"Occupation and chronic illness: Identifying correlations and risk"

James Lapthorne

Domain

The domain of the study covers health and employment.

Question

The aim of this study is to identify correlations between occupation and the risk of developing a chronic illness. Both industry of employment (E.g. mining, food services) and occupation (E.g. labourer, manager) were considered, and the rates of several common chronic illnesses were studied. Additionally, health risk factors such as smoking and excessive alcohol consumption were taken into account as these may influence the results. Data for the Greater Melbourne area was used, but should be relevant elsewhere as geographical area will be abstracted away in the final results.

Datasets

The following datasets were used:

- ABS Census 2011 - B44 Industry of employment by occupation (**Figure 1**): Contains information about industry of employment and occupation for census respondents in each SA2[1]. It is a comprehensive dataset and is sufficiently detailed to enable granular analysis. As the ABS is a highly reputable source, this data can be relied upon to be accurate and correct. URI: `http://stat.abs.gov.au/Index.aspx?DataSetCode=ABS_CENSUS2011_B44`
- University of Adelaide - SA2 Chronic disease modelled estimate (**Figure 2**): Modelled estimates of several common chronic illnesses by SA2, including an indication of error for each estimate. Available from AURIN.
- University of Adelaide - SA2 Health risk factors modelled estimate (**Figure 3**): Modelled estimate of several health risk factors by SA2, including an indication of error for each estimate. Available from AURIN.

These datasets were chosen as they come from reliable sources and have the common SA2 attribute, which simplifies successful integration.

region_id	Ag_For_Fish_OcMngr	Manufact_OcSalesWk	Construc_OcTechTrdW	Construc_	Construc_
201011001	18	27	199	0	40
201011002	34	22	214	0	33
201011003	45	58	554	3	91
201011004	13	57	533	0	59
201011005	27	28	180	0	31

Figure 1: Sample employment data

[1]See Australian Bureau of Statistics: SA2

region_id	asthma_n	hyperten	hyperten	respirtry_	mntl_bh_	hg_chole
206011105	3	1783.277	3	7491.494	1154.916	32.30322
206011106	3	664.1753	3	2790.181	430.1445	32.30322
206011107	3	1276.895	3	4390.58	711.3781	32.48
206011108	3	2359.132	3	8284.811	1182.855	31.96419
206011109	3	925.8168	3	3183.404	515.787	32.48

Figure 2: Sample chronic disease data

region_id	smokers_	obese_f_r	smokers_	obese_m	ovrwght_	alcohol_c
208031187	841.7244	3	468.463	418.1654	757.7722	3.039347
213011339	1761.69	3	1075.815	886.2804	1624.508	2.879908
213011340	2530.394	3	1518.831	1546.166	2712.471	2.719277
213021341	1793.419	3	1081.169	1110.431	1948.128	3.065071
213021342	3250.329	3	1870.808	1630.262	2648.11	3.007004

Figure 3: Sample health risk factor data

Preprocessing

Preprocessing of the datasets was done using the `pandas` library for Python. The `DataFrame` object is designed for working with large matrices of data, so was chosen to internally represent the datasets.[2]

Given three reasonably large datasets, it was crucial to clean the data and discard all but what was necessary to answer the question. As well as extraneous data, there were other problems preventing good integration. Some data was given as a percentage while others were raw numbers, some of the modelled estimates had high error, and some data was missing or zero. The following was done to prepare the datasets for successful integration:

- Employment data was normalised to a percentage. This was done using the sum over an entire row (which is the number of employed people in an SA2). Although this doesn't account for unemployment, it was a sufficient choice with the given datasets.
- For the modelled estimates, data with a high error was removed as it was unreliable and could give false results.[3] This necessitated removing all data for diabetes and chronic obstructive pulmonary disease. It was unfortunate to lose potentially interesting results, but accuracy is foremost.
- Unnecessary data was removed. This included the raw numbers where rate per 100 was preferred, and error information as it was no longer needed.
- Where there was missing data, the entire row (SA2) was removed from the study. This choice was made as having areas with partially present data could skew the results. Additionally, SA2s with zeros recorded were also removed from the dataset. This is because it was unclear whether it was an error in the data, or a true value. Missing and zero values were not inferred using column mean or similar method, as that would artificially lower the diversity of the data. There was sufficient data, and sufficiently little missing data to safely remove it.
- Although there were outliers in the data, it was decided to leave them unchanged. This choice was made since outliers can provide interesting information, and the data sources are reputable. Removing outliers would artificially improve the final results, which could lead to false correlations being identified.

Integration

To reduce the number of columns in the dataset and make it easier to work with, employment data was aggregated by industry and occupation. E.g. instead of "Mining manager", "Mining technician", "Agriculture Sales", etc., the data was grouped into "Mining", "Agriculture", "Managers", "Technicians", etc. This simplifies future analysis, and also allows for discovery of correlations shared by the same occupation across different industries (Managers with mental health issues, for example).

To connect the three datasets, an inner join was performed on the common SA2 attribute. This served a secondary purpose of removing the areas which were not present in all three datasets. Finally, the data was condensed into one contiguous set ready for analysis.

[2]Note the change from Python's dictionary type as mentioned in Phase 2, this became unmanageable in later stages of processing.

[3]Error was defined in the datasets as an integer from 1-3, where 3 is reliable enough for scientific application and 1 is unsuitable for use. Only data with an error value of 3 was used.

Due to the multidimensionality of the data, there was no way to tell whether it actually held any useful information. To determine if it was worth investigating further, a lag plot of the whole dataset was generated. This is a measure of autocorrelation, and shows how random the data is[4]. The lag plot showed quite a low level of randomness (**Figure 4**), which encouraged further enquiry.

Figure 4: Lag plot of entire dataset

The next step was to calculate the Pearson correlation of each pair of variables, and look for significant values. The correlation coefficients were plotted in a matrix to show the results visually, disregarding irrelevant pairs. This was done using the `seaborn` Python library, and showed some interesting positive and negative correlations (**Figure 5, Figure 6**). At this point, several pairs with particularly interesting correlations were chosen for further investigation.

For each chosen pair, a scatter plot was drawn to examine correlation. Alongside each plot, the Pearson correlation coefficient and p-value was provided to indicate significance of each correlation (**Figures 7-10**).

Results

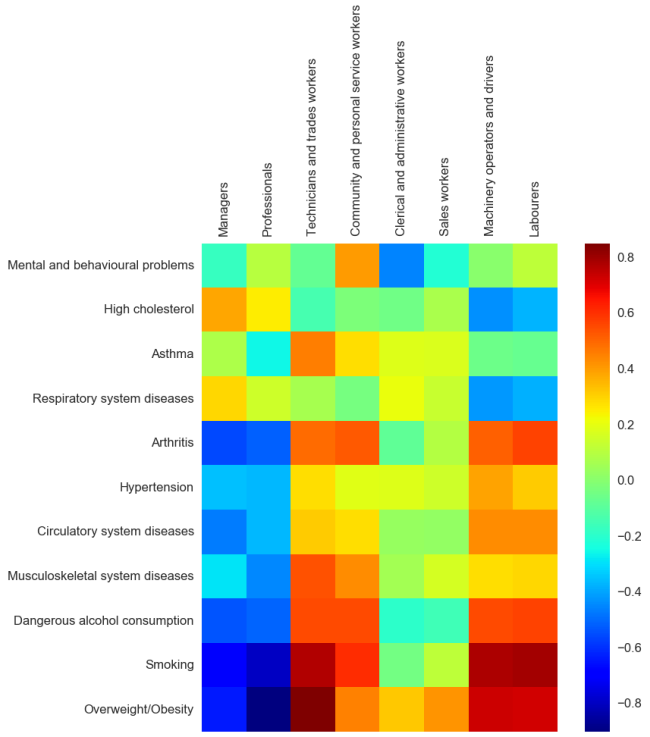

Figure 5: Pearson correlation by occupation

Upon examining the correlation matrices (**Figure 5, Figure 6**) it's immediately obvious that there are several strong correlations in the data. While many pairs appear to have little interaction, both strong negative and positive ends of the spectrum are present. A selection of interesting results:

- Areas with a higher percentage of manufacturing and construction workers tend to have a higher percentage of overweight and obese people ($r \approx 0.7$)
- Areas with a higher percentage of construction workers tend to have a higher percentage of people with asthma ($r = 0.65$) and musculoskeletal diseases[5] ($r = 0.56$)
- Areas with a higher percentage of accommodation and food services workers tend to have a higher percentage of people with mental and behavioural problems ($r = 0.51$)
- Areas with a higher percentage of professional workers tend to have a lower percentage of overweight and obese people ($r = -0.90$) and less smokers ($r = -0.80$)

[4]For a random dataset, points on the lag plot would be randomly distributed.
[5]Musculoskeletal diseases consist of strain injuries such as carpal tunnel syndrome and tendonitis.

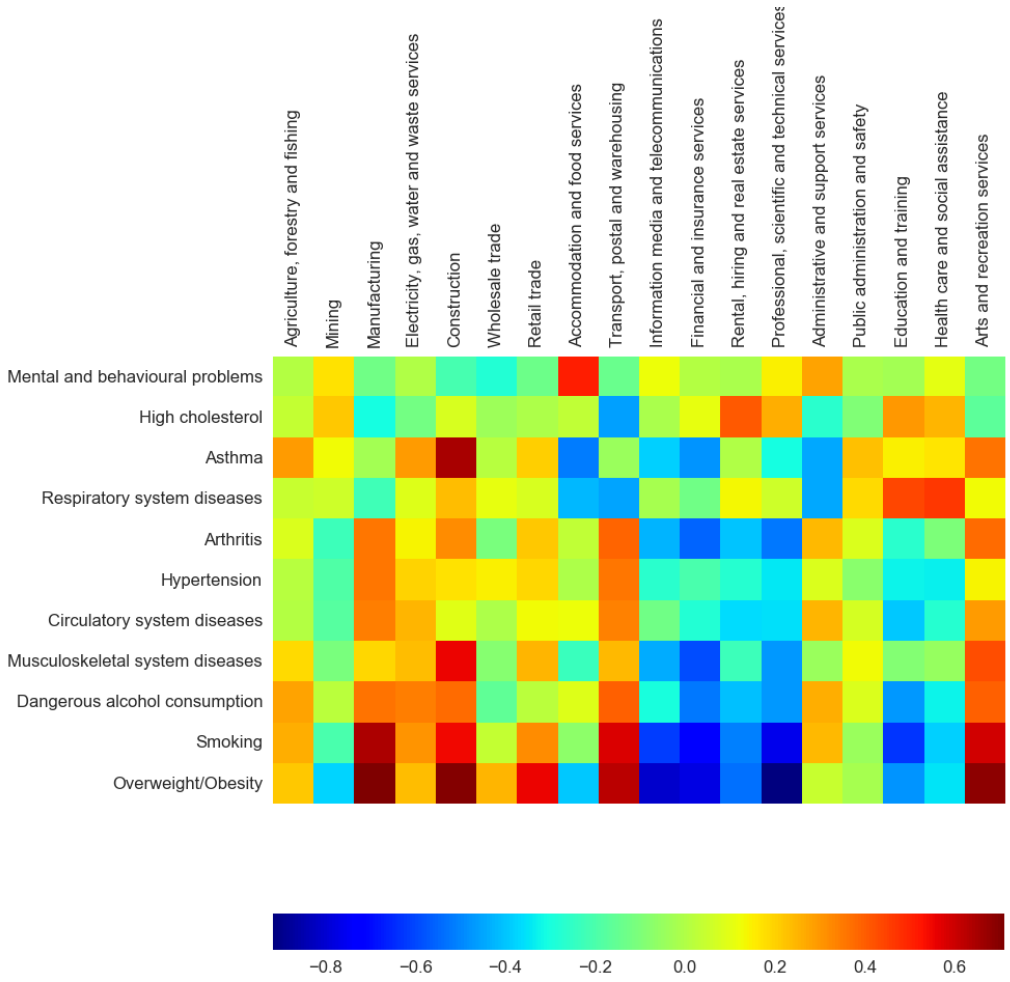

Figure 6: Pearson correlation by industry

Although there are some compelling correlations between occupation and health risk factors (Smoking, overweight/obesity, and dangerous alcohol consumption), they are not the focus of the study so won't be investigated further.

Since it was impractical to analyse every correlation, several with high r-values were chosen arbitrarily. These were plotted on scatter plots where each point represents a single SA2, and a regression line drawn to illustrate each suspected trend (**Figures 7-10**).

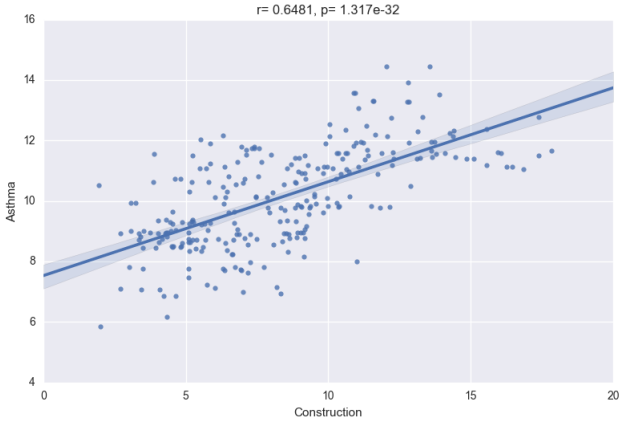

Figure 7: Construction workers vs. asthma (%)

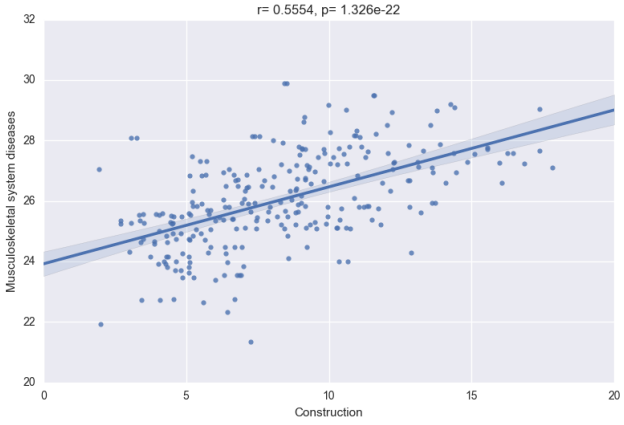

Figure 8: Construction workers vs. musculoskeletal system diseases (%)

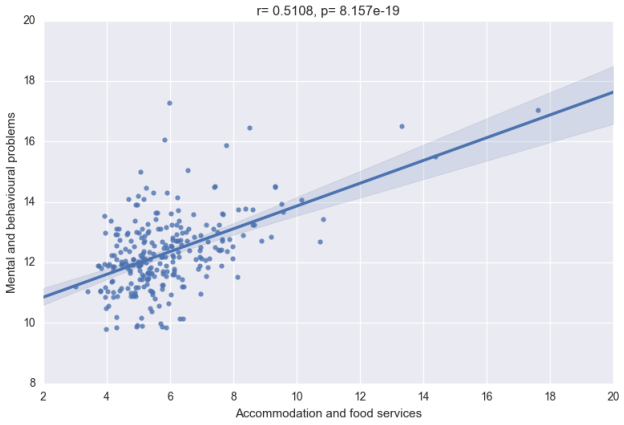

Figure 9: Accommodation and food services workers vs. mental and behavioural problems (%)

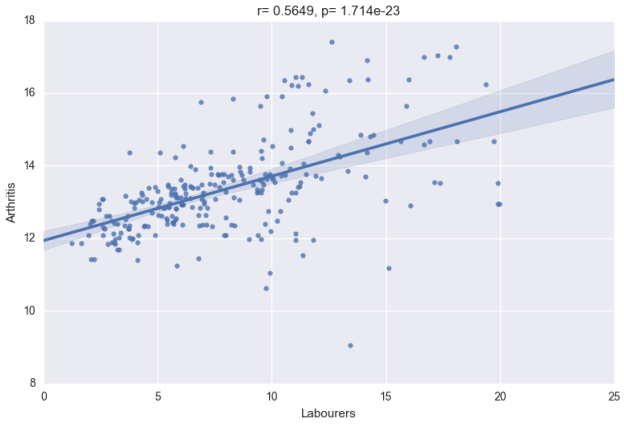

Figure 10: Labourers vs. arthritis (%)

Figure 9 illustrates how outliers affect the results, where several outliers in the top-right section of the plot are influencing the trend line. It is possible that the results would have been more accurate had outliers been removed in the preprocessing stage, but the presence of outliers was decided to be preferable to suppressing potentially significant results.

As evident from **Figures 8-10**, there is a clear relationship between labourers and arthritis, construction workers and musculoskeletal system diseases, and construction workers and asthma. The first two of these aren't particularly unexpected, as physical jobs seem likely to cause strain on muscles and joints. However, the link between construction workers and asthma is less obvious and arouses further background investigation.

Arif et al. found that construction workers in the US have a higher risk of developing work-related asthma than the general population[1]. Furthermore, Sauni et al. had a similar result in a study of Finnish males, suggesting that the fine particulates that construction workers are exposed to increase their risk of asthma[2]. This reinforces the results of this study and suggests them to be applicable in general, not just for the geographical areas addressed by the original datasets.

In order to investigate the possibility of inferring chronic disease and health risk factors for areas where employment data is available, a linear regression model was built using the results of the study. Once the model was trained on training data which consisted of the results excluding the test data, an accurate prediction was able to be generated (**Figure 11**). With more diverse training data and tweaks to the model, a higher accuracy model is possible. Of course, this incorrectly assumes that there are no unknown confounding variables, but further research could identify and incorporate them into the model.

	True	Pred.	MSE
Mental and behavioural problems	12	12.201	0.039
High cholesterol	35.02	33.838	1.39
Asthma	9.246	10.335	1.19
Respiratory system diseases	30.21	30.144	0.004
Arthritis	12.4	12.819	0.177
Hypertension	9	9.536	0.288
Circulatory system diseases	13.94	15.046	1.22
Musculoskeletal system diseases	24.71	25.721	1.02
Dangerous alcohol consumption	2.426	2.5737	0.022
Smoking	11.77	12.742	0.953
Overweight/Obesity	51.36	53.421	4.245

Figure 11: Linear regression model: true vs predicted values

Value

The advantage of this study is that information and knowledge was extracted from a dataset which was incomprehensible in its original form. Given the multidimensional nature of the datasets, it was impossible to infer any useful information about correlations within. By taking advantage of computer-aided data processing, over 110,000 data points were condensed into human-readable and easily visualised results. Furthermore, an accurate model that can predict chronic illness is invaluable to any party who needs to quickly analyse risk without the expense of obtaining data from traditional sources such as surveys or a census.

Challenges and reflections

It is important to note that while the study found correlations with low p-values and moderate r-values, causality is not directly implied. There is a vast realm of possible unknown confounding variables, which need to be identified and carefully analysed before acting on the results of this study. While a few possible variables are indicated as health risk factors which could cause chronic disease, these represent a tiny proportion of potential candidates for confounding variables. For example, socio-economic status is unknown in the data and could be a large factor in some of the diseases investigated.

Throughout the course of this study, there were several issues encountered:

- Multidimensional data was difficult to visualise during the preprocessing and integration stages, causing uncertainty whether the direction chosen was worthwhile. It wasn't known until the results stage if the data would show any useful information.
- There was no known general way to represent the data visually, hence the need to select several pairs to plot against each other. While this still provides useful visualisations, a meaningful way to display the inherent correlation over the whole dataset would be valuable.
- The correlation matrices suitably display relationships between variables, but are cluttered with low correlations thus need a certain level of comprehension to see the useful information.
- The linear regression model, while a good proof-of-concept, needs tweaking and work beyond the scope of this study to be truly useful.

Question resolution

The question posed was answered sufficiently, and the goal of abstracting away geographical area and relating employment directly with chronic diseases was achieved. However, more accurate results and novel ways of interpreting the data are possible with deeper statistical analysis. Further research would identify unknown variables and refine the results to be more general, possibly including more data from geographical areas outside Greater Melbourne.

The results of this study could be valuable to doctors diagnosing illnesses given a patient's work history, workplace safety organisations determining appropriate PPE for employees to wear, insurance companies in risk analysis, and workers' unions in negotiating for a fair workplace. Additionally, an accurate model for predicting chronic disease rates would be useful for a government body deciding where to build medical facilities.

Code

The Python libraries used were `pandas`, `scikit-learn`, `matplotlib`, and `seaborn`, and `numpy`.

The following functionality was implemented in Python and is available in `load_data.py`, `plot.py`, and `model.py`:

- Read data from `.csv` files and store internally in `pandas DataFrame` object
- Clean missing and zero data points from set
- Inner join datasets on SA2
- Calculate Pearson coefficient for each pair of variables
- Generate lag plot, correlation matrices, and scatter plots using `matplotlib` and `seaborn`
- Train and test linear regression model using `scikit-learn LinearRegression`

Unused code written for Phase 2 is available in `old.py`.

References

[1] Ahmed A. Arif et al. "Occupational exposures associated with work-related asthma and work-related wheezing among U.S. workers". In: *Am. J. Ind. Med.* 44.4 (2003), pp. 368–376. DOI: 10.1002/ajim.10291.

[2] R. Sauni et al. "Increased risk of asthma among Finnish construction workers". In: *Occupational Medicine* 53.8 (2003), pp. 527–531. DOI: 10.1093/occmed/kqg112.