Analysis of Sorafenib therapy on patient survival

Alex Pegg and Harsha Payidiparty

Abstract

We are tasked with interpreting a dataset of cancer therapy, of which is formatted as survival analysis — meaning we have data on events and censoring over a period of time.

We primarily used multivariate linear regression for our method, with output as the expected time that a patient will stay on treatment.

We find that, for a given patient, their province, sex, and if they are on other anti-cancer therapy, are significant factors in their survival, with males and those on other therapies having a longer expected survival time. For provinces with a higher output, we also find a positive correlation with GDP per capita, and negative correlation with the percent of smokers and obese persons. While we find age has some negative impact on survival, its small regression coefficient and high variance that leads us to conclude it is negligible in describing the expected time a patient remains on treatment.

However, we also find the dataset to be insufficient to draw strong conclusions, noting the low sample size, missing data, and ambiguity of the term 'event' — which may refer to a death, recovery or otherwise. Admittedly, our model's ability to describe the data is limited, with a low r^2 coefficient and high variance.

Methodology

All calculations are done in python, using numpy and sklearn for regression. Plotting is done in python using matplotlib and seaborn.

Patients at any given month

This value was trivially generated using the aggregate of all patients.

Rate of discontinuation

We assume that discontinuation refers to patients who are censored.

Using a log regression, where the input is time, and the output is the amount of people still in therapy (excluding patients who encounter events), we were able to get a function that mapped strongly to the ground truth values. Given this function, we can get the rate of discontinuation from the derivative.

However, this log derivative algebraic expression is not necessarily a “rate”. Another method of finding the rate of growth used in Econometrics is by taking the log of the output instead of the input. This means that for a unit change of our input, we can find the percent change in our output. This method is only an approximation, but it may be a more helpful variable to describe the rate than an expression. Using this method, we were able to get a value that is an approximation of the percent change in people taking therapy over any given month. Because this value was not perfect, using a brute force algorithm, we fine-tuned the value to minimise the sum of squares, ultimately generating a percentage rate that describes the data well.

Predicting successful therapy and discontinuation

Questions #2 and #4 are similar because they ask for an explanation of the data. To do this, we used linear regression once again. This is because unlike popular methods of predicting data such as neural networks, linear regression is interpretable because we can directly see the effects of a unit change of a parameter through the coefficients.

The number of data points is low, and many of the rows are aggregations of the data which are redundant — providing no information gain. The true number of rows is 242, and of those 242 rows, 175 have incomplete data in the form of null fields for the Con_ACT variable. Whereas typically you might divide research 60/20/20 into training/validation/test data, because of the low number of samples, we decided to simply use all data for training to increase our chances of finding signal. The cost of this choice is that our model may overfit the training examples, but because we are using linear regression (polynomial degree 1), the low complexity of our model makes it more difficult to overfit. Furthermore, as we found, the data is so noisy that any small sample of test examples are sure to produce high loss.

Our choice of regression output is the expected length of time a patient is expected to remain in therapy. Our input parameters are the province, which were encoded as dummy variables. To prevent the dummy variable trap, we excluded the province “UNKWN”. We also used dummy variables for if the patient is on other cancer therapy and if the patient is male. A decision that we made was to convert null data for concurrent cancer therapy to false, so that we may use a binary value for if patients are using other cancer therapies. When using age as a parameter, the minimum age is used in an interval.

Because the term ‘event’ is ambiguous between implying a death or a recovery, success is incredibly difficult to define due to the fact that death and recovery have opposing implications. Because this is survival analysis data, we will make the assumption that the objective is to survive as long as possible — meaning a successful therapy is a therapy is a longer therapy. This means maximising the output.

We judged the efficacy of our models using the r^2 values of the regression, using the adjusted r^2 values to compare potential models.

Given the coefficients for different provinces, we used outside data on factors we thought could explain provincial trends: prevalence of alcoholism, prevalence of obesity and the GDP per capita.

Results

1. How many patients stay on treatment for at least 9 months?

On the 9th month, there are 704 / 1441 patients still on treatment.

2. What might predict successful therapy?

For predicting successful therapy, we will be using regression on the expected month an event occurs.

Comparing figures 1a) and 1b), where 1a) includes age as a parameter and 1b) does not, we can conclude that age does not play a statistically significant role in determining output in this regression. The coefficient is small, and also the adjusted r^2* are the same or higher when excluding age. This adjusted r^2* value diminishes the r^2 based on the number of parameters you include in the model, to reduce overfitting. It's surprising that age is not important, as we thought that age would have a negative impact on expected time on treatment, given that older people are more frail.

Despite that, we have significant coefficients on the other parameters. For example, living in the Atlantic versus living in PC, you might see a 5 month increase in your time on therapy (0.469 versus 5.631 coefficient). This geographical relationship is interesting, because assuringly these patients are receiving the same level of treatment and care. To explain this geographical relationship, we consulted 3 factors: alcoholism, obesity and income.

[see figure 2a)] One might predict that higher levels of alcoholism would lead to a faster expected event time, because alcoholism is a strong factor in the liver cancer. However, this is not the case, and we actually see a positive relationship, where provinces with a higher level of alcoholism have a higher expected event time.

[see figure 2b)] For obesity, we see a clear inverse relationship, where provinces with higher levels of obesity have a lower expected event time. This is not a novel idea, obesity is well known to bring about health complications.

[see figure 2c)] For income, we see another positive trend, where provinces with higher GDP per capita experience a higher expected event time. This may be explained by the fact that provinces with higher income can afford better healthcare and lifestyle choices associate with higher income such as diet and stress.

The coefficient that demonstrates the affects of being on other therapies has a positive relationship with expected event time. This is not a surprising result, as trivially you are receiving more help. But it does reveal that multiple treatments may not be redundant, and for a patient with this cancer they can increase their survival prospects through taking other therapies as well. It is important to note that there is not much data on those who are on other therapies, as a lot of the data in this column is null. Perhaps this factor is correlated with the income of province, but it is difficult to tell with a small sample size.

Gender seems to be an important factor here, where males have a higher expected event time. This is an interesting result because this cancer primarily affects males (cancer.org, 2019), so one might think that they are especially weak to this cancer.

It is important to note that whilst some correlation exists, these r^2 values are low. Regression was not sufficient to accurately predict this model. That may be endemic to the problem, which has few samples and parameters that may or may not be relevant.

3. What is the monthly rate of discontinuation?

Figure 3a) shows a logarithmic regression of time on remaining uncensored patients, where any patients that eventually experience an event have been removed from the totals. We can see a strong fit here, and a high r^2 value of 0.945. The rate of this function, or its derivative, is 25.33/x. However, this is an algebraic expression and may not be considered a rate.

If we perform a regular linear regression on the natural log of the output, we can get an approximation of the percent change with our regression coefficient. This is found to be -11.41%, which is shown by the blue dots in figure 3b). However, from inspection we can see this could be improved. Using a brute force algorithm that minimises the sum of square error, we were able to arrive at a new value -8.50%, which has a much lower sum of square error than the result found using the regression coefficient, and as you can see from the orange dots in figure 3b), it fits the data much better.

4. What are some potential reasons for discontinuation of therapy?

Comparing censor regressions to the event regressions in figure 1b), we can see a lot of similarities. Polarity is the same, which means any parameter that has a positive effect on expected months until an event also has a positive effect on the expected months until a censored occurrence. Being male and on other therapy leads to a higher expected censor time. However, for some reason Québécois last the longest. This is strange because they have the third-lowest coefficient for the event regression, after “UNKWN” and Atlantic. As seen in figure 4), Quebec has some of the cheapest insurance in Canada (although this figure is for auto insurance). While healthcare is free in Canada, this does not preclude people from private health care, especially when it comes to deadly diseases like cancer. However, cheaper insurance would not explain why Ontario has a similarly high coefficient, as it has by far the most expensive insurance in Canada.

Conclusion

In conclusion, we find a number of factors that predict how long one may stay on Sorafenib using linear regression. The province that a patient is in has a large effect, where provinces with low obesity and higher income are desirable. Being on other therapies and being a male will also increase a patients estimated time on therapy.

We also find that the regression was not strong. We have low r^2 values from 0.1 to 0.2, which are not statistically significant. This might be the nature of the data, which is highly sporadic. With better data, namely with more samples and less missing values, we might be able to perform better inference.

Bibliography

(n.d.). Retrieved November 9, 2020, from https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310009610&pickMembers[0]=1.11&pickMembers[1]=3.1&cubeTimeFrame.startYear=2016&cubeTimeFrame.endYear=2019&referencePeriods=20160101,20190101

Bedford, P. B., & 13, A. (2019, August 13). Frequency of alcohol consumption Canada by province 2018. Retrieved November 9, 2020, from https://www.statista.com/statistics/895635/frequency-of-alcohol-consumption-by-province-canada/

Liver Cancer Risk Factors. (n.d.). Retrieved November 9, 2020, from https://www.cancer.org/cancer/liver-cancer/causes-risks-prevention/risk-factors.html

Groupement des assureurs automobiles: Statistics. (n.d.). Retrieved November 9, 2020, from https://gaa.qc.ca/statistics/automobile-insurance-rates/comparison-by-province

Figures

1a) Coefficients and r^2 of regression including age

	y-intercept	Province						on other therapy	is male	age	r^2	r^2*
		BC	Atlantic	AB	ON	QC	Praries					
total	2.468	6.371	3.552	6.789	6.130	6.322	2.951	2.045	2.474	0.028	0.138	0.109
events	4.724	5.712	0.518	5.092	3.114	2.949	3.852	3.590	1.365	-0.021	0.145	0.102
censored	2.723	8.170	4.830	7.951	8.763	9.710	2.337	0.797	1.903	0.035	0.12	0.082

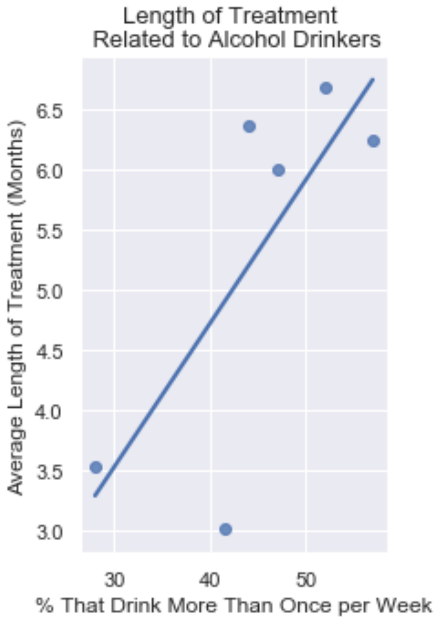
1b) Coefficients and r^2 of regression excluding age

	y-intercept	Province						on other therapy	is male	r^2	r^2*
		BC	Atlantic	AB	ON	QC	Praries				
total	3.707	6.37	3.53	6.69	6.00	6.24	3.02	2.13	2.57	0.135	0.109
events	3.844	5.631	0.469	5.058	3.105	2.911	3.726	3.558	1.313	0.158	0.142
censored	4.409	8.233	4.785	7.803	8.576	9.635	2.395	0.890	1.972	0.117	0.083

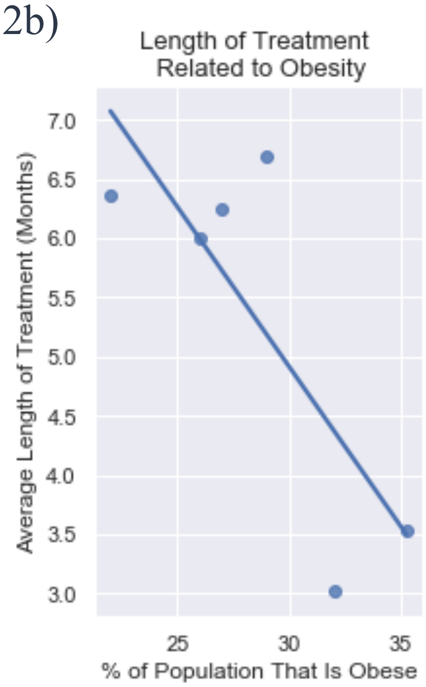
1c) Coefficients and r^2 of regression excluding age and excluding null

	y-intercept	Province						on other therapy	is male	r^2	r^2*
		BC	Atlantic	AB	ON	QC	Praries				
total	4.098	5.528	3.173	7.379	6.592	7.602	3.693	1.684	1.615	0.186	0.151
events	3.494	6.730	1.075	4.322	3.427	4.574	4.518	2.789	2.012	0.185	0.138
censored	4.820	6.352	5.095	9.888	9.202	9.834	3.163	0.982	0.238	0.117	0.083

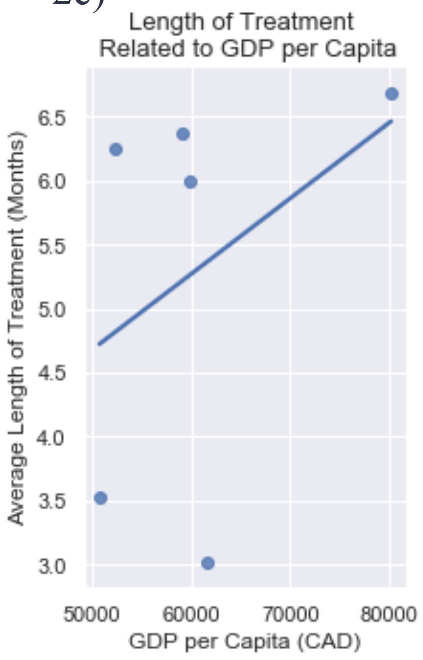
2a)



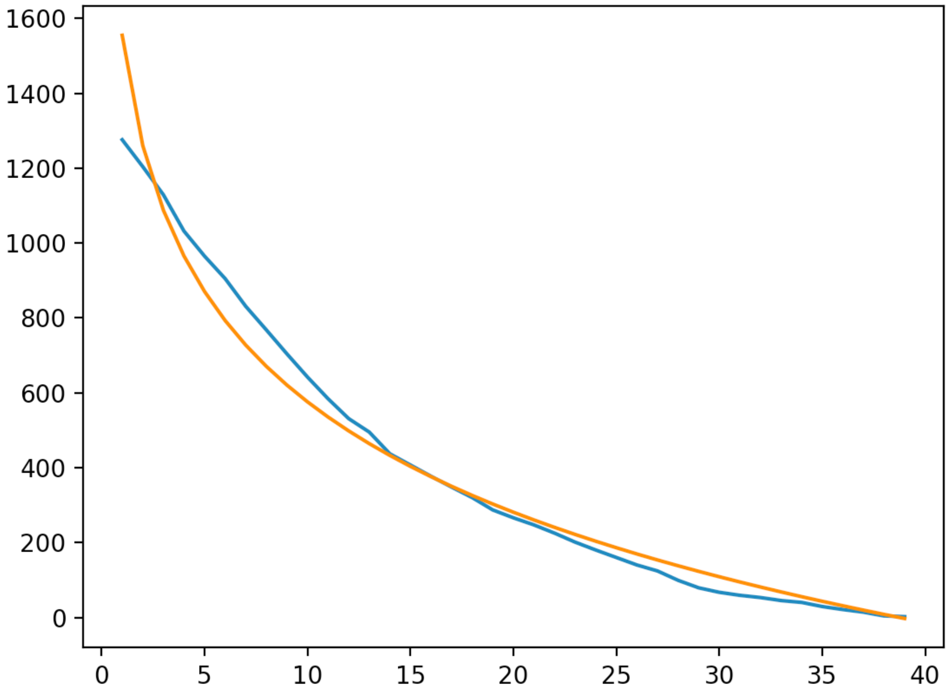
2b)



2c)



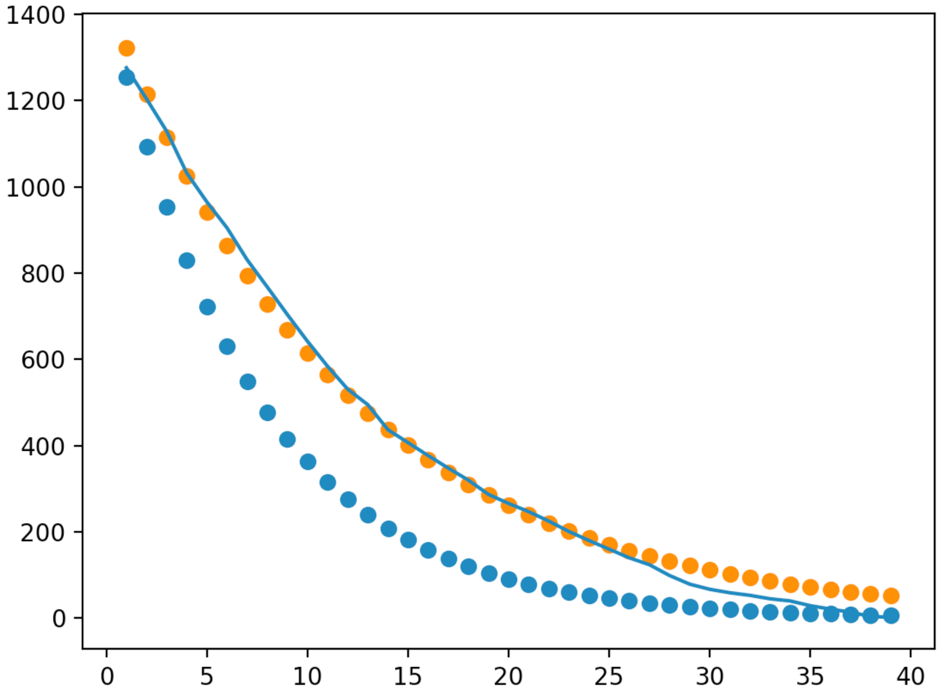
3a)



(blue line) Time (months) against patients in therapy

(orange line) Logistic regression of time (months) against patients in therapy

3b)

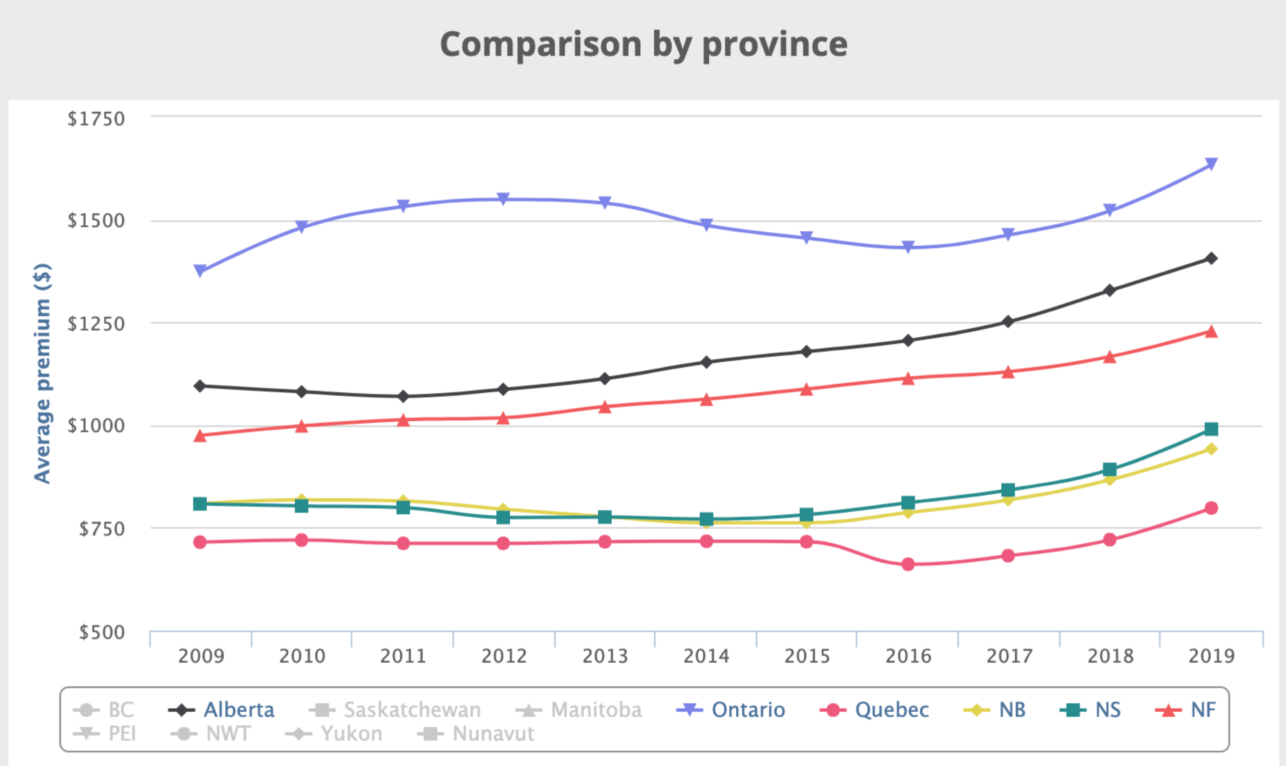


(blue line) Time (months) against patients in therapy

(blue dots) Growth rate of -11.41% (calculated)

(orange dots) Growth rate of -7.05% (brute forced to minimise SSE)

4a)



Price of (auto) insurance across Canada (gaa.qc.ca, 2019)