# Introduction

The dataset I have chosen relates to the incidence and prevalence of chronic like conditions throughout Canada from the year 1995 to 2011. Using classification and regression, we can help identify the most common disease in Canada. Whether the most common disease in Canada is linked to any other condition? Is this based on gender or age? Is it based on any other factors? Predicting which will be the most prevalent chronic condition in the coming years? Using multivariate regression, multivariate logistic regression, correlations, in R, shall help predict & provide a comprehensive understanding of these chronic diseases.

# Literature Review

The following two articles portray the various chronic diseases in Canada and the manner in which they affect Canadians. Both articles show correlations with various variables such as age, gender, chronic disease condition. The second article also shows how often patients go to which kind of doctor (i.e. general physician or specialist).

Article 1

From the article, Chronic Diseases in Canada, it was found that Alzheimer’s and vascular dementia were the two prevalent conditions among the elderly. As per the authors, Koutsavlis and Wolfson, these chronic diseases were mainly in men and those in older age bracket. Females on the other hand, were found to live longer than their male counterparts and had milder dementia. In addition, in this article, it was also found that those who were in institutions were prone more to strokes and Parkinson’s disease.

**Bibliography**

Koutsavlis, T., & Wolfson, C. (2000, February). Chronic Diseases in Canada; 3 Elements of Mobility as Predictors of Survival in Elderly Patients with Dementia: Findings from the Canadian Study of Health and Aging. Retrieved from https://www.researchgate.net/profile/William\_Pickett/publication/12246119\_Estimation\_of\_youth\_smoking\_behaviours\_in\_Canada/links/5494e07b0cf29b94482102c5.pdf

Article 2

In this article, it was found that diabetes, heart diseases (or cardio-vascular diseases) and asthma were the prevalent chronic diseases affecting Canadians. In regards to visiting doctors, it was concluded that Canadians visited their primary doctors two times more than they visited a specialist. It was also noted that those with no or at least one chronic disease had the largest number of visits to their primary doctor as compared to those with three or more than three chronic diseases. It was clearly found that most of the responsibility fell on the primary doctors as they were highly sought as compared to specialists in the field.

**Bibliography**

Muggah, E., Graves, E., Bennett, C., & Manuel, D. G. (2012, December 10). The impact of multiple chronic diseases on ambulatory care use; a population based study in Ontario, Canada. Retrieved from http://bmchealthservres.biomedcentral.com/articles/10.1186/1472-6963-12-452

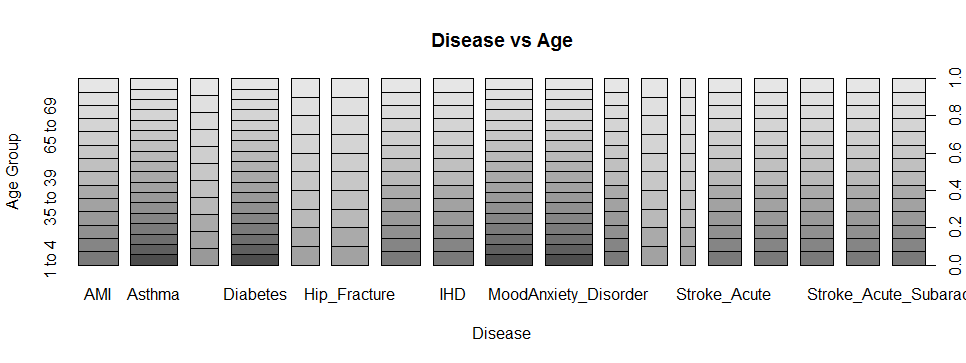
## Dataset

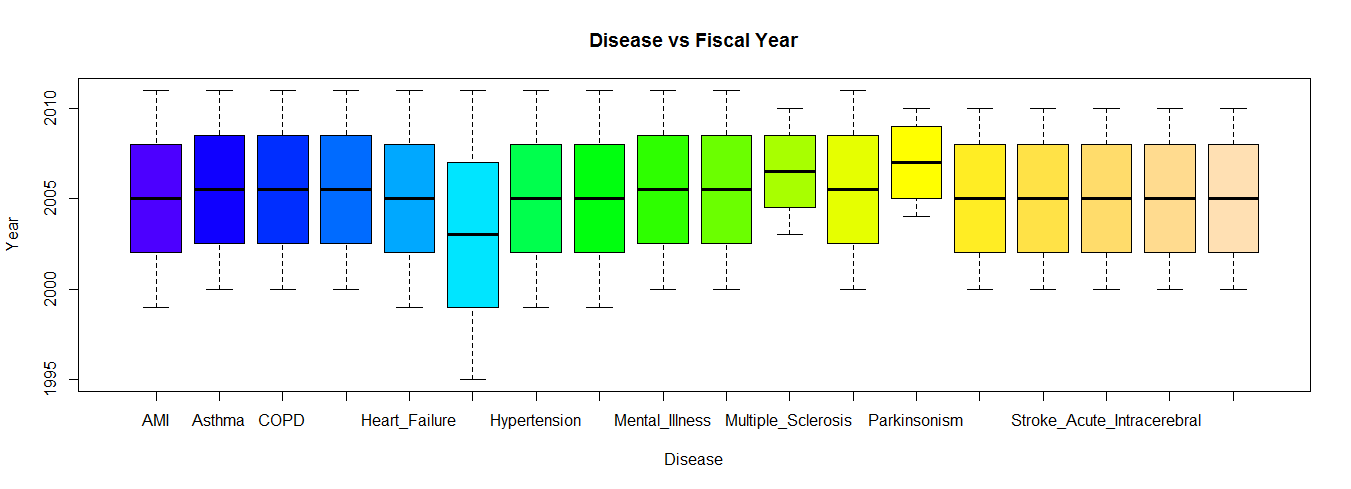
The dataset that is being used is the Canadian Chronic Disease Surveillance System 1996/1997-2011/2012 (source: <http://open.canada.ca/data/en/dataset/9525c8c0-554a-461b-a763-f1657acb9c9d>). In this dataset, there are 27 attributes. Besides the first four attributes, all the other attributes are basically count numbers or the number of instances that particular attribute has occurred. The attributes1 are listed in the below:

|  |  |  |
| --- | --- | --- |
| **#** | **Attribute** | **Description** |
| 1 | Disease | There are 18 unique diseases in this dataset. |
| 2 | Fiscal Year | There are 17 unique fiscal years in this dataset. |
| 3 | Gender | There are two unique values: Males and Females. |
| 4 | Age Group | There are 18 different levels, starting from 1-4 years to 85+. |
| 5 | Population | Number of people |
| 6 | Incident Cases | People with a new diagnosis of the disease case definition in the current year |
| 7 | Prevalent Cases | People with a diagnosis of the disease case definition prior to or during the current year |
| 8 | All cause mortality among people with a diagnosis of the disease case definition | Number of deaths among prevalent disease cases |
| 9 | All cause mortality among people without a diagnosis of the disease case definition | Number of deaths among people without a diagnosis of the disease |
| 10 | Hospitalization with the disease case definition person count | Number of people with the disease who were hospitalized |
| 11 | Hospitalization with the disease case definition (Separations). | Number of hospitalizations among people with the disease |
| 12 | Hospitalization with the disease case definition (Days Stayed). | Number of days stayed in hospital among people with the disease |
| 13 | Hospitalization with the disease case definition person count. | Number of people without the disease who were hospitalized |
| 14 | Hospitals without the disease (Separations). | Number of hospitalizations among people without the disease |
| 15 | Hospitals without the disease case definition (Days Stayed). | Number of days stayed in hospital among people without the disease |
| 16 | Physician Visits with the disease case definition person count | Number of people who visited a physician (General Practitioner and Specialists) with the disease |
| 17 | Physician Visits with the disease case definition (Visit Count) | Number of physician visits (General Practitioner and Specialists) among people with the disease |
| 18 | Physician Visits without the disease case definition person count | Number of people who visited a physician (General Practitioner and Specialists) without the disease |
| 19 | Physician Visits without the disease case definition (Visit Count) | Number of physician visits (General Practitioner and Specialists) among people without the disease |
| 20 | General Physician Visits with the disease case definition person count | Number of people who visited a physician (General Practitioner) with the disease |
| 21 | General Physician Visits with the disease case definition (Visit Count) | Number of physician visits (General Practitioner) among people with the disease |
| 22 | General Physician Visits without the disease case definition person count | Number of people who visited a physician (General Practitioner) without the disease |
| 23 | General Physician Visits without the disease case definition (Visit Count) | Number of physician visits (General Practitioner) among people without the disease |
| 24 | Specialist Physician Visits with the disease case definition person count | Number of people who visited a physician (Specialist) with the disease |
| 25 | Specialist Physician Visits with the disease case definition (Visit Count) | Number of physician visits (Specialist) among people with the disease |
| 26 | Specialist Physician Visits without the disease case definition person count | Number of people who visited a physician (Specialist) without the disease |
| 27 | Specialist Physician Visits without the disease case definition (Visit Count) | Number of physician visits (Specialist) among people without the disease |

1The list of attributes were found as part of the dataset found on (I have formatted it and presented it above)

* Out of the attributes listed above, I do not plan on using attributes # 11, 14, 18.
* After checking the class of this dataset in R, it was determined that this its class is a data frame.
* One of the trends visualized, below, is the relationship between the Disease and Age Groups.



* Another trend visualized below, is the relationship between the Disease and the Fiscal Year.

To give some more insight on the dataset (only the first four attributes within the dataset are shown below, as there are 27 attributes) below are the basic characteristics (using R) of the dataset (df – being the name of the dataset in R). The class, summary, structure, head (the first 6 records) and tail (the last 6 records) of the dataset are shown below:

* class(df):

"data.frame"

* summary(df)

Disease Fiscal.Year Gender Age.Group

Asthma : 432 Min. :1995 F:2914 40 to 44: 422

Diabetes : 432 1st Qu.:2002 M:2914 45 to 49: 422

Mental\_Illness : 432 Median :2005 50 to 54: 422

MoodAnxiety\_Disorder: 432 Mean :2005 55 to 59: 422

AMI : 364 3rd Qu.:2008 60 to 64: 422

Hypertension : 364 Max. :2011 65 to 69: 422

(Other) :3372 (Other) :3296

* str(df):

'data.frame': 5828 obs. of 27 variables:

$ Disease : Factor w/ 18 levels "AMI","Asthma",..: 4 4 4 4 4 4 4 4 4 4 ...

$ Fiscal.Year : int 2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...

$ Gender : Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...

$ Age.Group : Factor w/ 18 levels "1 to 4","10 to 14",..: 1 10 2 3 4 5 6 7 8 9 ...

* head(df):

Disease Fiscal.Year Gender Age.Group

1 Diabetes 2000 F 1 to 4

2 Diabetes 2000 F 5 to 9

3 Diabetes 2000 F 10 to 14

4 Diabetes 2000 F 15 to 19

5 Diabetes 2000 F 20 to 24

6 Diabetes 2000 F 25 to 29

* tail(df)

Disease Fiscal.Year Gender Age.Group

5823 Stroke\_Acute\_Intracerebral 2010 M 60 to 64

5824 Stroke\_Acute\_Intracerebral 2010 M 65 to 69

5825 Stroke\_Acute\_Intracerebral 2010 M 70 to 74

5826 Stroke\_Acute\_Intracerebral 2010 M 75 to 79

5827 Stroke\_Acute\_Intracerebral 2010 M 80 to 84

5828 Stroke\_Acute\_Intracerebral 2010 M 85+

# Approach

## Step 1: Cleaning

Cleaning the dataset is basically done to assist mapping raw data into a convenient format for it to be consumed. This can entail various ways such as finding if there is any extra data which is not required; ensuring that computations can be carried out easily, the attributes have the appropriate classes. Also if there are NAs within the dataset which need to be there or if they can be removed. The main essence of cleaning the data is to organize the data in such a way that it can be fitted into a consumable structure which can then be used for data analysis, aggregation or visualization.

In this dataset, there are quite a few NAs which need to be removed as any analyses cannot be carried out. Furthermore, certain visualizations cannot be presented due to the NAs present. It was also found that certain diseases had many more NAs for certain attributes that other diseases. It was difficult perform any analysis on those diseases and hence they were removed from the dataset. In addition, there were two columns (attributes 8 and 9, as mentioned above) were also removed as these columns had a lot of missing values, which would skew any analysis. Furthermore, these columns did not play a significant part in any manner relating to the research questions being analysed.

Also, for some computations and for visualizing them in certain graphs, the class for these attributes would have to be changed from its default class – factor to numerical for instance (especially for the continuous variables).

## Step 2: Relationships between/among attributes, trends; Analysis; Predictions

Once the data has been cleaned, finding trends and understanding the relationships between attributes is key. For instance, some examples could be understanding the relationship between/among:

* age and disease (2 attributes);
* disease and gender (2 attributes); disease and general physician visits with the disease (Visit Count) (2 attributes);
* disease and gender and fiscal year (3 attributes);

These relationships and trends can be better understood through correlations and correlation matrices. Similar to the instances indicated above, a combination of various attributes would have to be examined to have a better understanding of the dataset, to get more data insight and help predict various factors relating to chronic diseases in the future.

Analysis – Regression

For numerical analyses, regression testing can be performed on the dataset.

Firstly, using multivariate linear regression shall help provide a better picture regarding the relationships among the attributes. With this approach, it shall be much easier to view the various relationships with different variables and how they are shaped, i.e. whether a certain relationship has a positive/negative slope, if it’s completely horizontal/vertical (no relationship per se).

Secondly, for more than 2 variables (or attributes), a multivariate regression approach can be used. With a multivariate regression approach, there will be one or more outcome variables and predictor variables. This shall help provide a better idea on the effect the independent variables shall have on the dependent variable. In this scenario, several attributes (such as Hospitalizations with the Disease that stayed (person count), the number of general physician visits to those with the disease, specialist visits to those who had the disease) can be analyzed with the dependent variable (i.e. Disease, Gender, Age group, fiscal year). This, for instance, will give an idea of the number of visits different doctors visit the patients for each disease. Further, this can also be analyzed based on Gender or Age Group to have better data insights with the other attributes (independent variables).

To predict what the disease would be in the coming years, a training and testing set would have to be chosen. Choosing the fiscal year along with other continuous variable attributes (i.e. basically all the other attributes besides Disease, Age Group, Gender, Fiscal Year) in various combinations shall help predict how diseases could impact Canadians in the future. To further validate these predictions, cross validation techniques could be to help verify and indicate the accuracy of the predictions.

## Step 3: Visualizations

After carrying out various analyses as mentioned in Step 2, visualizing them would help enhance the understanding of the various trends and relationships of the different attributes in the dataset.

Having different kinds of graphs demonstrating the relationships such as bar plot, box plot, line graphs shall help in explaining the analysis done in Step 2.

In addition, graphs such as for the multivariate linear regression almost immediately shows how the variables are related to one another and justify the analysis. Similarly, for logistic regression, having visualizations helps provide a better idea of the different points in the graph, i.e. where each variable will be affected at various points on the graph.

Lastly, having identified what the dataset from 1996-2010 looks like, with the testing and training sets & cross validation, predicting the most common chronic disease with a variety of factors (such as Age, Number of visits by a physician and surgeon) can help put the predicted data in Step 2 in a much more concise manner in a graph. This shall also help in comparing data, i.e. the predicted data vs the data from the dataset (through graphs/visualizations from 1996-2011) to identify any differences in trends and relationships among the attributes.

# Results

It is important to note that in this dataset (all analysis carried out in R), analyzing the attributes must be carried out in certain sets. This is because for each category per se there are associations with a physician, general physician and a specialist.

(i.e. person count of those who have the disease and went to their physician, those who did not have a disease but still went to see their physician and similarly for their general physicians and specialists)

After carrying out a correlation among various sets of attributes, I found that the correlation among the person counts that those who visited the hospital (with the disease), those who saw a physician, general physician and specialist (all person count and had a disease) had the highest correlation as compared to other attribute sets, i.e. a correlation of 0.988.

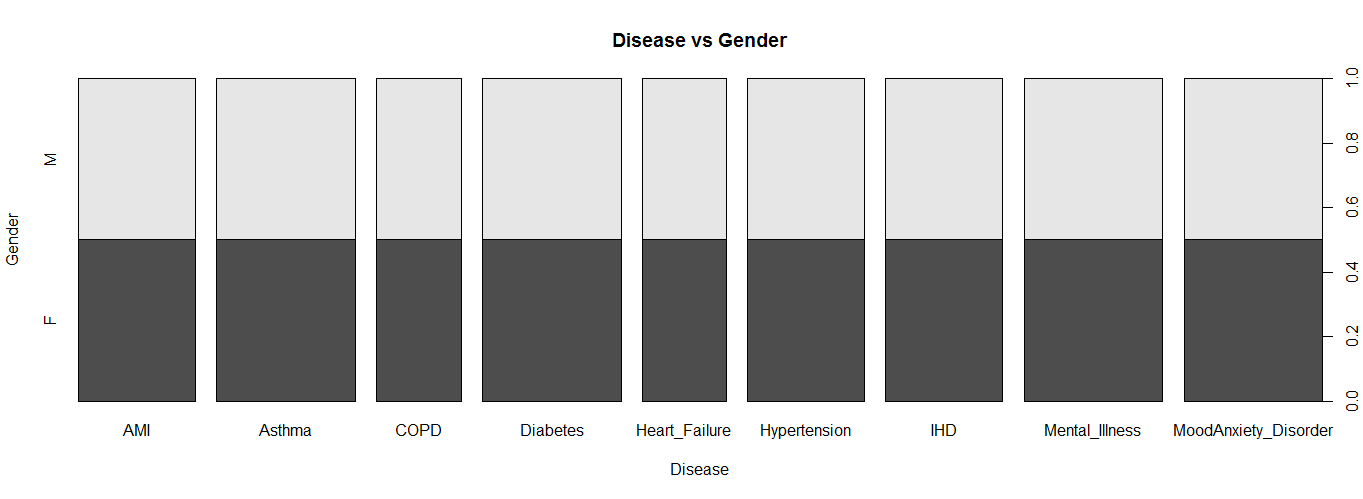
Hence, throughout the analysis and predictions, these three attributes have been used. They are (as per the wording from the dataset):

* Physician Visits with the disease case definition person count
* General Physician Visits with the disease case definition person count
* Specialist Physician Visits with the disease case definition person count

To further enhance the analysis with the three attributes mentioned above, I have also used the following attributes, as it pertains to the person count and relates to hospital visits:

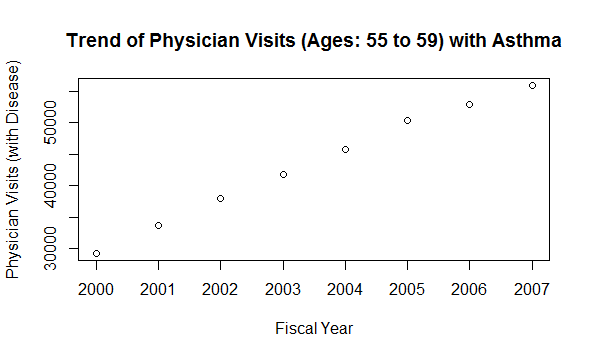
* Hospitalizations with the disease case definition person count

While analysing the dataset, it was discovered that Asthma was one of the diseases that had the highest count of patients throughout the dataset. Hence the analysis focuses on asthma in this results section of the report. Also, in this dataset, the number of males and females for each disease were equal, as can be seen in the diagram below, allowing us to randomly choose the gender to be studied for purposes of this project.



To have a firm understanding of Asthma, I decided to study those in the following age groups: 20 to 24; 55 to 59; 75 to 79. This allowed me to have a holistic perspective - those who were just starting their life to those who would be retiring soon to those who are more elderly. In this report, the gender male was randomly chosen.

For each age group, a separate training and testing sets were created. Hence, based on age group and for each output variable (each of the 4 attributes mentioned above), there were separate multi linear models and predictions. One of the trends is identified between the Physician Visits and Fiscal Year are shown below. In this plot, it can be clearly seen that since 2000, the number of males with Asthma visiting the Physician has been increasing almost at a steady rate.



Similar to the plot mentioned above, all of the trends for various age groups mentioned and attributes have a similar ascend as the years progress. In addition to the Physician or General Physician or Specialist vs Fiscal Year, other models and predictions have also been created for the same output but with different input variables. For instance, in one model – Fiscal Year and Age Group is used while in another model – Fiscal Year and Gender (male and female) are used, while in another model – Fiscal Year, Gender and Age Group are used. These three different models have been used with each of the different types of doctors. Even for these plots, there is a trend which increases linearly and steadily year over year.

Regarding the predictions of the various models, for Physicians, the Age Group of 55 to 59 had the least RMSE as shown in the table/chart below. This trend was consistent for all the various attributes (Hospitalization, Physician, General Physician, Specialist Person Count). For those who went to the Hospital, the 20 to 24 age group has the second least RMSE, making the 75 to 79 age group having the highest RMSE.

In terms of the different types of doctors (person count):

* Physician: the 55 to 59 had the lowest RMSE and the predicted data was very similar to the actual data. For the age group 20 to 24 – the predicted data was not close actual data initially but as the years progressed, the test data was much closer to the predicted values. The 75 to 79 age group predicted data had less wider differences (than the 20 to 24 age group) with its corresponding test data but as the years progressed, the predicted values came closer to the actual data.
* General Physician: the 55 to 59 had the lowest RMSE and the predicted data was very similar to the actual data. For the age group 20 to 24 – the predicted data was not close actual data initially but as the years progressed, the test data was much closer to the predicted values. The 75 to 79 age group predicted data had less wider differences (than the 75 to 79 age group in the Physician and General Physician set) with its corresponding test data but as the years progressed, the predicted values came closer to the actual data.
* Specialist Physician: the 55 to 59 had the lowest RMSE and the predicted data was very similar to the actual data. For the age group 20 to 24 – the predicted data was not close actual data initially but as the years progressed, the test data was much closer to the predicted values, (In comparison with the Physician and General Physician for the 20 to 24 age group), the Specialists had the least RMSE in this age category, as the predicted data and actual data were closest. The 75 to 79 age group predicted data had less wider differences (than the 75 to 79 age group in the Physician and General Physician set) with its corresponding test data but as the years progressed, the predicted values came closer to the actual data.

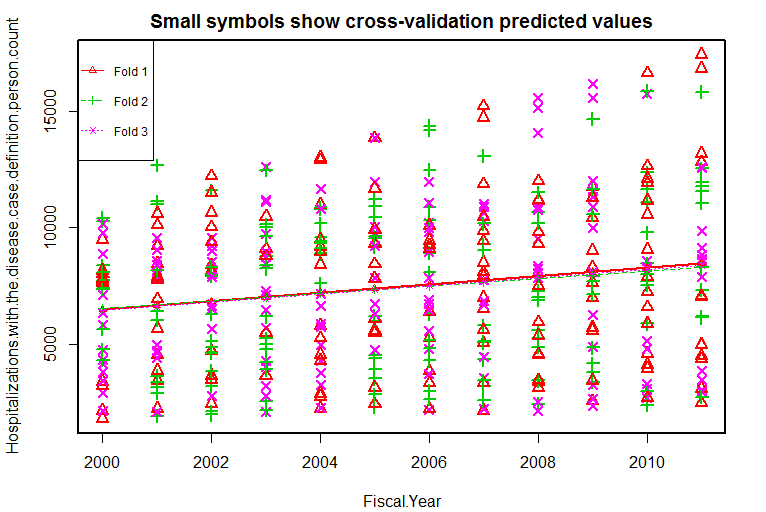
All the data mentioned above points can be viewed in the table/chart below (to see more in depth information such as the formulas used, kindly see the Microsoft Excel file attached to D2L -dropbox for this course):



In addition to predicting the data and comparing them to the actual data (i.e. using training and testing sets), values for person count for the various age groups and different doctors were also predicted for additional upcoming years. (As in the dataset, the information provided is till the year 2010/2011). Hence, using the predict function in R, I have predicted the value for the next three years (2012, 2013, 2014), which can be seen in the table/chart above. From this chart, it is shown that in the year 2012, the highest number of people were visiting the Physician and it is in the 20-24 age group. On the other side of the spectrum, the lowest number was to the Specialist in the 75-79 age category. Also, the hospitalization visits for asthma were least in the 20 to 24 age group category and the age group that went the most were those in the age range of 75-79.

The table further indicates, that this trend is consistent till 2014. The number of those suffering from asthma continue in high number to visit the Physician primarily (as the numbers in the 20 to 24 age group increases every year and is the highest number among all the other attributes in the table).

To further enhance the accuracy of the training and testing models, I used the cross validation function. In R, with a package known as ‘DAAG’, I installed the package which helped me run a cross validation technique known as k-fold cross validation. A graph below shows the use of the k-fold cross validation (the Y axis reads “Hospitalizations with the disease case definition person count):



Using the k-fold cross validation, samples are randomly chosen and divided into various equal k sets (also known as folds). Using all the samples, a model is fit (not including the first subset), where the prediction error (pertaining to the fitted model) is computed using the first set of sample that were not included. This operation is repeated for the number of folds as inputted and hence the performance of the model is calculated by averaging the errors across the different tests sets. This cross validation technique was carried out for various number of models (as seen in the R code) to help present a better idea of the test error for each model. In this scenario, a fold of 3 was used as with more folds, it becomes very difficult to identify any sort of trend due to the data being between only 1999/2000 to 2010/2011. If there were more years, a higher fold would have been more suitable given that there would be much more data. In this particular case, as seen above in the diagram, all the three folds run very close to each other throughout the data. From the graph, it can be seen that as the years progress and the cross validation increases in a very linear fashion, albeit slowly but steadily, throughout all the three folds.

It is also important to note that although most of the plots being analyzed were linear in nature (as indicated by the graphs) in R, there was also some form of non-linearity towards the end of the predictions. As the years progressed, the values were gradually closer and it could be seen as a car that is slowing down after running for a while. To validate this, I created polynomial regressions, which included squares and cubes. Use these polynomial models, I predicted the results and computed the RMSE. Further, I compared these RMSE results with the RMSE results computed earlier (with multivariate linear regression models). Below is table highlighting the RMSE of the polynomial regression and the difference between the RMSE (above) vs the polynomial RMSE differences.

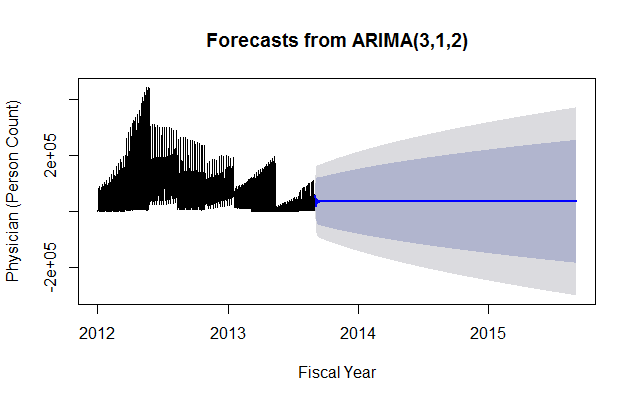
Based on the table/chart above, the analysis on the different doctors (person count):

* Hospitalization: As it can be seen above (the polynomial regression), for the Hospitalization the RMSE as compared to the multivariate linear regression RMSE is much higher for the 55-59 and 75-79 age categories. While analyzing the plot of these two age categories, it was found that midway, the predicted values started to decrease instead of increasing. These predictions were not accurate as the dataset showed constant increasing values as the years progressed.
* Physician: On the other hand, for the physicians although the polynomial RMSE is overall lower than the linear RMSE, the RMSE for 55-59 age category is much higher than the linear RMSE. With the majority having a lower RMSE with the polynomial regression, it seems like the Physicians category are more inclined towards a polynomial regression model.
* General Physician: Similar to the Physician category, the polynomial RMSE was significantly lower than the linear RMSE but the polynomial RMSE for the 50-59 age category is quite high as compared to the linear RMSE.
* Specialist: While both age groups, 20-24 and 75-79 had a much lower polynomial RMSE compared to the linear RMSE, the outlier was the 50-50 age category. The polynomial RMSE was much higher than the linear RMSE.

Overall, from the polynomial and linear RMSEs, it can be seen that within the various Physician categories, the age group from 55 to 59 had more of a linear model (due to a lower RMSE – in the linear regression) while the other two age groups (20 to 24 and 75 to 79) seemed more inclined towards a more polynomial regression model (due to a lower RMSE – in the polynomial regression).

It is important to note that throughout all the various polynomial regression models, taking out the cube factor, did not play any significant role in the predicted results.

Lastly, only for the basis of ensuring that the predicted data is moving towards the trend identified earlier, I used a similar function to predict, that is the forecast function. The forecast function is at time known as the subset of the prediction function as the forecast function focuses mainly on time time-series and the future. Using the forecast function, below is the graph of the people who have Asthma and visit their physician (person count). In the graph below, till mid 2013, it is clearly indicated how the forecast is predicted for Physician visits with Asthma. Beyond mid 2013, this function makes an estimate of the values in the future where the light grey area is the 95% interval and the purplish-blue area is the 80% interval consisting of the predicted values. Hence most of the predicted value shall lie in these highlighted areas and it seems to be increasing (i.e. the count) as the years progress, which is similar to the prediction made in the first table (in the results section above) regarding the future years. Similar to this Physician (person count) - Hospitalization (person count), General Physician (person count) and Specialist (person count), all share a similar trend, i.e. as time go ahead, the count also increases, further validating the predictions made earlier (as mentioned in the table above).



# Conclusion

Hence, it can be concluded that overall, from a holistic perspective, the number of people with asthma will increase in the next coming years (i.e. from 2012 to 2014) as this dataset is limited to information from 1996/1997 till 2010/2011. Asthma seems to quite prevalent in age groups that are more towards the elderly but it seems that the younger generation (ages 20 to 24), although they don’t have the disease (not as high in number as the elderly, ages 75 to 79), the younger ones visit their Physician many times. This is quite insightful as further research can be carried out to understand how the psychology/behaviour of the younger generation vs the elderly to better gauge those visiting the hospitals for certain doctors in future years.

The GitHub link to view these documents pertaining to this Capstone Project is:

<https://github.com/vidurpathak1/ryerson>