Nursing Home Survey Data: what are some characteristics of a successful nursing home?

By Caitlin Salerno

Executive Summary:

With advanced technology and medicine steadily increasing the average life span, people in their old age, or perhaps those suffering from a mental or physical ailment, tend to go to nursing home. Many different variables need to be balanced to keep the a nursing home business in good standing, for both residents and staff. A few factors for a nursing home (other than profit) can be deemed as “success” variables, and using data from a 2004 National Nursing Home Survey, I aim to find the most significant predictors to focus on when building the success of these businesses.

I have created multiple models using sampled data from nursing homes, and the residents and nursing assistants in these nursing homes, finding that some variables such as being accredited, having good staff training, and ensuring the safety of their residents (through preventing falls) can truly strengthen the standing and overall success of a nursing home.

A nursing home taking predictors for success variables into account can increase its reputation and profits, along with perhaps getting funding from outside sources such as the government. Struggling nursing homes can see what the most important factors are to change, and nursing homes with good success can keep a stronger watch on these important variables than others so that they do not have to be analyzing all variables and get lost in too much data. Most importantly, these findings will also make nursing homes who follow them more comfortable and safer for those living in them. With nursing home abuse and neglect being a sad reality of many facilities, perhaps the level of these unfavorable factors can be depleted over time.

Table of Contents

Business Relevance/Problem Significance:

Nursing homes often depend on their ability to keep residents, visitors and nurses happy to be successful, along with avoiding characteristic traits of neglect (something sadly common in nursing homes). These success measurements are often important for a facility to get funding or a good profit, if a private business. Using survey data from nursing homes will allow current businesses to reflect on past information and find factors from the most successful, highest rated and most profitable nursing homes to use for making future decisions. Additionally, analyzing traits of nursing homes that are most successful will also help those looking to find a nursing home for themselves or family/friends, which is difficult to find.

The Data:

I am using raw data from the National Nursing Home Survey of 2004 (which was the last time a nursing home study was done on a national scale with the United States Department of Health). There are three datasets that come from this raw data. One dataset is focusing on nursing home data, another dataset is focusing on residents, and the last dataset is focusing on the nursing assistants. Since the two datasets not directly about the nursing homes can still be grouped by nursing home and tied back to the first dataset, we can use all three pieces of raw data to create a full picture of the success of a nursing home. The survey allowed up to 8 nurses from each facility to be sampled via a computer-assisted telephone interview (at most 4 working there under a year and at most 4 working there for a year or more). The survey also allowed up to 12 residents to be sampled per facility using the same method. The Nursing Home dataset has 1174 rows and 306 columns. The Nursing Assistant dataset has 3017 rows and 473 columns. The Resident data has 13507 rows and 248 columns. Unfortunately, Nursing Assistants were only surveyed from the first 582 facilities, so models using anything from the nursing assistant dataset have to have about half of the samples taken out. For all other y values, we leave out the nursing assistant data to get a larger sample set.

Data Preparation/Cleaning:

To prepare the data, I needed to clean it first. To make the data easier to understand and model, here are the steps I took:

-I either got rid of rows with unknown/missing data, or replaced it with the most frequent (if categorical) or the mean of the responses (if numerical). I found these means using the proc means function.

- Created columns that were a combination of others, such as finding the average retention rate of CNAs, LPNs, and RNs to make the variable AVGRETENTION, which stood for the average percent of staff who have been there longer than a year at the facility.

-Relabeled ordinal factors that originally were descending, and changed them to ascending, so that I could use the ordinal factors as pseudo-numerical in my models.

Once my data was clean, I needed to combine the three datasets I had, but also needed to ensure that I was only choosing variables that I was interested in. My steps for this section of data preparation are as follows:

-Started by putting each dataset into proc SQL to narrow down the variables I was interested in, along with being able to change the names of the variables easily.

-There was another reason to use proc SQL regarding the residents and nursing assistants datasets; we needed to group these datasets by facility and only take the averages of every variable over each facility.

-I created a second dataset that did not include the nursing assistant data so we would have a larger sample size for any y variables that were not from the nursing assistant dataset, due to the nursing dataset only having 582 facilities and when combined with the other tables, would only give information on those facilities and not all facilities (I used a left join).

Variable Choice/Hypothesized Predictor Table:

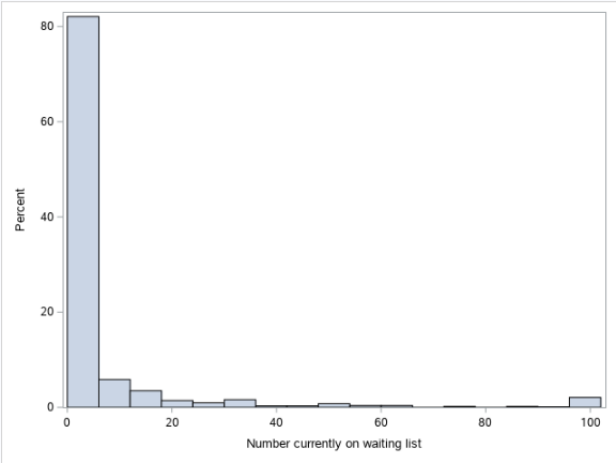
The response variables used in models: The number of people on a waitlist in a facility (NUMWAIT), the average retention rate of the staff (AVGRETENTION), and the average of how likely a nursing assistant would recommend the facility to their friend or family (RECOMMEND).

The predictor variables of interest (hypothesized variables that would affect the response variables, against a null hypothesis that these variables have no relationship to the response variables):

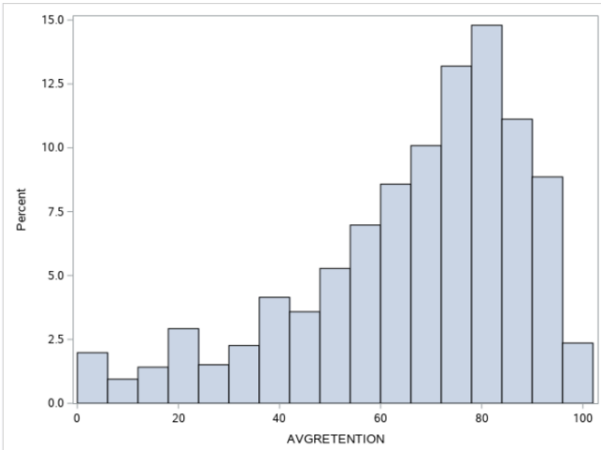
Dataset	**Variables**
Nursing Home Dataset	**NUMWAIT**, **PRIVBASE** (base charge for a resident), **ANYFACAC** (whether the facility was accredited), **VACANCY** (whether any nursing staff vacancies), **AVGWAGE** (average wage of nursing staff), **AVGRETENTION** (average percent of nursing staff there longer than a year), **INJURY** (number of staff on leave or on light duty due to an injury gotten at work), **VOLUNT** (whether the facility has volunteer workers come in)

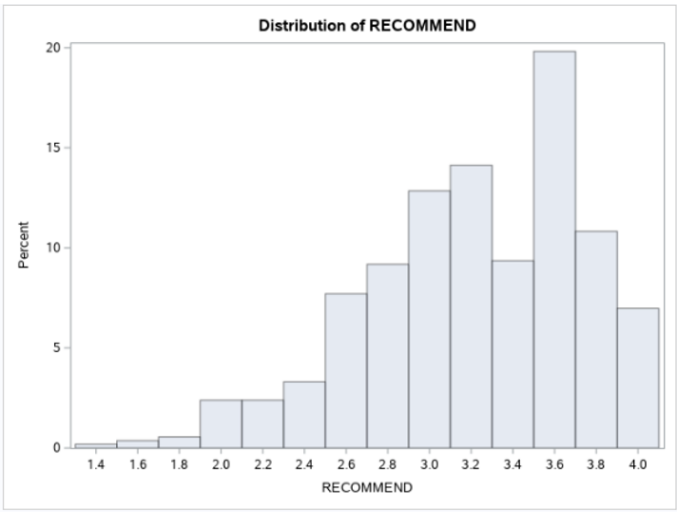
Nursing Home Residents Dataset	**ANYFALLS** (average proportion of residents who have fallen at a facility in the last 6 months, **WGTLOSS** (average proportion of residents who have had major weight loss recently at a facility), **WGTGAIN** (average proportion of residents who have had major weight gain recently at a facility), **CHARGESLASTMONTH** (average charges last month for the residents at a facility)
Nursing Assistants Dataset	**RECOMMEND (average recommendation a nursing assistant would give their friend or family to work at a facility (1-4), TRAINEDWELL (average rating that a nursing assistant gives their training for how well it prepared them for the job (1-3), STAFFAGE (average staff age), STAFFEDUCATE (average staff education level, ordinal level variable: 0-17)**

Descriptive Analysis



The histogram of the number of people currently on the waiting list is not a normal distribution, and is in fact very similar to a poisson distribution. However, we cannot take the log of this variable due to it having so many zeros, so we have to keep the variable as is and take note of the limitations to the models it might have.





The AVGRETENTION and RECOMMEND variables were both relatively normally distributed but were both slightly skewed to the left.

Models and Robustness Checks

Model 1:

Model 1 aimed to explore that factors lead to a rise in people on the waitlist for a facility. Due to the large number of zeros, we will be using a zero-inflated negative binomial regression model. We originally went with a zero-inflated poisson model, but there was a lot of overdispersion (unexpectedly high Pearson Chi-Squared value) so we changed it to the negative binomial regression with zero-inflation.

```
proc genmod data = PartMerge;

class ANYFACAC VOLUNT;

model NUMWAIT = VACANCY PRIVBASE ANYFACAC AVGWAGE AVGRETENTION INJURY
VOLUNT ANYFALLS WGTLOSS WGTGAIN CHARGESLASTMONTH /dist=zinb;

ZEROMODEL VACANCY PRIVBASE ANYFACAC AVGWAGE AVGRETENTION INJURY VOLUNT
ANYFALLS WGTLOSS WGTGAIN CHARGESLASTMONTH;

run;
```

The Pearson Chi-Squared is 1.12, which is much closer to what is expected, so we were correct in changing this to a negative binomial regression model. Based on the statistical significance of certain variables, this was my driving factor to re-running the model with only significant predictors:

```
proc genmod data = PartMerge;
```

```
class ANYFACAC VOLUNT;
model NUMWAIT = VACANCY ANYFACAC AVGRETENTION /dist=zinb;
ZEROMODEL ANYFACAC AVGRETENTION VOLUNT;
run;
```

Though some variables are still not statistically significant at a 0.05 value, the p-values are still relatively small, so we can call both parts of the model relatively strong.

PARAMETER INTERPRETATIONS FOR BINOMIAL REGRESSION MODEL:

The expected change in log(NUMWAIT) for a one unit increase in VACANCY is 0.2711. The expected count in the number of people on the facility's waiting list is 1.31 (e^0.2711=1.31) times higher for each additional vacancy in staff the facility has.

Facilities not accredited had an expected log(NUMWAIT) 0.4688 LOWER (as the number was negative) than facilities not accredited. The expected count in the number of people on the facility's waiting list decreases by about 37% (1-e^-0.4688=0.3742) when the facility is NOT accredited.

The expected change in log(NUMWAIT) for a one unit increase in AVGRETENTION is 0.004. The expected count in the number of people on the facility's waiting list is 1.004 (e^0.004=1.004) times higher for each additional percentage of average staff who stay for a year or more. This does not seem to be like much but a large difference in percentages (90% retention rate vs 0% retention rate) could make a big difference.

PARAMETER INTERPRETATIONS FOR LOGISTIC REGRESSION MODEL FOR ESTIMATING PROBABILITY OF BEING AN EXCESS ZERO:

The log odds of being an excessive zero would increase by 0.2526 if the facility was not accredited (ANYFACAC=0) compared to if it was accredited, meaning that facilities with no accreditations have a higher chance of having 0 people on their wait list (and therefore not having a wait list).

The log odds of being an excessive zero would DECREASE by -0.0093 for each unit increase in AVGRETENTION, meaning that the higher the average retention rate in a facility, the more likely the chance that the facility has a wait list (of any value).

The log odds of being an excessive zero would increase by 0.3411 if the facility does not have volunteer workers come to the facility (VOLUNT=0) compared to if it did have volunteer

workers, meaning that facilities not having volunteer workers have a higher chance of having 0 people on their wait list (and therefore not having a wait list).

CHECKING ASSUMPTIONS:

The observations are independent of one another as there are no repeated measurements and there is no matched data. Each facility/resident is different.

```
proc reg data = PartMerge;
model NUMWAIT = VACANCY ANYFACAC AVGRETENTION VOLUNT /vif;
run;
```

All vif values are very low and much less than 5. No multicollinearity.

```
data PartMerge2;
set PartMerge;
ln_AVGRETENTION = log(AVGRETENTION);

proc genmod data = PartMerge2;
class ANYFACAC VOLUNT;
model NUMWAIT = VACANCY ANYFACAC ln_AVGRETENTION|AVGRETENTION /dist=zinb;
ZEROMODEL ANYFACAC ln_AVGRETENTION|AVGRETENTION VOLUNT;
run;
```

Interaction term of the regular AVGRETENTION and the ln_AVGRETENTION is not significant, so the assumption of linearity is met because a significant p-value would indicate non-linearity.

Our final assumption is that there is a large sample size, and with over 1000 samples in the dataset it is safe to say this assumption is met.

Model 2:

Model 2 is aimed at exploring factors leading to a higher retention rate in the facility. AVGRETENTION is a continuous variable and it is a proportion (but measured as a percent, a number between 0 and 100. We will use a linear regression model for this, but will caution the reader that a beta model may be better if the response variable was transformed from a percentage back to a proportion.

```
/* Using Partial Model */
```

```
proc reg data = PartMerge;

model AVGRETENTION = VACANCY PRIVBASE ANYFACAC AVGWAGE NUMWAIT INJURY
VOLUNT ANYFALLS WGTLOSS WGTGAIN CHARGESLASTMONTH;

run;

/* Best to take out PRIVBASE, WGTLOSS and WGTGAIN as these variables had very high p
values in the model */

proc reg data=PartMerge;

model AVGRETENTION = VACANCY ANYFACAC AVGWAGE NUMWAIT INJURY VOLUNT
ANYFALLS CHARGESLASTMONTH;

run;

/*It is best to take out the AVGWAGE, CHARGESLASTMONTH and VOLUNT variables as they
are not significant. */

proc reg data=PartMerge;

model AVGRETENTION = VACANCY ANYFACAC NUMWAIT INJURY ANYFALLS /vif;

run;
```

Though some predictors are still not exactly statistically significant, all parameters here would be significant with an alpha value of 0.8. The model seems to do well according to the f-value but the r-squared is extremely low, indicating that the model is not the best fit for the data and we might be better off using a beta model.

PARAMETER INTERPRETATIONS FOR MODEL:

The average retention rate for a facility is expected to DECREASE by 2.38927 if there are vacancies for positions in the facility (VACANCY=1).

The average retention rate for a facility is expected to increase by 4.52744 if the facility has been accredited(ANYFACAC=1).

The average retention rate for a facility is expected to increase by 0.06983 for every one unit increase in the number of people on the waiting list for the facility.

The average retention rate for a facility is expected to increase by 0.92369 for every one unit increase in the number of nursing staff who are currently on sick leave or doing light duty due to an injury sustained at the facility.

The average retention rate for a facility is expected to DECREASE by 5.95667 for every one unit increase to the percent of residents that have fallen in this facility in the last 6 months (comparing 0% to 100% of residents).

CHECKING ASSUMPTIONS FOR LINEAR REGRESSION:

```
proc univariate data = PartMerge;
var VACANCY ANYFACAC NUMWAIT INJURY ANYFALLS AVGRETENTION;
histogram VACANCY ANYFACAC NUMWAIT INJURY ANYFALLS AVGRETENTION;
qqplot NUMWAIT INJURY ANYFALLS AVGRETENTION;
run;
```

1) Relationship between outcome variable and predictor variable is linear - the residuals between all variables and the retention rate do not have a definitive pattern, indicating that this assumption is met.

2) All variables present have normality - While VACANCY and ANYFACAC are binary, NUMWAIT and INJURY seem to be non-normal. However, the residuals present normality. We cannot change the NUMWAIT and INJURY variables (they seem to be exponential) by taking the log of them because there are many zeros present. The ANYFALLS variable is normally distributed.

3)No multicollinearity- by using the vif function in the final model, all variables have very small variance inflation (values much less than 5) which means there is no multicollinearity present.

4) No autocorrelation- residual data values are independent of each other, shown by the lack of patterns taking place with the residual plots given with the proc reg.

5) Data is homoscedastic (residuals are equal across the regression line) - the residuals for each variable, including the response variable, seem to reflect homoscedasticity.

ALL ASSUMPTIONS MET.

Model 3:

The third model aims to explore factors leading to a higher average if nurses recommending their friends and family to stay at the nursing home (Using the FullMerge data as this variable is regarding responses form the Nursing Assistant dataset). Note that to include data from the Nursing Assistant dataset, we have to drop nearly half of our rows, as nursing assistants were only surveyed at the first 582 nursing homes surveyed in the study. RECOMNH is a continuous

variable between 1 and 4. We will use a linear regression model for this, but will caution the reader that a beta model may be better.

```
proc reg data = FullMerge;

model RECOMMEND = VACANCY PRIVBASE ANYFACAC AVGWAGE NUMWAIT INJURY
VOLUNT ANYFALLS WGTLOSS WGTGAIN CHARGESLASTMONTH AVGRETENTION
TRAINEDWELL STAFFAGE STAFFEDUCATE;

run;

/* Best to take out CHARGESLASTMONTH, STAFFAGE, INJURY, WGTLOSS and WGTGAIN to
start, as they seem to be

insignificant variables due to their high p-values. */

proc reg data = FullMerge;

model RECOMMEND = VACANCY PRIVBASE ANYFACAC AVGWAGE NUMWAIT VOLUNT
ANYFALLS AVGRETENTION TRAINEDWELL STAFFEDUCATE;

run;

/* The only variables with smaller p values seem to be ANYFACAC, ANYFALLS,
AVGRETENTION and TRAINEDWELLL, so these

are the only variables we will keep in our final model. */

proc reg data = FullMerge;

model RECOMMEND = ANYFACAC ANYFALLS AVGRETENTION TRAINEDWELL /vif;

run;
```

Though some predictors are still not exactly statistically significant, all parameters here would be significant with an alpha value of 0.8. The model seems to do well according to the f-value but the r-squared is extremely low, indicating that the model is not the best fit for the data and we might be better off using a beta model.

PARAMETER INTERPRETATIONS FOR MODEL:

The average recommendation of nursing assistants to give to their friends and family (1 is worst, where the nursing assistant will definitely not recommend it, and 4 is best, where the nursing assistant will definitely recommend it) for a facility is expected to increase by 0.09738 if the facility is accredited compared to if it was not accredited. This is not a very large increase, so although the variable might have a good p-value, it might not be that significant.

The average recommendation of nursing assistants to give to their friends and family for a facility is expected to increase by 0.16397 for every one unit (100%) increase to the percent of residents that have fallen in this facility in the last 6 months (comparing 0% to 100% of residents).

The average recommendation of nursing assistants to give to their friends and family for a facility is expected to increase by 0.00101 for each unit increase in the retention rate of the staff, which is a percentage (not a proportion). For facilities with high retention rates, this can make a large impact.

The average recommendation of nursing assistants to give to their friends and family for a facility is expected to increase by 0.26919 for each unit increase in the average rating of the nurses feeling as if their training prepared them for the job (there are three ordinal levels, the baseline is not feeling as if the training prepared them at all).

CHECKING ASSUMPTIONS FOR LINEAR REGRESSION:

```
proc univariate data = FullMerge;
var ANYFACAC ANYFALLS AVGRETENTION TRAINEDWELL RECOMMEND;
histogram AVGRETENTION ANYFALLS TRAINEDWELL RECOMMEND;
qqplot AVGRETENTION ANYFALLS TRAINEDWELL RECOMMEND;
run;
```

1) Relationship between outcome variable and predictor variable is linear - the residuals between all variables and the retention rate do not have a definitive pattern, indicating that this assumption is met.

2) All variables present have normality - While ANYFACAC is binary, all other values are normally distributed. The residuals present normality too.

3)No multicollinearity- by using the vif function in the final model, all variables have very small variance inflation (values much less than 5) which means there is no multicollinearity present.

4) No autocorrelation- residual data values are independent of each other, shown by the lack of patterns taking place with the residual plots given with the proc reg.

5) Data is homoscedastic (residuals are equal across the regression line) - the residuals for each variable, including the response variable, seem to reflect homoscedasticity.

ALL ASSUMPTIONS MET.

Actionable Insights:

Based on model one, to increase people on a waitlist, facilities should become accredited and form measures to increase the retention rate of staff. Though the model showed that having vacancies for staff positions did increase the wait list, this is probably due to beds not being filled due to understaffing so this is not a sound recommendation.

Based on model two, to increase retention rate, facilities should not have vacancies for staff often (though this could be a correlation due to a lower retention rate meaning that there will be a higher chance of a vacancy occurring). Facilities should also become accredited, increase the number of people on the waiting list, and decrease the number of residents having falls. There is a significant variable regarding the injuries that staff have, but this positive correlation does not seem sound or ethical so we are leaving this out on recommendations. It is likely a result of limitations regarding the model.

Based on Model 3, to increase staff recommending the facility to their family and friends, the facility should become accredited, increase the average retention, and ensure that their methods of training the staff are helpful and realistic to the functions of the nursing home. There is a significant variable regarding the proportion of residents having falls, but seeing as this would increase the recommendation, this does not seem like a sound or ethical recommendation to give. It is likely a result of limitations regarding the model.

Limitations of Analysis:

-Many factors in the models were binary, so we can not easily check to see if there is normality in the variables. Additionally, some variables had an poisson distribution, but we were unable to take the log due to zeros in the variable. This was especially prevalent for the the NUMWAIT variable, the dependent variable for our first model.

-Our third model used proportion data, in which a beta model might perform better.

-We got very low r-squared values for our second and third models even with significant variables, indicating that perhaps linear models are not the best for predicting these response variables.