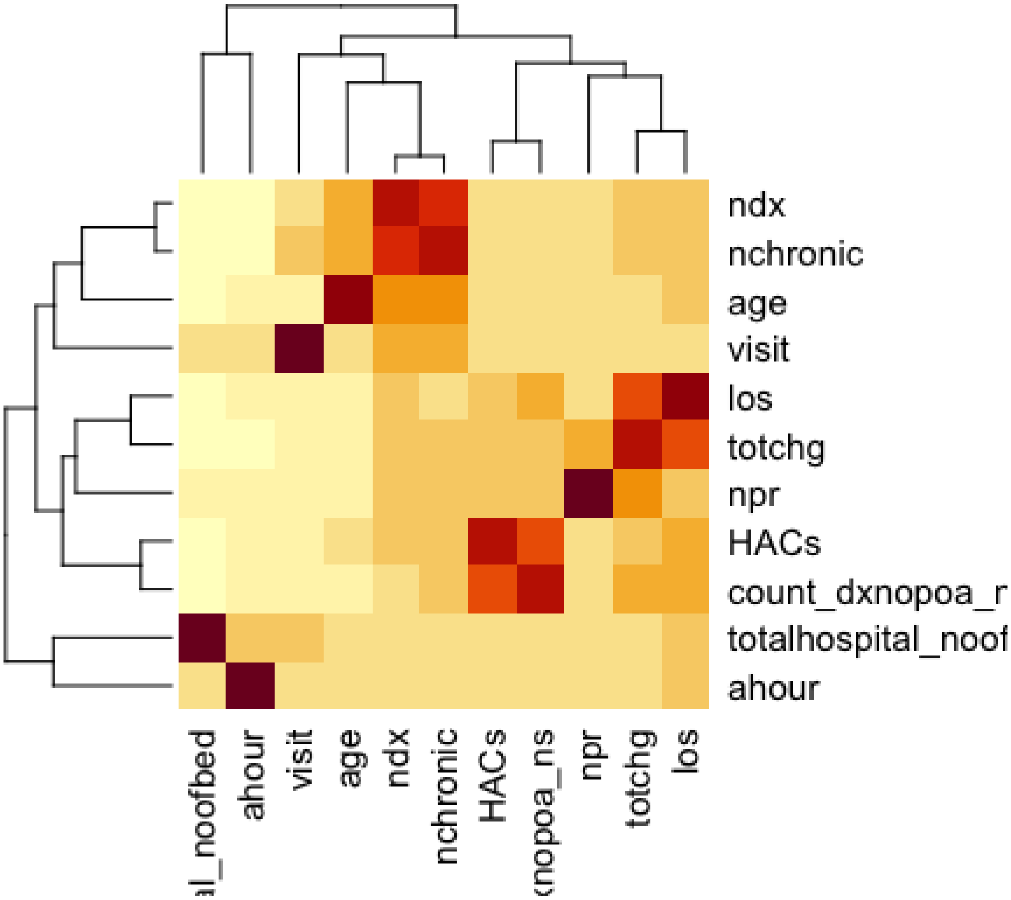
Problem statements

Patients in the US experience Hospital-acquired conditions (HACs) in 3% of the cases with 11% of patients dying during hospitalization. The US government has provided incentives to hospitals to prevent HACs, and therefore understanding the relationship between HACs and outcome measures is crucial in improving patient safety. In this study, we aim to study the relationship between two outcomes measures, the length of hospital stay and total charges to HACs as stated below.

i) The detection of relationship and impact/effect of HACs on length of stay for the corresponding index hospitalization

ii) The detection of relationship and impact/effect of HACs on total charges incurred for corresponding index hospitalization.

Exploratory Data Analysis



```
## [1] 0.6942357
```

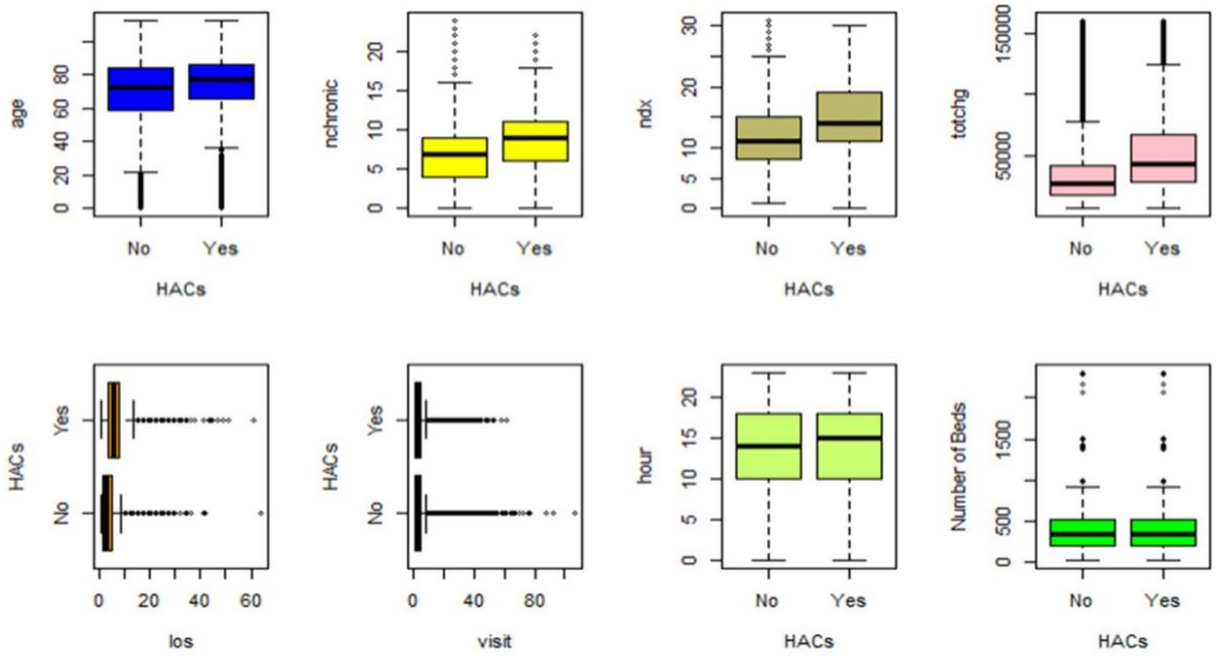
EDA Analysis I: Data preparation is an important aspect before developing a model. As found above, we do not encounter any missing observations in the dataset. The HAC variable is developed from ‘count_dxnopa_ns’ variable, which looks at the difference in the number of diagnostics at admission and discharge. Since, there are no missing values for the same, we do not encounter any missing values for Hospital Acquired Conditions (HACs). Looking at the given dataset, we find that there 72,968 cases of Hospital Acquired Condition.

The dataset contains a mix and numeric and factor variables which we try to analyze and understand before incorporating it in the model. Initially, we observe the heatmap with all the 10 numeric variables in the dataset along with HACs, since it is our main variable of interest for the given case study.

Looking at the heatmap, it does not come across as a surprise that HACs and 'count_dxnopa_ns' are heavily correlated with since the later variable forms the prior variable. Additionally, we see that total charge and length of stay at the hopital show strong association between them. This would be the general assumption as extended stay in hospital will lead to higher costs and vice-versa.

The other two variables which show strong association here are 'ndx' and 'npr'. While 'ndx' states the number of diagnoses, 'npr' states the number of procedures. It is prevelant that most of treatment diagnosed in hospital will lead to a procedure being performed on the patient.

Moderate association can be seen between the age of the patient with number of diagnoses (ndx) and 'nchronic' which states an ICD-9 chronic condition, relatable to a chronic disease or illness.



EDA Analysis II: For maximizing our insights about the given datset and looking at key structures/patterns, we try to see what effect Hospital Acquired Condition has in relation with other variables when separated by its existence i.e,HAC = “No” meaning there’s absence of Hospital Acquired Condition and “Yes” stating it is existent.

There is a certain underlying structure when HCA criteria is existence. We can observe from the above boxplots that Hospital Acquired condition is prevelant in people with higher age. Additionally, number of diagnoses and procedures is higher for existing HAC condition. Similarly, total charge and length of stay in hospital is greater for HAC patients. However, on the contrary we do not find any significance difference in terms of visit and hours of admission when differentiated by HAC.

```
## # A tibble: 2 x 3
##   HACs  `Heart Failure Patients` Percentage
##   <chr>                    <int>      <dbl>
## 1 No                      113069      0.769
## 2 Yes                      33909      0.231
```

```
## # A tibble: 2 x 3
##   HACs  `Heart Attack Patients` Percentage
##   <chr>                   <int>      <dbl>
## 1 No                      39674      0.772
## 2 Yes                     11733      0.228
```

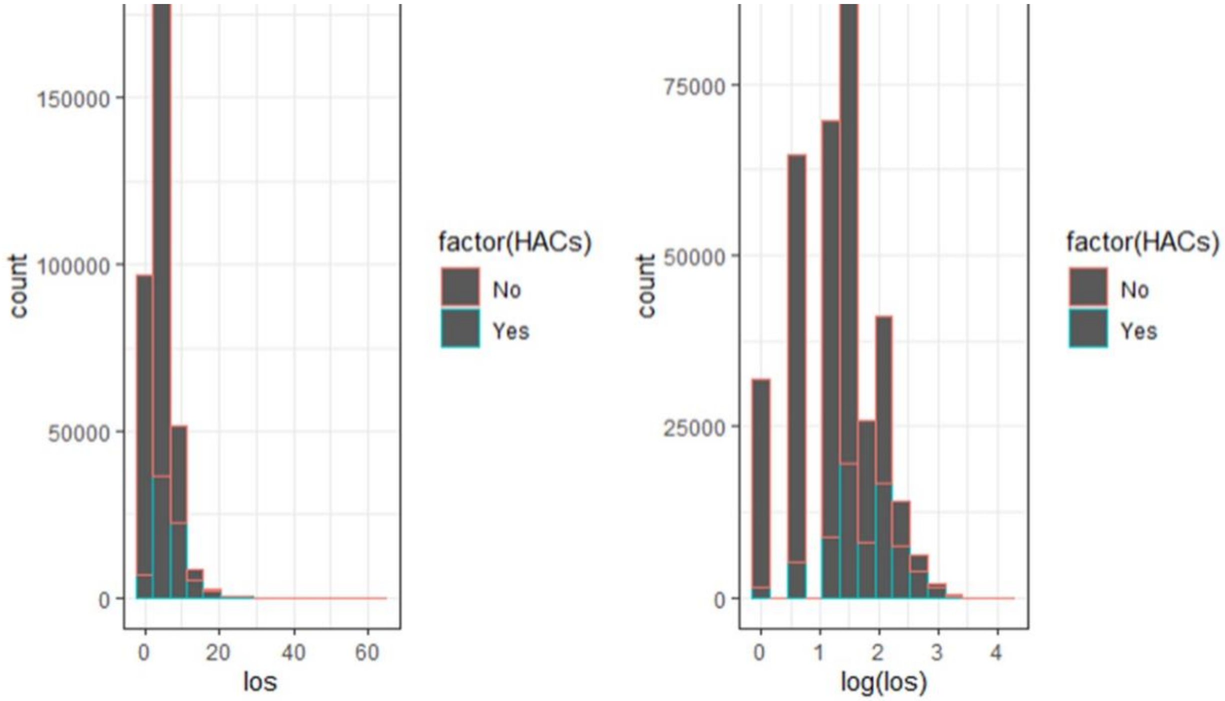
```
## # A tibble: 2 x 3
##   HACs  `Pneumonia Patients` Percentage
##   <chr>                <int>      <dbl>
## 1 No                  120974      0.816
## 2 Yes                  27326      0.184
```

EDA Analysis III: For the given dataset, our focus lies on three different types of patients - Heart Failure Patients, Heart Attack Patients, and Pneumonia patients. In the above tables, we are examining the prevelance of Hospital acquired condition for these three different types. Looking at the percentage values, we can say that HAC exists for roughly about 18-24% of the patients in each of its category.

The highest number of patients affected through HAC are heart failure patients (33909) while the least affected are heart attack patients (11733). However, when we take sample size into consideration for each of the categories, the percentages of patients provide a better estimation stating that affected patients by HAC are almost within touching distance.

```
##     HACs      age nchronic      ndx       npr   totchg      los
## No    No 68.59579 6.771958 11.72450 0.5489283 33518.30 3.854108
## Yes  Yes 74.70417 8.871725 14.89945 1.0094014 51174.66 6.543663
##     count_dxnopoa_ns    visit    ahour totalhospital_noofbed teachstatus
## No          0.000000 3.070862 13.40213              483.9811    1.370664
## Yes         1.953404 3.281123 13.61746              448.0148    1.330446
##     tcontrol    rural     pay1 condition    atype  aweekend     dqtr
## No  3.875704 1.054469 1.609615  2.297022 1.072904 0.2575763 2.438303
## Yes 3.881099 1.051625 1.400970  2.213696 1.080131 0.2470946 2.417772
##     dshospid    female medincstq     year     race  tran_out zipinc_qrtl
## No   1253422 0.4959758  2.280637 2012.471 1.546093 0.3275756    1.945641
## Yes  1326666 0.5174734  2.322868 2012.406 1.456543 0.5956858    1.982307
##     cm_anemdef     cm_chf cm_chrnlung    cm_coag cm_depress     cm_dm
## No   0.2559870 0.06350355   0.3934356 0.04848438 0.09561335 0.2949506
## Yes  0.4013266 0.10610130   0.4698498 0.09262965 0.11979224 0.3078336
##        cm_dmcx  cm_htn_c cm_hypothy  cm_lytes   cm_neuro  cm_obese
## No  0.05798325 0.6710033  0.1557192 0.2542955 0.08196057 0.1571769
## Yes 0.09609692 0.7282919  0.1968260 0.5107170 0.11274805 0.1853963
##     cm_perivasc cm_renlfail     read  count
## No   0.09643172   0.2582850 1.183368 273717
## Yes  0.14062329   0.3904177 1.222056  72968
```

EDA Analysis IV: Interesting Observations: - Existence of HAC is attributed to people with higher age, about 75 years of age. - HAC is prevelant for patients with chronic diseases. - Higher number of ICD-9 discharge and procedures attributing to patients with HAC. - LOS is significantly higher for HAC = "1*Yes* vs HAC = *No*; almost twice. - HAC prevelant cases are attributed to small size of hospitals with less number of beds. - As mentioned earlier, Comorbidity which means simultaneous presence of two chronic diseases or conditions in a patient is prevelant for patients with HAC



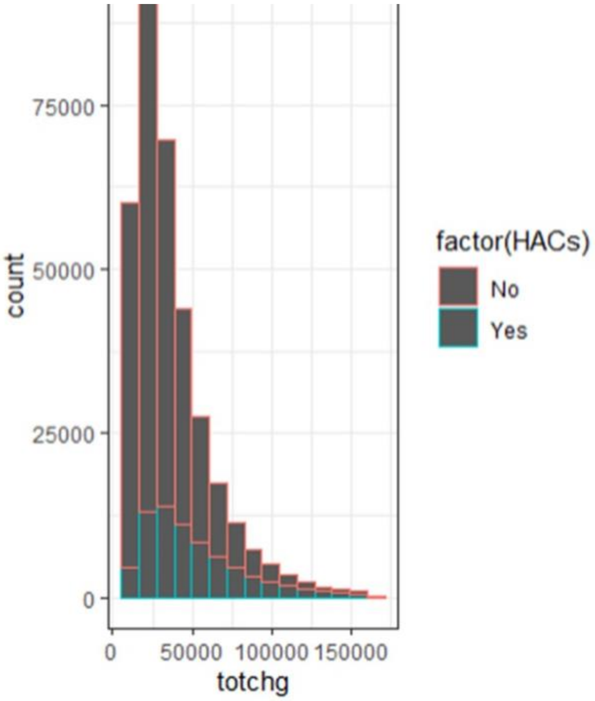
In the first problem, we are examining if Length of Stay has an association with HAC. Looking at the histogram for length of stay, we can say that it is skewed towards the right. A log transformation helps to better the distribution removing excessive skewness. For this study *we will use log(los + 1) and log(totchg +1) as our response variable*.

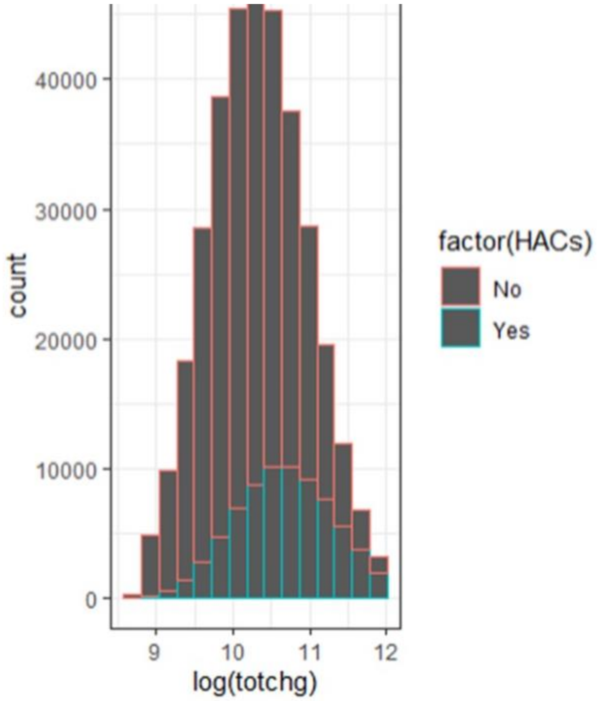
Focusing on the boxplot for length of stay, we have included an interaction term between HAC and year on the Y-axis. We conclusively observe the yeraly pattern of length of stay for all the patients. From the plot, it is evident that cases where patient has had a hospital acquired condition, the length of stay is longer as compared to the patients who haven't.

Interestingly, the described pattern is consistent and existent for all the years in the given dataset.

Moving to the three types of conditions described earlier, we see that length of stay follows a consistent pattern for patients with HAC across all three distinct disorders. It is evidently higher for patients with HAC.

HF - Heart Failure **AMI** - Heart Attack **PN** - Pneumonia





In the second problem, we are examining if Total Charge has an association with HAC. Looking at the histogram for length of stay, we can say that it is skewed towards the right. A log transformation helps to better the distribution as it makes the distribution close to normal.

Focusing on the boxplot for total charge, we have included an interaction term between HAC and year on the Y-axis. We conclusively observe the yeraly pattern of total charge for all the patients. From the plot, it is certain that cases where patient has had a hospital acquired condition, total charge is higher as compared to the patients who haven't. Total charge is considerably higher for patients with HAC.

Moving to the three types of conditions described earlier, we see that total charge does not follow a consistent pattern for patients based on HAC condition for the three categories. It is spread out at both the ends. It does not follow the same pattern as total length of stay.

HF - Heart Failure *AMI* - Heart Attack *PN* - Pneumonia

Naive lm model with transformed reponse

For all the model onwards, we will include the the variables given to us, except for drg. We believe that drg and condition are 2 similar variables and include both of them will be redundant. Condition is also the better variable, which is

We start with fitting a multiple linear regression model using all the predictors for each transformed response. The first model where we regress the length of stay variable ('los') on the whole train data set explains 64.3% of the variation in log(los+1). Most of the predictors are significant as their p-values are lower than 0.05. HACs is also statistically significant, with a coefficient estimation that is equal to 0.1085. In other words, assuming everything other predictor constant, presence of HACs increases the length of stay by 11.39%. Having RMSE on the test set around 0.2985827 also indicates that this model generalizes decently as it is almost the same with the one we get from train set. (0.2989)

Residuals vs. fitted plot of the naive1_log_train model shows that the linear model assumption holds to a reasonable degree. It also reveals a slight violation of constant error variance assumption (heteroscedasticity) due to a pattern in the bottom left. As seen in ACF plot, there seems to be no correlation in successive errors, suggesting that error terms are independent. The q-q plot, however, shows a slight violation of normality assumption of error terms due to left skewness. Yet, since it is a short-tailed distribution, the consequences of non-normality are not serious and can reasonably be ignored.

The model, where we regressed the log-transformed variable for total charge ('totchg') on the whole train data set, explains 76.84% of the variation in log(totchg+1). Most of the predictors are statistically signifant in this model as well. HACs also follows the same trend as its p-value is less than 0.05. Its estimated coefficient 0.07791 suggests that, assuming everything other predictor constant, presence of HACs increases the total charge by 8.10%. RMSEs drawn from both train and test data sets are also identical that indicates the model's efficiency.

Residuals vs. fitted plot of the naive2_log_train model also suggests that the linear model assumption holds to a reasonable degree. However, it reveals a slight violation of the constant error variance assumption due to increasing variance for the higher predicted values. It could be also detected through scale-location plot. ACF plot for this model shows that there is no autocorrelation among the error terms. Like in the first model, the residuals do not follow a normal distribution due to tails on both ends. Yet, having short tails could be ignored due to having such a reasonably large sample size.

Due to presence of multicollinearity in the previous models because of the correlation between the HAC variable and response variables (los and totchg), we also created another set of multiple linear regression models, where we excluded the 'totchg' and 'los' respectively. The first model explains 34.72% of the variation in log(los+1). Most of the predictors are significant as their p-values are lower than 0.05. HACs is also statistically significant, with a coefficient estimation that is equal to 0.1361. In other words, assuming everything other predictor constant, presence of HACs increases the length of stay by

14.6%. Having RMSE on the test set around 0.4048561 also indicate that this model generalizes decently as it is almost the same with the one we get from train set. (0.4037)

Having totchg variable excluded in the model, except the slight violation of constant error variance in the bottom left, all the assumptions holds to a reasonable degree with linearity of model, normality of error terms and non-existence of autocorrelation.

The model, where we regressed the log-transformed variable 'totchg' with the absence of 'los' predictor, explains 56.42% of the variation in log(totchg+1). Most of the predictors are statistically signifant in this model as well. HACs also follows the same trend as its p-value is less than 0.05. Its estimated coefficient 0.1200 suggests that, assuming everything other predictor constant, presence of HACs increases the total charge by 12.75%. RMSEs drawn from both train and test data sets are also similar that indicates the model's efficiency in generalizing.

Same conclusion could be drawn for the model where los variable is excluded. None of the assumptions are violated in a degree to impact parametric inferences from this model.

Based on this information, we can see that the presence of los in model predicting totchg result in a downward bias estimate for the effect of HACs on totchg (and same for the case of model predicting los). Given that HACs is likely to be associated with both los and totchg, having both HACs and totchg as predictors or HACs and los as predictors is likely to result in multicollinearity, and we can say that the estimated coefficients of HACs is biased in these naive model.

Fit a lasso regression with totchg and los as response

Lasso regression (Least Absolute Shrinkage and Selection Operator) provides a better prediction accuracy relative to the models presented earlier in the report as it shrinks and removes certain variables thereby reducing variance without a substantial increase in bias. It also reduces overfitting by eliminating certain variables from the model. Lasso regression was performed individually with our subject of interest (totchg: total charges and los: length of stay) to study the influence of HACs (Hospital-acquired conditions.

When totchg was regressed with all other predictors excluding los, we found that the coefficient from the log lasso model for HACs is 0.1040708 at lambda 1se and HACs coefficient of 0.1070342 min lambda (Assuming all else constant, HACs increase total charge by around 11.3%). At both levels of lambda, predictor HAC remains significant. We can safely conclude that HAC indeed plays a significant role in predicting totchg, and a potential causal relationship between the two might exist. The RMSE of the model is 0.6796682 at 1 sd lambda and 0.6807463 at min lambda, making it less accurate than the original model.

```
## 5 x 2 sparse Matrix of class "dgCMatrix"
##                       1           1
## cm_obese     0.07588515  0.088805516
## cm_perivasc -0.01669544 -0.027382329
## cm_renlfail  .          -0.008381564
## read         0.03657826  0.049894228
## HACs         0.11023243  0.114064015
```

We ran a second model to regress totchg with all predictors including los (length of stay). At lambda lambda1se, the HAC coefficient was 0.04336801 and at min lambda, the coefficient of HAC was 0.06502471 (Assuming all else constant, HACs increase total charge by around 4-6%). At both levels of lambda, the minimum value of lambda as well as 1 standard error from the minimum lambda level, HAC was found to be significant. This reiterates our earlier observation that HAC indeed plays some role in predicting los. The change in the coefficient in these two models can be attributed to the nature of the relationship between the length of stay, total charges and HACs. We can see that the presence of los greatly reduce the effect of HACs on totchg

```
## 5 x 2 sparse Matrix of class "dgCMatrix"
##                       1            1
## cm_obese     0.06271781  0.069559541
## cm_perivasc  .          -0.006914725
## cm_renlfail -0.02186901 -0.031114279
## read         0.01715391  0.024171617
## HACs         0.03747636  0.064151485
```

We perform a lasso regression on the response variable, log(length of stay) with other predictors besides total charges. The coefficient of the HAC from the log model was observed to be 0.1295934 at lambda 1se and 0.1315453 at min lambda (Assuming all else constant, HACs increase length of stay by around 13-14%). The predictor HAC is significant in either level of lambda. We can safely say that HAC plays some role in predicting the response variable, length of stay. The RMSE of the model is 0.5125953 at minimum lambda and 0.513499 at 1 SE lambda.

```
## 5 x 2 sparse Matrix of class "dgCMatrix"
##                       1           1
## cm_obese     0.01390720  0.02741708
## cm_perivasc -0.01556367 -0.02800442
## cm_renlfail  0.02153941  0.03004300
## read         0.01719407  0.02474895
## HACs         0.13629153  0.13873649
```

Lasso regression of the response variable log(length of stay) was performed against all other predictors including the total charge. We observed the coefficient of HAC to be 0.1669708 at lambda 1se and 0.1685096 at minimum lambda (Assuming all else constant, HACs increase total charge by around 18%). At both levels of lambda, HAC was found to be significant. It will be a fair assessment to say that the HAC predictor plays a significant role in predicting total charges. The RMSE at minimum lambda is 0.574021 and at 1 SE lambda is 0.5750868.

```
## 5 x 2 sparse Matrix of class "dgCMatrix"
##                     1            1
## cm_obese    .          -0.007561804
## cm_perivasc .          -0.021667618
## cm_renlfail 0.02943131  0.037687245
## read        .          -0.003173789
## HACs        0.16878624  0.170816169
```

Regression adjustment

By examining the data, we realized that there are only 170 hospital IDs with HACs = 1. As such, we need to drop observations with hospital ID where there is no patients with HACs = 1.

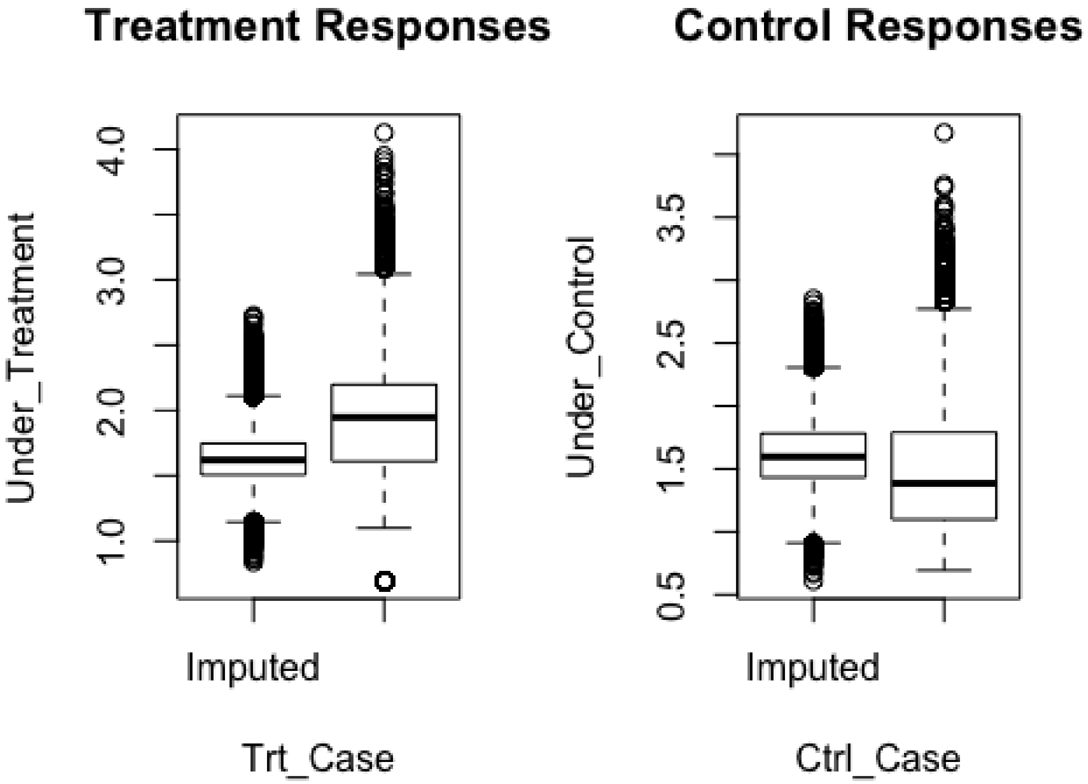
Similar to the previous parts, we run 4 regression adjustment. They are:

- los as the response, excluding totchg as a predictor
- los as the response, including totchg as a predictor
- totchg as the response, excluding los as a predictor
- totchg as the response, including los as a predictor

For the 1st case, we have the ETA = 0.1894895, which means HACs increase length of stay by 20.86% and has a 95% CI bracketted by [0.1880983, 0.1908806] (which indicates statistical significance). This estimation is higher than what we had previously. Based on the box plots, we can see that the imputed values for treatment is much less than the observed, while the imputed value of control case is slightly higher than the observed. These two obervations also support the observation that the causal effect may be slightly underestimated. Overall, the regression adjustment methods seems to yield reasonable results, perhaps conservative because we seem to overestimate control cases and underestimate treatment cases when we impute. So, this might lead to slightly downward biased estimate of the average treatment effect.

```
## treated_cases control_cases
##         72968        273717

##
##  Paired t-test
##
## data:  complete_data$Under_Treatment and complete_data$Under_Control
## t = 266.97, df = 345848, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1880983 0.1908806
## sample estimates:
## mean of the differences
##               0.1894895
```


Treatment Responses

Control Responses

With the addition of totchg in predictors, we can that the effect of HACs on los reduces to 0.1151178 (translated to 12.2% increase in los). This effect is statiscally significant, evidenced by the 95% CI of [0.1140730, 0.1161626]. We can also deduct that there is a downward biased estimate of the average treatment effect and the bias seems to be more prominent in this case, which is true as we have seen ealier that the presence of totchg reduces effect of HACs on los.

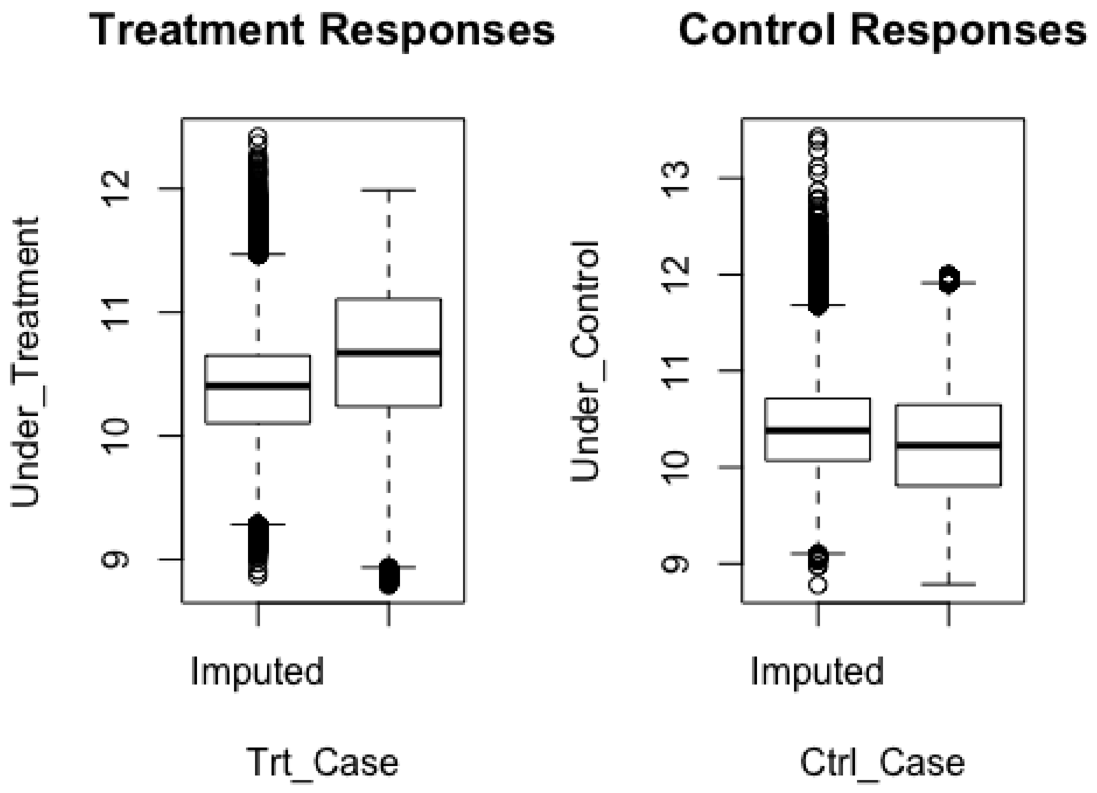
```
## treated_cases control_cases
##         72968        273717
```

```
## 
##  Paired t-test
## 
## data:  complete_data$Under_Treatment and complete_data$Under_Control
## t = 215.95, df = 345848, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1140730 0.1161626
## sample estimates:
## mean of the differences
##               0.1151178
```

The effect of HACs on log(totchg) is 0.174836, which is a 19% increase in total charge. The effect is statiscally significant. The imputed and observes for control cases are pretty close, while observed for treatment is higher than imputed values. Thus, there is some degree of downward bias here, but it’s not as strong like in the case of los

```
## treated_cases control_cases
##         72968        273717

## 
##  Paired t-test
## 
## data:  complete_data$Under_Treatment and complete_data$Under_Control
## t = 235.68, df = 345848, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.173382 0.176290
## sample estimates:
## mean of the differences 
##                0.174836
```



For the last case, we have the effect of HACs on totchg is 0.08718293, which means presence of HACs increases the total charge by 9.11%. The effect is statistically significant, given the 95% CI of [0.08605660,0.08830926]. The problem of downward bias persists, and we can see that presense of los as a predictor lowers the impact of HACs on totchg.

```
## treated_cases control_cases
##         72968        273717

##
##  Paired t-test
##
## data:  complete_data$Under_Treatment and complete_data$Under_Control
## t = 151.71, df = 345848, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.08605660 0.08830926
## sample estimates:
## mean of the differences
##              0.08718293
```

Conclusion

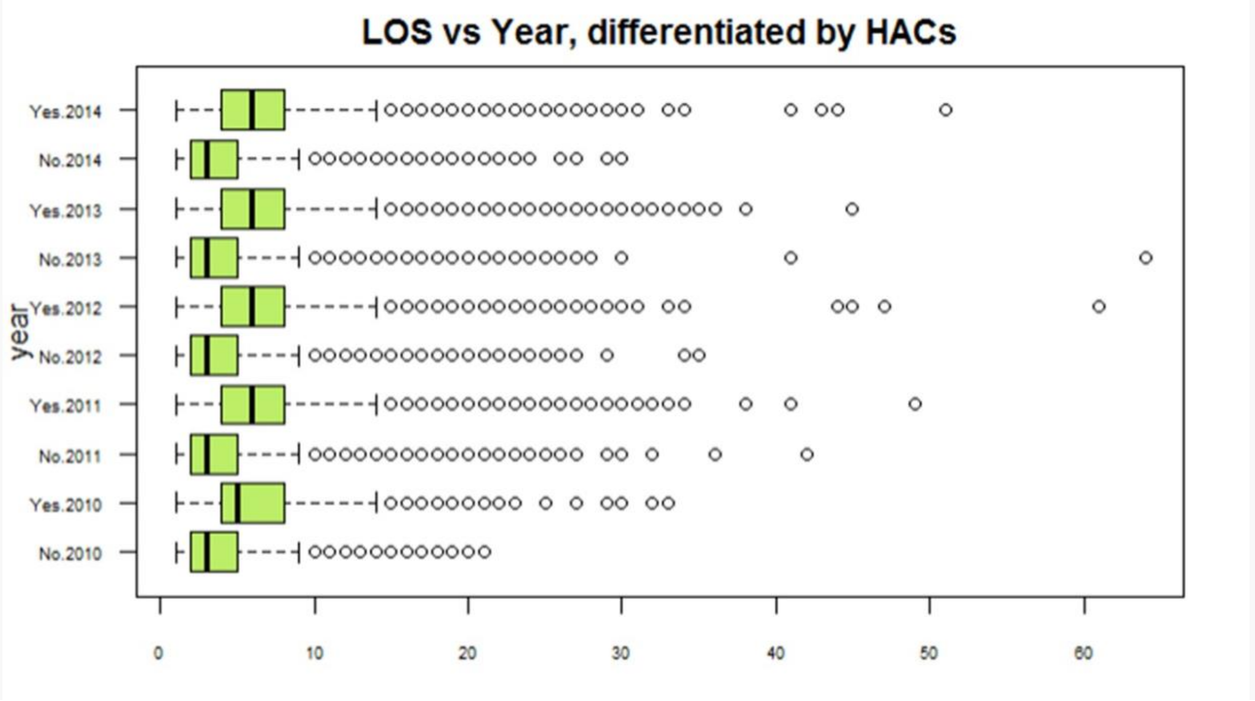
To quickly sum up, the three different types of model that we fitted support our observation that there are indeed associations between HACs and length of stay, as well as HACs and total charge during hospitalization.

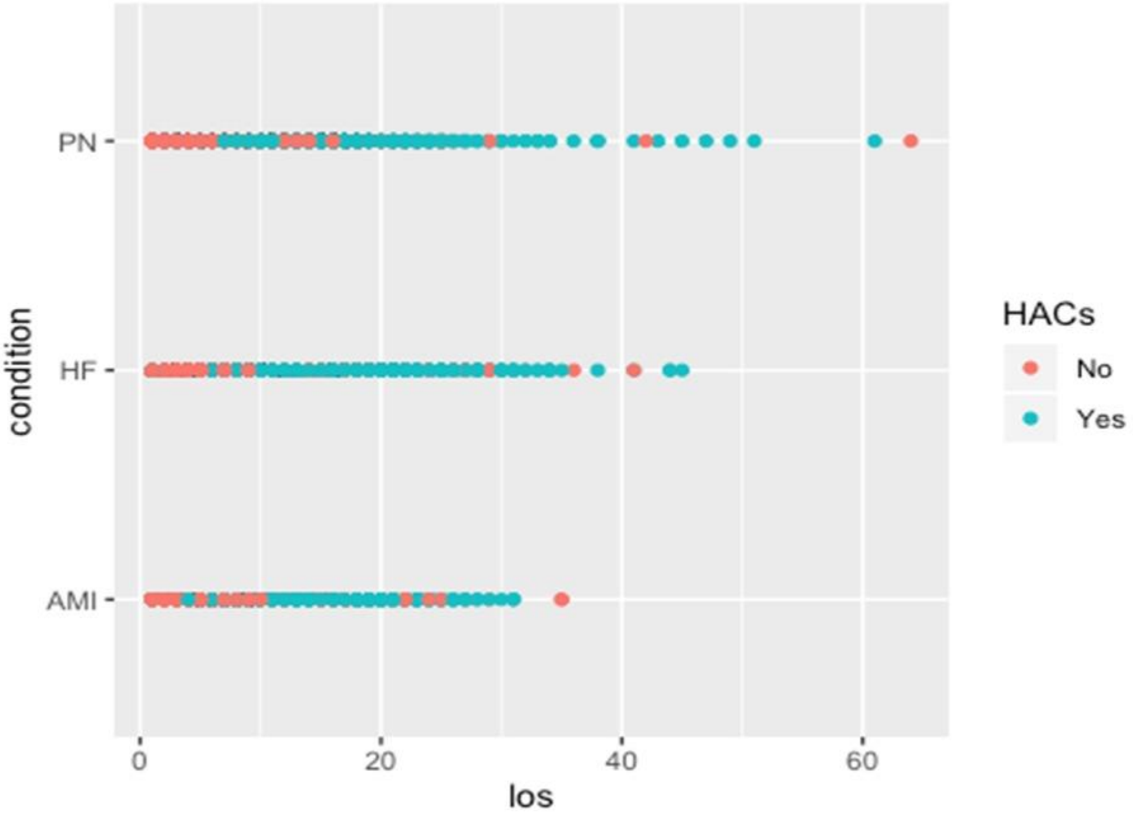
We also show that with or without presence of los(totchg) as predictors in model predicting totchg (los), HACs remains a significant predictor of our response, even though it effect might vary from 4-20% increase.

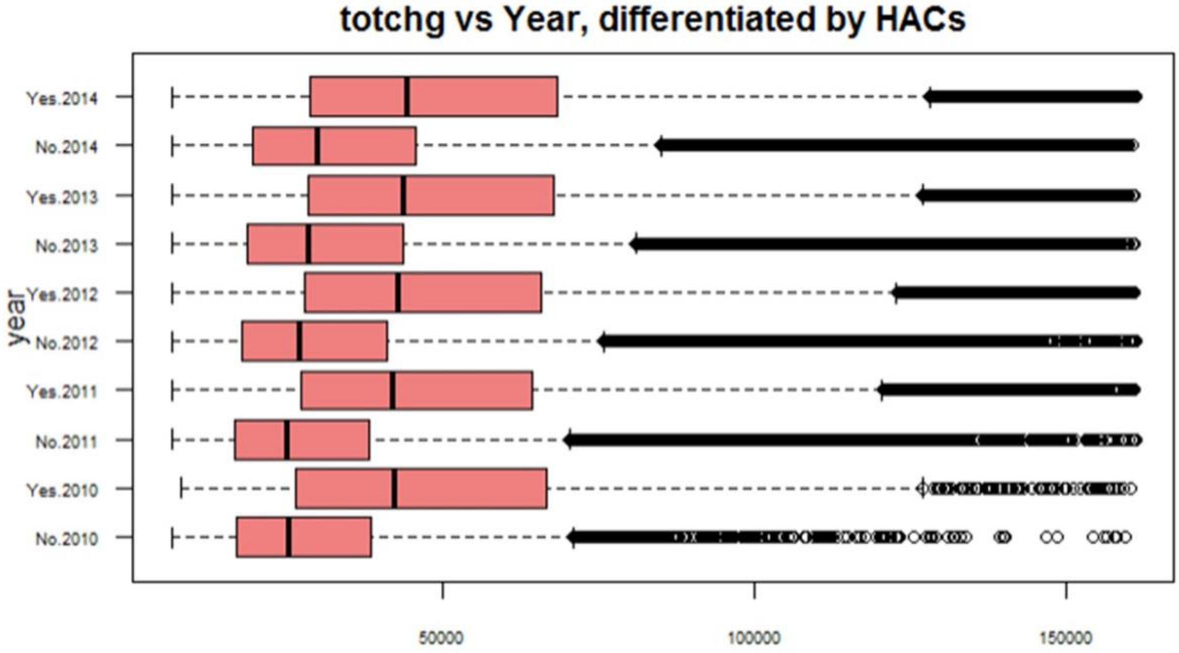
We believe that the relationship between HACs and length of stay or HACs and total charge can be best explained by a causal effect. However, such effects between HACs and los and HACs and totchg will be subjected to multicollinearity when we use both HACs and los to predict totchg or both HACs and totchg to predict los. Additionally, we also believe that it's not justified to excluded los in model predict totchg and vice versa. So we can safely say that the impact of HACs on los should be between 12.2% and 20.86% increase and impact of HACs on totchg should be between 9.11% and 19%. There is definite associations in form of causal effect between these variables.

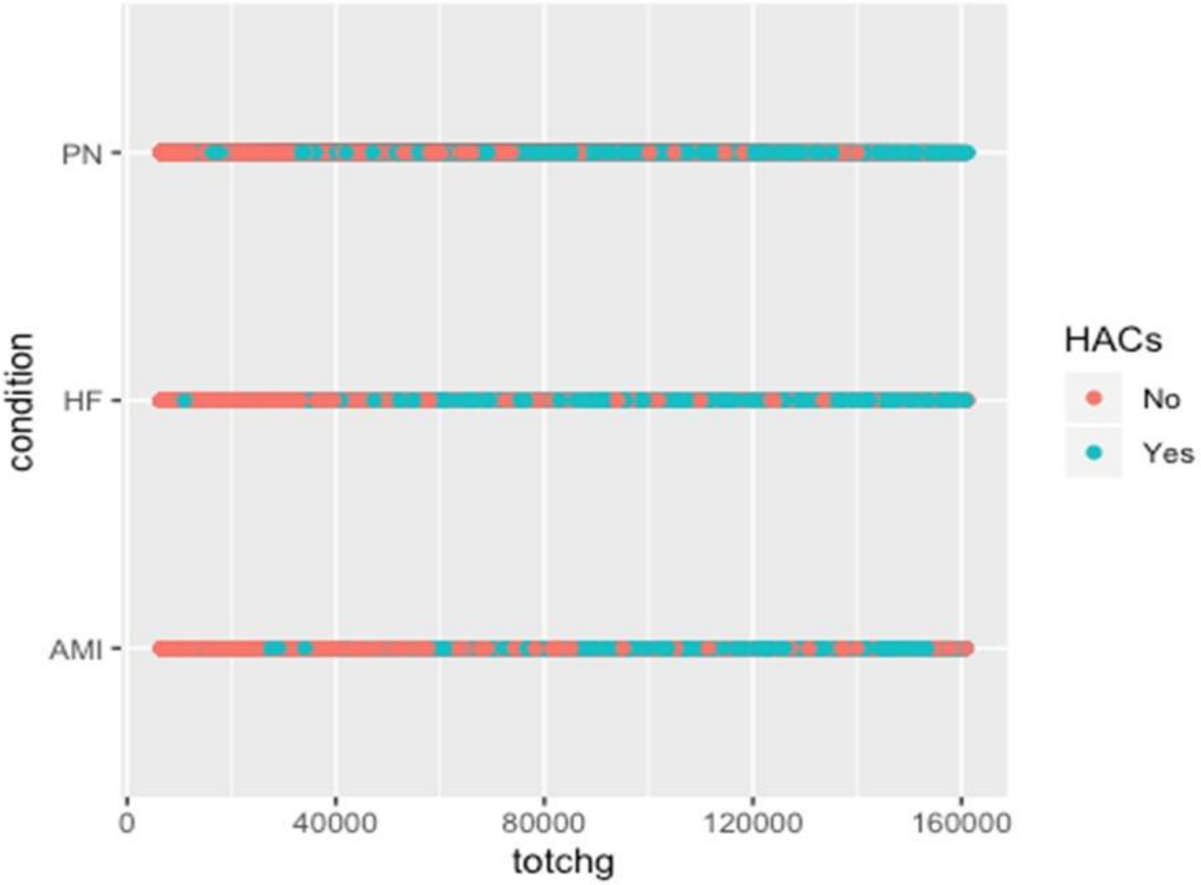
Appendix

Codes for EDA



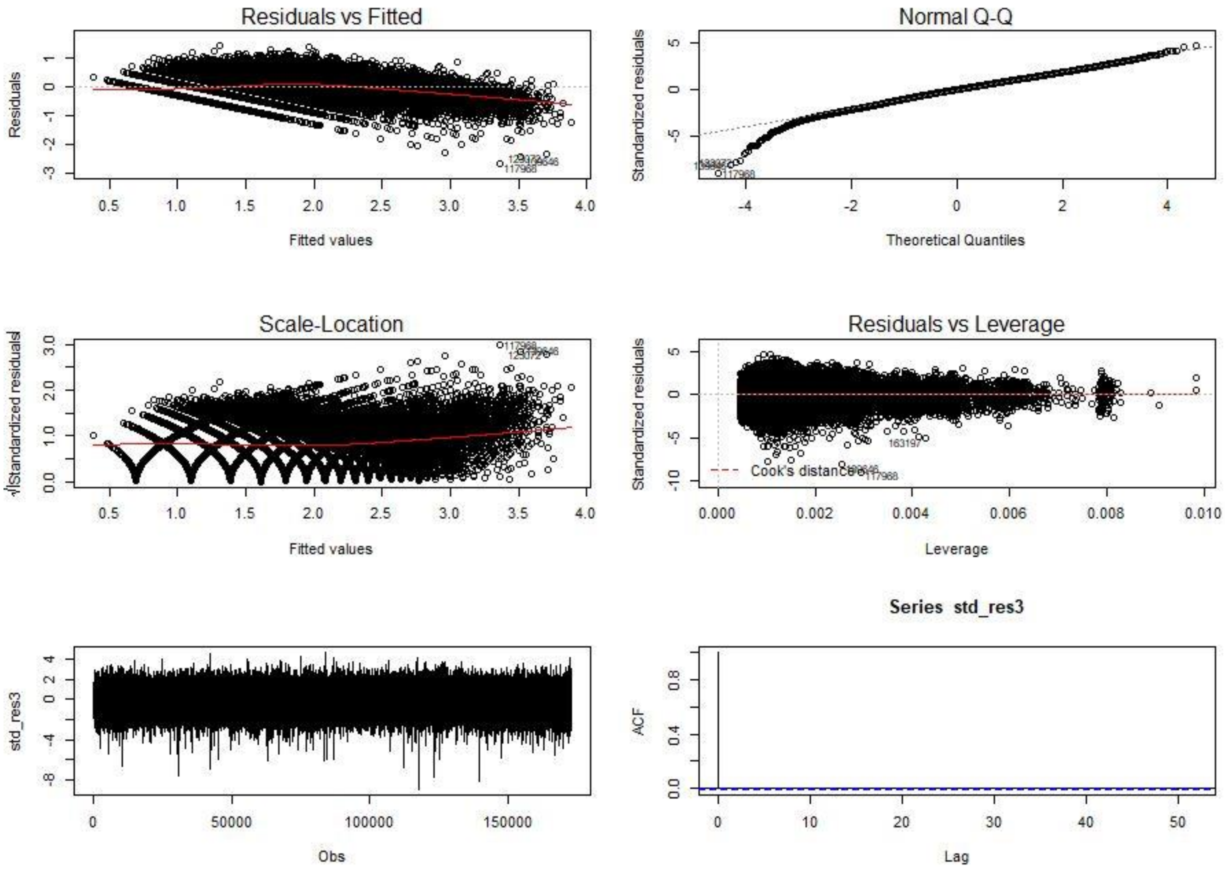



totchg vs Year, differentiated by HACs
year
Yes.2014
No.2014
Yes.2013
No.2013
Yes.2012
No.2012
Yes.2011
No.2011
Yes.2010
No.2010
50000
100000
150000

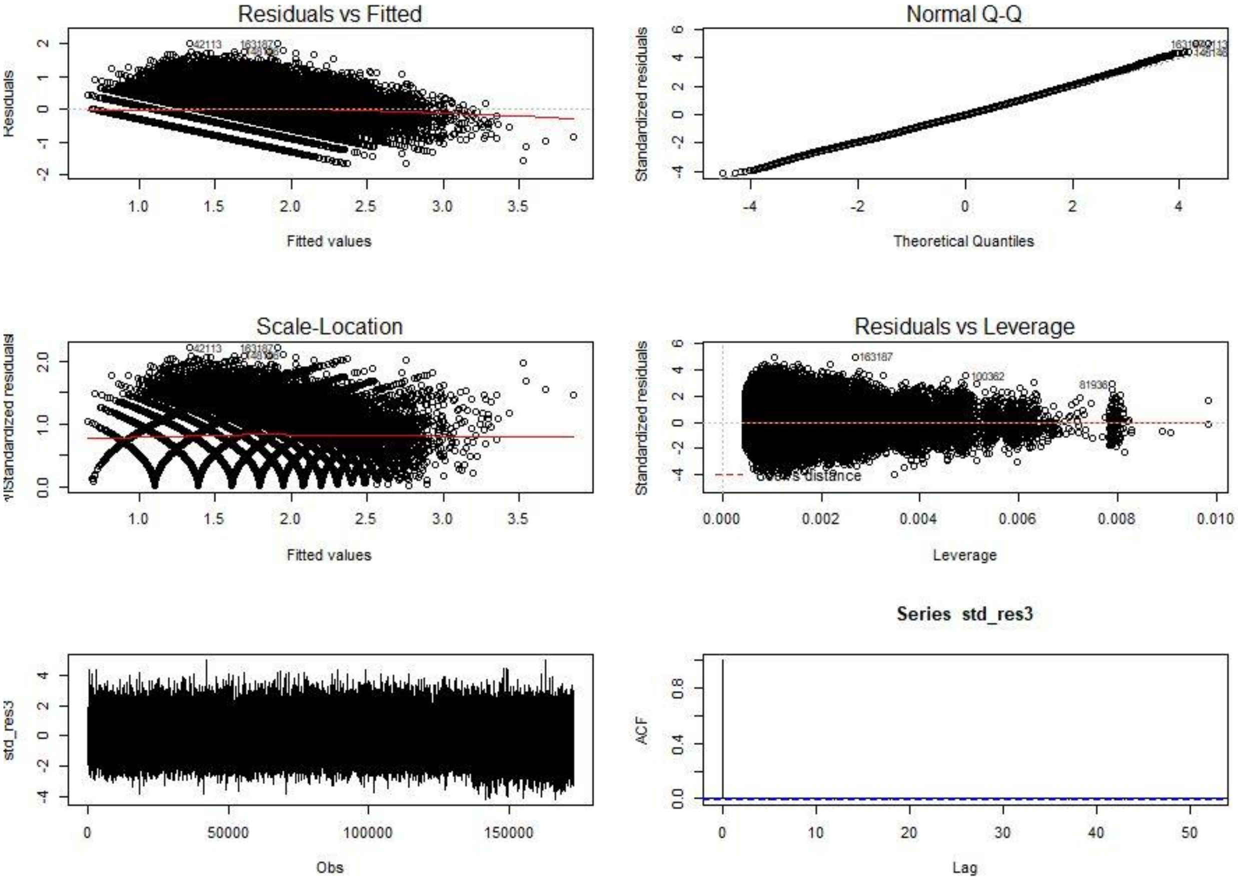

condition
PN
HF
AMI
0
40000
80000
120000
160000
totchg
HACs
No
Yes

Diagnostic plot for regression

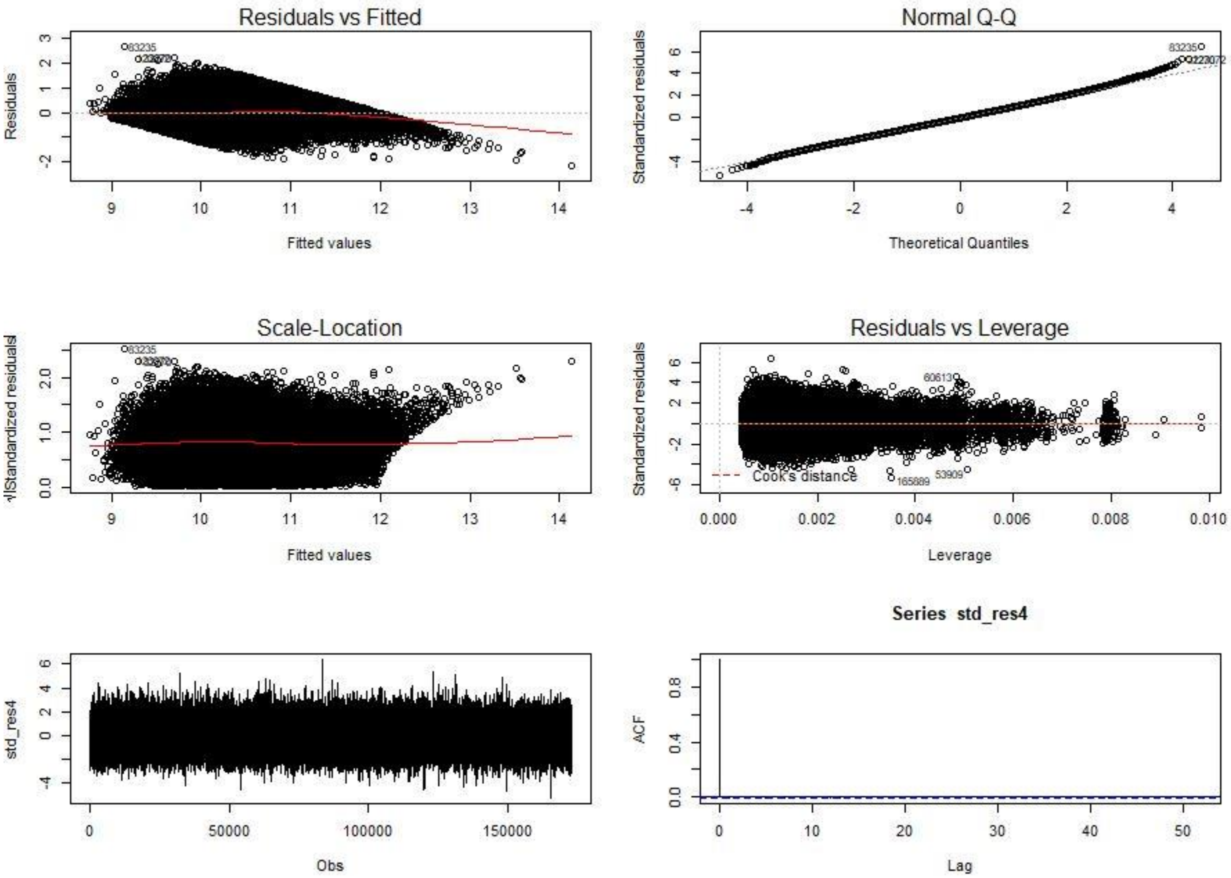
Naive model predict log(los+1) with all variable



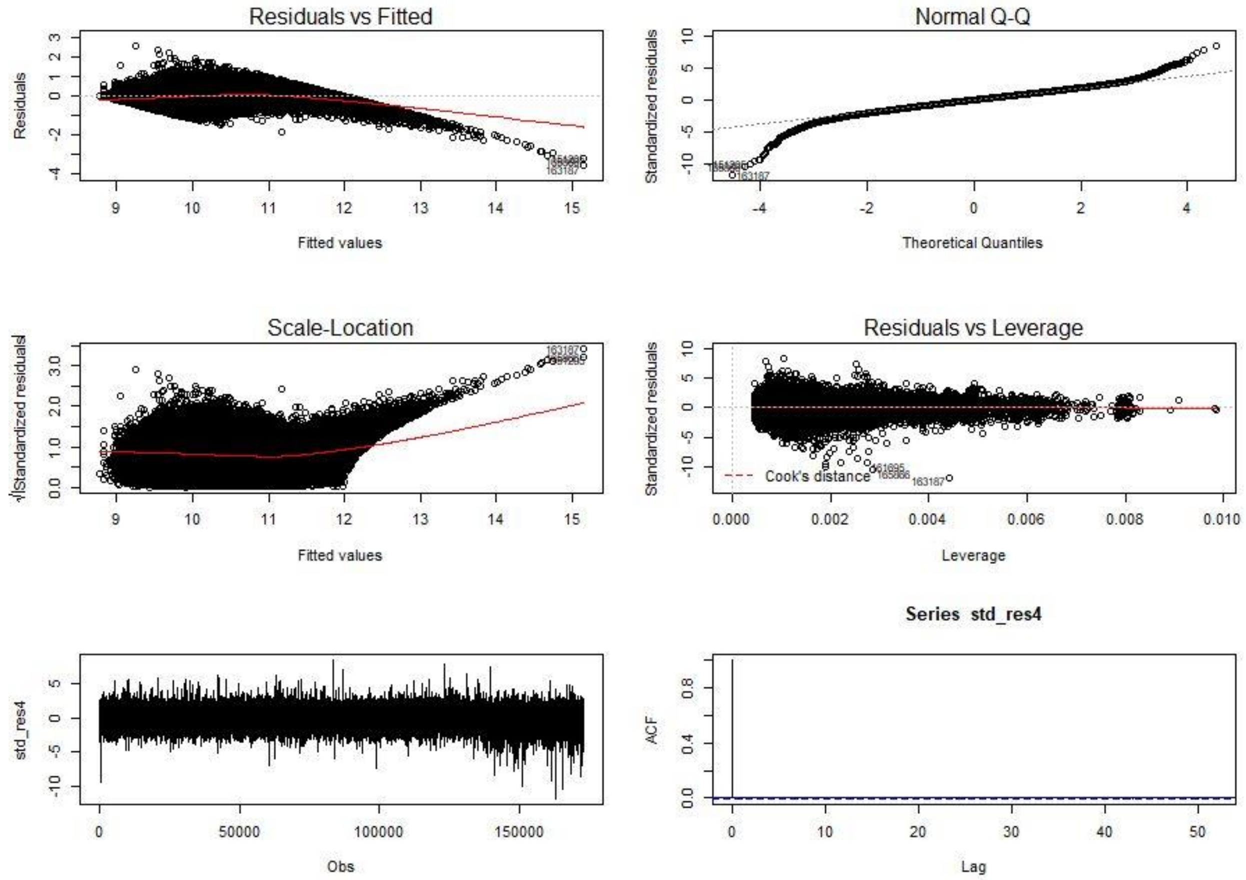
Naive model predict log(totchg+1) with all variable



Naive model predict log(los+1) excluding totchg



Naive model predict log(los+1) exlcuding totchg


Residuals vs Fitted
Residuals
Fitted values
9
10
11
12
13
14
15
3
2
1
0
-2
-4
163187
Normal Q-Q
Standardized residuals
Theoretical Quantiles
-4
-2
0
2
4
10
5
0
-5
-10
163187
Scale-Location
√|Standardized residuals|
Fitted values
0.0
1.0
2.0
3.0
163187
Residuals vs Leverage
Standardized residuals
Leverage
0.000
0.002
0.004
0.006
0.008
0.010
Cook's distance
161695
165886
163187
std_res4
Obs
0
50000
100000
150000
Series std_res4
ACF
Lag
0.0
0.4
0.8
0
10
20
30
40
50