California Office of Statewide Health Planning and Development: Modeling Accounts Receivable to Predict Abnormal Cost Centers

A Group Effort By:

John Benischeck
Tommy Baw
Jordan McIntyre

Table of Contents

Abstract

Through the exploration and linear modeling of the California office of Statewide Health Planning data, we look to find predictors of the dependent variable Accounts Receivable. Accurately predicting future Accounts Receivable and, by association, the Loss Reserves, afford hospitals a greater efficiency in financial planning. The paper will discuss variable selection and calculation, along with observation inclusion and deletion, as well as variable transformations. Simple linear regression modeling, including the Stepwise method, will be employed to model the variables. To judge the robustness of the model, we will analyze the residuals and outliers through various tests, along with the use of the model with a new, test dataset.

Introduction

Hospitals and, more importantly, those who work in them, face daily uncertainties. What are the chances of an incorrect diagnosis? Manifestation of side effects of a treatment? What about the financial uncertainties a hospital faces? The healthcare community is highly fractured in functionality and seemingly prices its respective products and services arbitrarily[1]. Financial planning, in the face of such uncertainty, is difficult. Without good financial planning, a hospital could go bankrupt very quickly. How can pro-forma – the estimated future financial performance of a company - be estimated accurately?

One way to make this pro-forma process easier is to control expenses. Hospitals know statistically how much a consultation or procedure or other service will cost. What they do not know is how many of each will be administered in a given time period. This means that revenue will fluctuate, as well as a cost variable known as accounts receivable. Accounts receivable is simply the outstanding debts owed to the hospital. For example, insurance may cover only some of the cost of a procedure, and a copay is owed, which has not yet been paid. The copay would be considered a part of the hospital's accounts receivable. Like all businesses, there will be a certain percentage of clients (patients) who either cannot or do not pay their bills. To cover these losses, hospitals set aside funds in a category called loss reserves. Loss reserves are usually set as a percentage of estimated accounts receivable for a period.

The aim of this paper is to explore the California Office of Statewide Health Planning and Development's (OSHPD) 2011-2012 data using regression analysis for potential predictors of accounts receivable, in order to more accurately predict the needed loss reserves. By more accurately predicting this cost, the hospital can more efficiently allocate funds to increase bottom dollar (profit).

Data Gathering

For this project, we have used the OSHPD online repository to collect the raw datasets for analysis. We chose to use the 2011-2012 dataset for our training data, and the 2012-2013 data for our test set. Each dataset contained two tabs: one named "Financial and Utilization Data" and a second called "Cost Allocation Data". Our focus was solely on the variables in the former tab. Due to the size of the dataset, we selected 20 variables we thought might be good predictors of accounts receivable:

[1] http://www.uta.edu/faculty/story/2311/Misc/2013,2,26,MedicalCostsDemandAndGreed.pdf

Table 1: Chosen Variables

Variable	Variable Name in Datasets	Type
X1	Type of Control_Church	Binary (1 = yes)
X2	Type of Control_Non-Profit Corporation	Binary (1 = yes)
X3	Type of Control_Non-Profit Other	Binary (1 = yes)
X4	Type of Control_Investor - Individual	Binary (1 = yes)
X5	Type of Control_Investor - Partnership	Binary (1 = yes)
X6	Type of Control_Investor - Corporation	Binary (1 = yes)
X7	Type of Control_State	Binary (1 = yes)
X8	Type of Control_County	Binary (1 = yes)
X9	Type of Control_City/County	Binary (1 = yes)
X10	Type of Control_City	Binary (1 = yes)
X11	Type of Control_District	Binary (1 = yes)
X12	Available Beds (Average)	Continuous
X13	Residents_Total	Continuous
X14	Trauma Center?	Binary (1 = yes)
X15	Income Statement_Gross Patient Revenue	Continuous
X16	Productive Hour Precentage	Continuous
X17	Avg Length of Stay (including LTC)	Continuous
X18	(Occupancy Rate (available beds))	Continuous
X19	operating rooms	Continuous
X20	Loss Reserves	Continuous
Y1	Accounts Receivable	Continuous

Data Cleaning

To clean the dataset, where the value of '0' could not be imputed (namely in the dependent variable column), the observation was deleted. This makes sense intuitively, because hospitals should have a revenue greater than zero, or else it would imply that the hospital accepted zero patients. Further, certain functions of R do not handle blank observations well and would otherwise remove the observation from calculations. Systemically, the Kaiser hospital and Shriner hospital observations has missing Accounts Receivable in both training and test datasets, and were removed from the dataset. Select others (non-systemic) were removed under the same criteria. The end result was a training dataset of 387 observations and a test dataset of 390 observations.

Data Calculation

Some of the variables in our dataset were created through formulas contained in the data guide from the OSHPD website[2], and others were simply collapsed variables. The Trauma Center variable was

[2] http://www.oshpd.ca.gov/HID/Products/Hospitals/AnnFinanData/HAFDDoc2013.pdf

collapsed from 4 binary variables to 1. We made the assumption that, for our model predictions, there would be no difference between the types of trauma centers, and that it would suffice to say simply whether or not a hospital had a trauma center. Productive Hour Percentage was calculated as (PROD_HRS / (PROD_HRS + NON_PRD_HR)), where PROD_HRS and NON_PRD_HR were calculated based on the instruction sheet. Avg Length of Stay (including LTC), Loss Reserves, Accounts Receivable, and Operating Rooms were all calculated based on the instruction sheet.

Data Exploration

With the variables relevant to the study chosen, calculated, and cleaned, histograms of the variables were run to observe their distributions. Below are the variables that appeared to require transformations. Figure 1 shows the dependent variable, Accounts Receivable, which is highly right skewed. A log transformation was performed to make it more normally distributed (Figure 2). From there, the descriptive statistics were run to explore the independent variables. Referring to Table 1, only non-binary variables were taken into account. In order to avoid scaling issues, a closer look was given to x15 and x20. The variable x15 represents Gross Patient Revenue where the mean is much larger than the other variables. In order to have the model place a more equal weight to each independent variable, a log transformation was performed as referenced below. Variable x20 will be discussed in the following section.

```
> summary(hos$x15)
   Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
2.239e+06 9.455e+07 3.610e+08 7.472e+08 1.065e+09 8.989e+09
> summary(log(hos$x15))
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 14.62   18.36   19.70   19.46   20.79   22.92
```

Figure 1: Histogram of Y

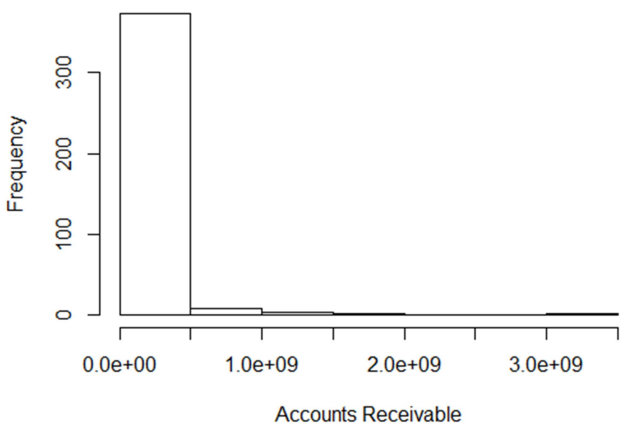


Figure 2: Histogram of Transformed Y

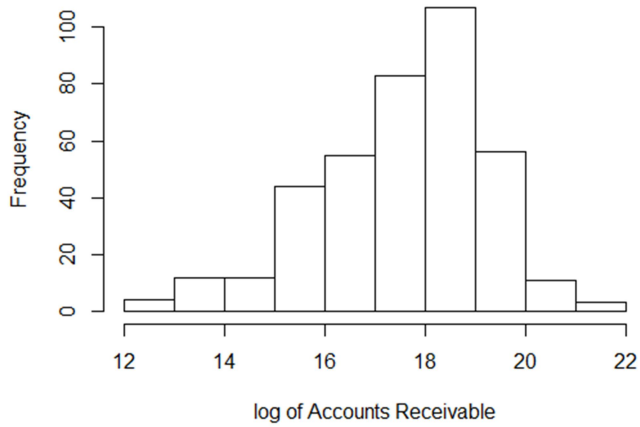


Figure 3: Histogram of X15

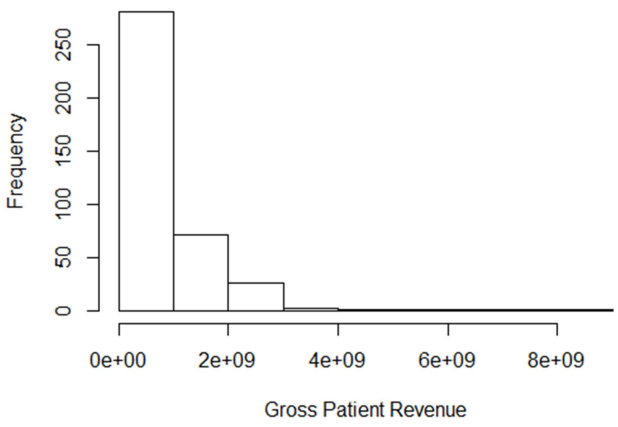


Figure 4: Histogram of Transformed X15

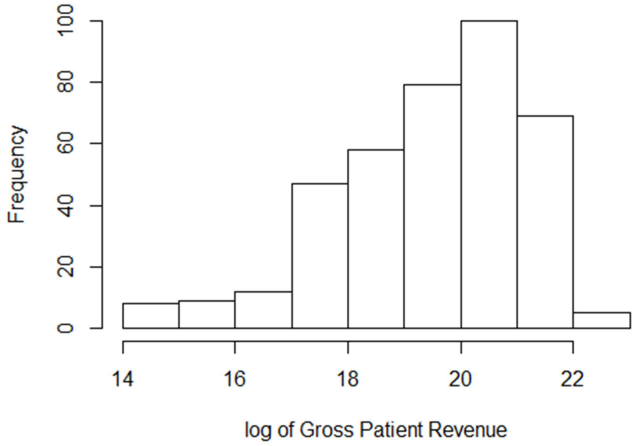


Table 2: Descriptive Statistics

x12	x13	x15	x16	x17
Min. : 10.0	Min. : 0.00	Min. :2.239e+06	Min. :0.5491	Min. : 1.198
1st Qu.: 69.5	1st Qu.: 0.00	1st Qu.:9.455e+07	1st Qu.:0.8459	1st Qu.: 4.047
Median :153.0	Median : 0.00	Median :3.610e+08	Median :0.8696	Median : 5.047
Mean :194.6	Mean : 18.48	Mean :7.472e+08	Mean :0.8660	Mean : 16.244
3rd Qu.:277.0	3rd Qu.: 0.00	3rd Qu.:1.065e+09	3rd Qu.:0.8927	3rd Qu.: 8.307
Max. :892.0	Max. :661.05	Max. :8.989e+09	Max. :1.0000	Max. :775.842

x18	x19	x20	y	yt
Min. : 5.434	Min. : 0.000	Min. :0.000e+00	Min. :2.220e+05	Min. :12.31
1st Qu.:50.216	1st Qu.: 2.000	1st Qu.:6.720e+06	1st Qu.:1.411e+07	1st Qu.:16.46
Median :61.775	Median : 5.000	Median :3.338e+07	Median :5.169e+07	Median :17.76
Mean :60.536	Mean : 6.995	Mean :9.212e+07	Mean :1.225e+08	Mean :17.53
3rd Qu.:73.964	3rd Qu.:10.000	3rd Qu.:9.795e+07	3rd Qu.:1.380e+08	3rd Qu.:18.74
Max. :98.925	Max. :45.000	Max. :3.032e+09	Max. :3.101e+09	Max. :21.85

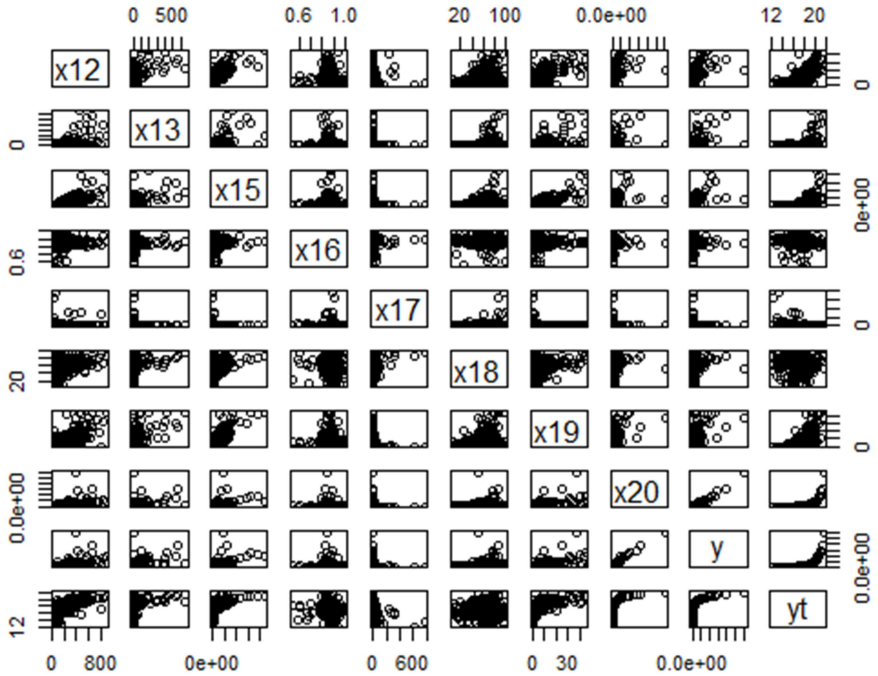
We explored further by taking a look into the correlation and scatterplot matrix. The purpose is to visualize the relationships between variables and to avoid multicollinearity. X20 was discussed above when looking at the scalability of the data. In the correlation matrix, x20 is highly correlated with the dependent variable, with a value of 0.987. X20 is the amount put aside in loss reserves while the dependent variable is accounts receivable so it is logical that there is a high correlation between the two. The decision was made to remove X20 from the model and confirmed when running the linear regression. The other variable of concern was the relationship between gross revenue (x15) and number of operating rooms (x19). This positive relationship exists as number of operating rooms normally correlates to the size of the hospital, and larger hospitals typically will produce more revenue. However,

both were kept in the model as it was observed that not all hospitals had operating rooms and both could be important factors when seeing how they affect accounts receivable. The scatterplot matrix did not indicate any new information that was not already addressed.

Table 3: Correlation Matrix

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18	x19	x20	y	yt
x1	1.000	-0.023	-0.056	-0.015	-0.027	-0.025	-0.015	-0.015	0.066	-0.021	0.018	-0.048	-0.029	-0.005	-0.013	0.009	-0.037	-0.101	-0.012	-0.004	-0.007	0.022
x2	-0.023	1.000	-0.149	-0.039	-0.161	-0.470	0.066	-0.165	-0.079	-0.055	-0.276	0.136	-0.019	0.100	0.172	0.031	-0.065	-0.018	0.253	0.073	0.092	0.193
x3	-0.056	-0.149	1.000	-0.010	-0.046	-0.118	-0.010	-0.042	-0.020	-0.014	-0.026	0.146	0.277	0.094	0.218	-0.009	0.032	0.022	0.173	0.124	0.146	0.078
x4	-0.015	-0.039	-0.010	1.000	-0.012	-0.031	-0.003	-0.011	-0.005	-0.004	-0.018	-0.058	-0.013	-0.023	-0.036	0.011	-0.012	-0.125	-0.033	-0.021	-0.025	-0.082
x5	-0.027	-0.161	-0.046	-0.012	1.000	-0.121	-0.012	-0.051	-0.024	-0.017	-0.086	-0.076	-0.060	-0.079	-0.070	-0.026	-0.026	-0.129	-0.054	-0.034	-0.043	0.008
x6	-0.025	-0.470	-0.118	-0.031	-0.121	1.000	-0.031	-0.131	-0.062	-0.044	-0.219	-0.235	-0.146	-0.202	-0.219	0.125	-0.001	0.073	-0.244	-0.171	-0.194	-0.274
x7	-0.015	0.066	-0.010	-0.003	-0.012	-0.031	1.000	-0.011	-0.005	-0.004	-0.018	-0.007	0.029	-0.023	0.029	-0.064	-0.011	0.029	0.033	0.005	0.007	0.040
x8	-0.015	-0.165	-0.042	-0.011	-0.051	-0.131	-0.011	1.000	-0.022	-0.015	-0.077	0.112	0.200	0.069	0.007	-0.044	0.179	0.084	-0.023	0.226	0.227	0.082
x9	0.066	-0.079	-0.020	-0.005	-0.024	-0.062	-0.005	-0.022	1.000	-0.007	-0.037	0.207	0.258	0.156	0.029	-0.077	0.086	0.090	0.000	0.020	0.039	0.081
x10	-0.021	-0.055	-0.014	-0.004	-0.017	-0.044	-0.004	-0.015	-0.007	1.000	-0.026	-0.044	-0.018	0.062	-0.034	-0.128	-0.016	-0.140	-0.042	-0.016	-0.021	-0.029
x11	0.018	-0.276	-0.026	-0.018	-0.086	-0.219	-0.018	-0.077	-0.037	-0.026	1.000	-0.156	-0.090	-0.035	-0.151	-0.032	0.017	-0.012	-0.162	-0.089	-0.102	-0.172
x12	-0.048	0.136	0.146	-0.058	-0.076	-0.235	-0.007	0.112	0.207	-0.044	-0.156	1.000	0.470	0.415	0.784	-0.141	-0.030	0.195	0.726	0.476	0.556	0.701
x13	-0.029	-0.019	0.277	-0.013	-0.060	-0.146	0.029	0.200	0.258	-0.018	-0.090	0.470	1.000	0.364	0.517	-0.087	-0.044	0.188	0.539	0.394	0.480	0.334
x14	-0.005	0.100	0.094	-0.023	-0.079	-0.202	-0.023	0.069	0.156	0.062	-0.035	0.415	0.364	1.000	0.424	-0.022	-0.082	0.059	0.401	0.247	0.295	0.357
x15	-0.013	0.172	0.218	-0.036	-0.070	-0.219	0.029	0.007	0.029	-0.034	-0.151	0.784	0.517	0.424	1.000	-0.137	-0.127	0.186	0.834	0.514	0.616	0.662
x16	0.009	0.031	-0.009	0.011	-0.026	0.125	-0.064	-0.044	-0.077	-0.128	-0.032	-0.141	-0.087	-0.022	-0.137	1.000	0.020	-0.043	-0.116	-0.095	-0.107	-0.160
x17	-0.037	-0.065	0.032	-0.012	-0.026	-0.001	-0.011	0.179	0.086	-0.016	0.017	-0.030	-0.044	-0.082	-0.127	0.020	1.000	0.221	-0.156	-0.075	-0.085	-0.300
x18	-0.101	-0.018	0.022	-0.125	-0.129	0.073	0.029	0.084	0.090	-0.140	-0.012	0.195	0.188	0.059	0.186	-0.043	0.221	1.000	0.084	0.076	0.105	-0.008
x19	-0.012	0.253	0.173	-0.033	-0.054	-0.244	0.033	-0.023	0.000	-0.042	-0.162	0.726	0.539	0.401	0.834	-0.116	-0.156	0.084	1.000	0.533	0.616	0.658
x20	-0.004	0.073	0.124	-0.021	-0.034	-0.171	0.005	0.226	0.020	-0.016	-0.089	0.476	0.394	0.247	0.514	-0.095	-0.075	0.076	0.533	1.000	0.987	0.525
y	-0.007	0.092	0.146	-0.025	-0.043	-0.194	0.007	0.227	0.039	-0.021	-0.102	0.556	0.480	0.295	0.616	-0.107	-0.085	0.105	0.616	0.987	1.000	0.577
yt	0.022	0.193	0.078	-0.082	0.008	-0.274	0.040	0.082	0.081	-0.029	-0.172	0.701	0.334	0.357	0.662	-0.160	-0.300	-0.008	0.658	0.525	0.577	1.000

Table 4: Scatterplot Matrix



The Linear Model and Model Reduction

Several linear regression models were completed to test the assumptions referenced above and to find the best fitting model. First, the full model was run to see the initial R^2:

$$lm(yt{\sim}x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 + x12 + x13 + x14 + x15 + x16 + x17 + x18 + x19, data = hos)$$

With an R^2 value of 0.6399, there was clearly room for improvement. As stated during our data exploration, a log transformation on Gross Patient Revenue (x15) was performed to normalize the disproportionately weighted independent variables. The following model, now using log(x15), ran with an improved R^2 of 0.8663 where the significant variables at $\alpha = 0.05$ are County Hospitals (x8), Average Available Beds (x12), Gross Patient Revenue (log[x15]), Average Length of Stay (x17), and Occupancy Rate of Beds (x18).

$$lm(yt{\sim}x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 + x12 + x13 + x14 + log(x15) + x16 + x17 + x18 + x19, data = hos)$$

To see how well this model performs, a Step-wise regression model was also performed to see how many variables could be removed from the model while still retaining a relatively high R^2. The model is as follows:

$$Y_t = 1.607 + 0.201x1 + 0.121x2 + 0.245x5 - 0.131x6 + 0.894x8 + 0.001x12 + 0.001x13 + 0.823\log(\mathrm{x15}) - 0.002\mathrm{x17} - 0.005\mathrm{x18}$$

The Step-wise regression gave the same exact significant variables at $\alpha = 0.05$ while removing eight variables from the model. By significantly reducing the complexity of the model and only losing 0.001 variation explanation, The R^2 of 0.865 makes this a reasonable model to use.

Model Diagnostics

Looking at the residual plot in Figure 5 of the next page shows the observations evenly around the 0-axis. The residual and QQ plot does not appear to violate any assumptions, though observation 252 might be a potential outlier. To detect these possible outliers, we next calculated DFFITS, DFBETAS, and Cooks Distance.

Figure 5: Residual Plot

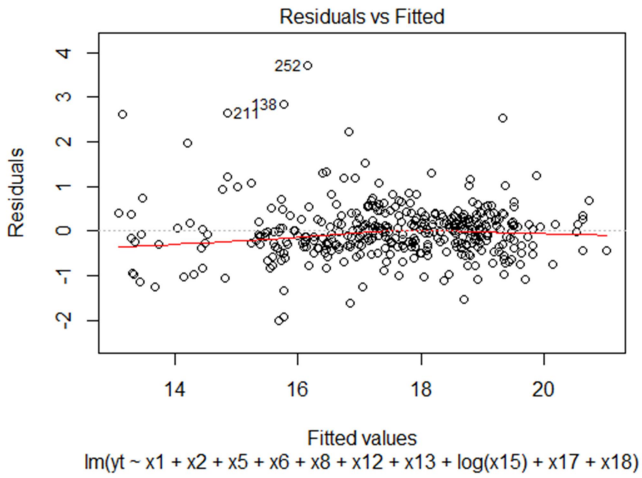
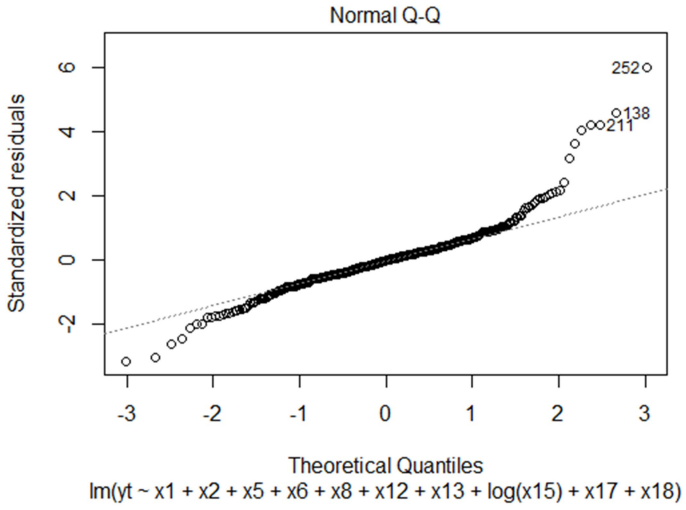


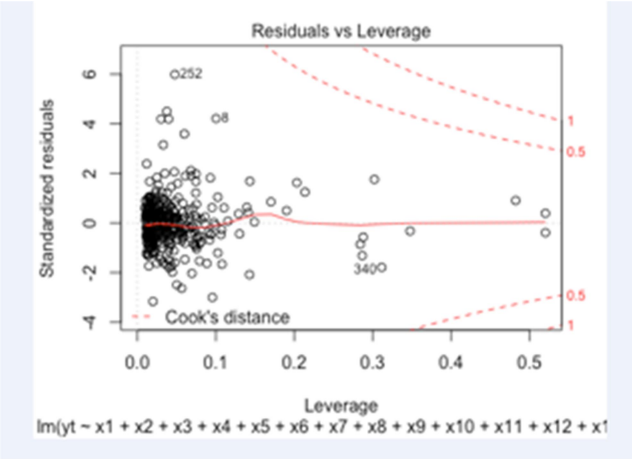
Figure 6: QQ Plot



Several outlier tests were run (DFFITS, DFBETAS, Cooks Distance), and some influential points were identified, but ultimately not dropped from the data. DFFITS outlier test revealed several hospitals that could be seen as outliers under the definition of |DFFITS| > 2. |DFFITS| >0.44315057141 gave us the six potential outliers. However, if using the less strict |DFFITS| >1 method, only one observation, ALTA BATES SUMMIT MEDICAL CENTER, would be seen as an outlier.

Taking a look at Cooks Distance, several observations appear to be influential using the Cooks Distance > 4/n. However, and as observed in the DFFITS test, only one hospital had a value greater than 1- ALTA BATES SUMMIT MEDICAL CENTER (6.590936).

Table 7: Leverage Plot



Since, ALTA BATES SUMMIT MEDICAL CENTER is a very large hospital, we did not want to excluded it from the data set. Additionally, and supporting its inclusion in the model, the residual vs. fitted plot did not identify Alta Bates as a potential outlier.

Robustness Check

To further test the robustness of the trained model, we used the 2012-2013 test set to test the model's predictive power. As a measure of how well the trained model performed with the test set, we used the Root Mean Square Error (RMSE) metric. Running the model on the train set yielded an RMSE of 17.60591, which fell around the middle of the range of the dependent observations. Using the test data, RMSE yielded a value of 13.71048, close to the bottom of the dependent observation range. Generally speaking, the lower RMSE value on our test set means that the model performed well in predicting the test set.

Conclusion

Given the independent variables used, we found a relatively strong and robust model to predict the value of accounts receivable. The model has a relatively accurate accounting of variance with an R^2 of 0.865, and seems to perform well with new data. However, many of the variables which our analysis found to be important are relatively intuitive to interpret. Our final model is as follows:

$$Y = e^{1.607+0.201x1+0.121x2+0.245x5 \quad .131x6 \quad .894x8 \quad .001x12+0.001x13+0.823\log(x15)-0.002x17-0.005x18}$$

Table 5: List of Final Model Variables

Variable	Variable Name in Model	Type
X1	Type of Control_Church	Binary (1 = yes)
X2	Type of Control_Non-Profit Corporation	Binary (1 = yes)
X5	Type of Control_Investor - Partnership	Binary (1 = yes)
X6	Type of Control_Investor - Corporation	Binary (1 = yes)
X8	Type of Control_County	Binary (1 = yes)
X12	Available Beds (Average)	Continuous
X13	Residents_Total	Continuous
X15	Income Statement_Gross Patient Revenue	Continuous
X17	Avg Length of Stay (including LTC)	Continuous
X18	(Occupancy Rate (available beds))	Continuous

In interpreting these results, we see, for instance, that Available Beds and Gross Patient Revenue point to the size of the hospital. Intuitively, the bigger the hospital, the more revenue they are likely to make, meaning that the accounts receivable will be higher.

Similarly, we understand the inverse relation of Average Length of Stay to mean that if more patients stay longer, there are less overall people the hospital can treat and subsequently less people who would not pay their bills. Another interpretation is that those staying longer may have insurance covering their stay, meaning that the hospital has less concern about collecting those related costs. Likewise, a higher Occupancy Rate (available beds) means that the more open beds there are, the less patients the hospital is serving, leading to a lower accounts receivable.

Number of residents can be either interpreted as a sign of the size of the hospital, or that it is a learning hospital. In either case, we may assume that there are extra costs associated with teaching, thus leading to a higher accounts receivable.

Perhaps the most difficult to interpret are how the types of hospitals (church, non-profit, etc) contribute to the amount of accounts receivable. For future studies, it may be worth looking at variables which are related to the size and location of these hospitals, as well as more detailed financial metrics

such as Receivables Turnover Ratio (how long it takes a hospital to collect its debts on average). This might provide hospitals with insight on how to more effectively handle their accounts receivable.

Given the roughly 17,000 variables contained in the datasets, having the help of a healthcare subject matter expert may help us better identify the more uncommon variables to test for correlation and causation of accounts receivable. Also, an SME would be good to consult regarding the variable of revenue. Revenue is typically used as a part of models calculating predicted accounts receivable. That means, however, using the previous year's revenue to calculate current needs. In practice, the question becomes whether or not to re-build the 2011-2012 dataset with the same variables and replacing patient revenue with that of the 2010-2011 (prior year) dataset. For this model, given the high correlation between the current year revenue and accounts receivable, we made the assumption that the current year revenue could stand in for prior year.

References

All data was use from the OSHPD website:

http://www.oshpd.ca.gov/HID/Hospital-Financial.asp

Interpretation and calculation of columns was guided by the OSHPD data manual:

http://www.oshpd.ca.gov/HID/Products/Hospitals/AnnFinanData/HAFDDoc2013.pdf

Interpreting RMSE:

Karen. (n.d.). *Assessing the Fit of Regression Models*. Retrieved from The Analysis Factor: http://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/

Project code for use in R:

```
####################################
##Install Packages
####################################

install.packages("hydroGOF")


####################################
##Call Libraries
####################################

library(readxl)
library(MASS)
library(caret)
library(hydroGOF)


####################################
##Import Train Set
####################################

hos <- read_excel("E:/STAT 628/Final Exam and Project/TrainData.xlsx") ### original dataset
names(hos) <-
c('name','x1','x2','x3','x4','x5','x6','x7','x8','x9','x10','x11','x12','x13','x14','x15','x16','x17','x18','x19','x20','
y')
hos <- hos[,-1]
yt <- log(hos$y)
hos <- cbind(hos,yt) #adding log(y) to data


####################################
##Import Test Set
####################################

hos12 <- read_excel("E:/STAT 628/Final Exam and Project/TestData.xlsx")
names(hos12) <-
c('name','x1','x2','x3','x4','x5','x6','x7','x8','x9','x10','x11','x12','x13','x14','x15','x16','x17','x18','x19', 'x20' ,
'y')
hos12 <- hos12[,-1]
yt2 <- log(hos12$y)
#hos12 <- hos12[ c(1:20)] ##drop Loss Reserves and the non-log column of y
hos12 <- cbind(hos12,yt2) #adding log(y) to data
```

```
####################################
##histogram of depdendant var
####################################

as.matrix(summary(hos))


####################################
##histogram of depdendant var
####################################

hist(hos$y, xlab="Accounts Receivable",main="Histogram of Y")
hist(hos$yt, xlab="log of Accounts Receivable",main="Histogram of Transformed Y")
hist(hos$x15, xlab="Gross Patient Revenue",main="Histogram of X15")
hist(log(hos$x15), xlab="log of Gross Patient Revenue",main="Histogram of Transformed X15")
hist(hos$x17)


################################################
##correlation matrix and scatter matrix
################################################

#correlation matrix
cor(hos$yt,log(hos$x15))
plot(hos$yt, log(hos$x15))
#scatter matrix
pairs(~x12+x13+x15+x16+x17+x18+x19+x20+y+yt, data=hos) #leaving out binary variables


################################################
##Regression Model no amendments
################################################

model <- lm(yt~.-y, data=hos) ## taking y out because yt is the log transformation of y, so it is redundent
to keep
summary(model) #high r^2 because of variable 20


################################################
##Regression Model - taking out x20
################################################

model1 <- lm(yt~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14+x15+x16+x17+x18+x19,
data=hos)
summary(model1) #lower r^2 but gets rid of multicollinearity


################################################
```

```
##Regression Model - taking log of revenue
################################################

model2 <- lm(yt~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14+log(x15)+x16+x17+x18+x19,
data=hos)
summary(model2)


################################################
##QQ Plot and Residual Plots
################################################

plot(model2)


####################################
##Stepwise Regression
####################################

fit <- lm(yt~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12+x13+x14+log(x15)+x16+x17+x18+x19,
data=hos)
step <- stepAIC(fit, direction="both")
step$anova
summary(step)

plot(step)


################################################
##Outlier detection tests
################################################

#DFFITS
dffits(model2, infl=lm.influence(model,),)

#DFBETAS
dfbetas(model2)

#Cooks Distance
cooks.distance(model2, infl=lm.influence(model,),)


####################################
##Cross-validation
####################################

#match the datasets used in the models above
```

```
hos13 <- hos12[-20]
hos13 <- hos13[-20]
hos13$x15 <- log(hos13$x15)

## classify the predictor model and calculate RMSE for the train model and the cross-corr

pred <- predict(step, hos13)
pred <- as.vector(pred)
rmse(yt2, pred) ## 13.71048
rmse(yt, model2$residuals) ## 17.60591
```