Stepwise Regression Tutorial in Python

Finding meaning in data using stepwise regression

Ryan Kwok Mar 10 · 9 min read

Photo by Franki Chamaki on Unsplash

How do you find meaning in data? In our mini project, my friend @ErikaSM and I seek to predict Singapore's minimum wage if we had one, and documented that process in an article over here. If you have not read it, do take a look.

Since then, we have had comments on our process and suggestions to develop deeper insight into our information. As such, this follow-up article outlines two main objectives, finding meaning in data, and learning how to do stepwise regression.

The Context

In the previous article, we discussed how the talk about a minimum wage in Singapore has frequently been a hot topic for debates. This is because Singapore uses a progressive wage model and hence does not have a minimum wage.

The official stance of the Singapore Government is that a competitive pay structure will motivate the labour force to work hard, aligned with the value of Meritocracy embedded in Singapore culture. Regardless of the arguments for or against minimum wages in Singapore, the poor struggle to afford necessities and take care of themselves and their families.

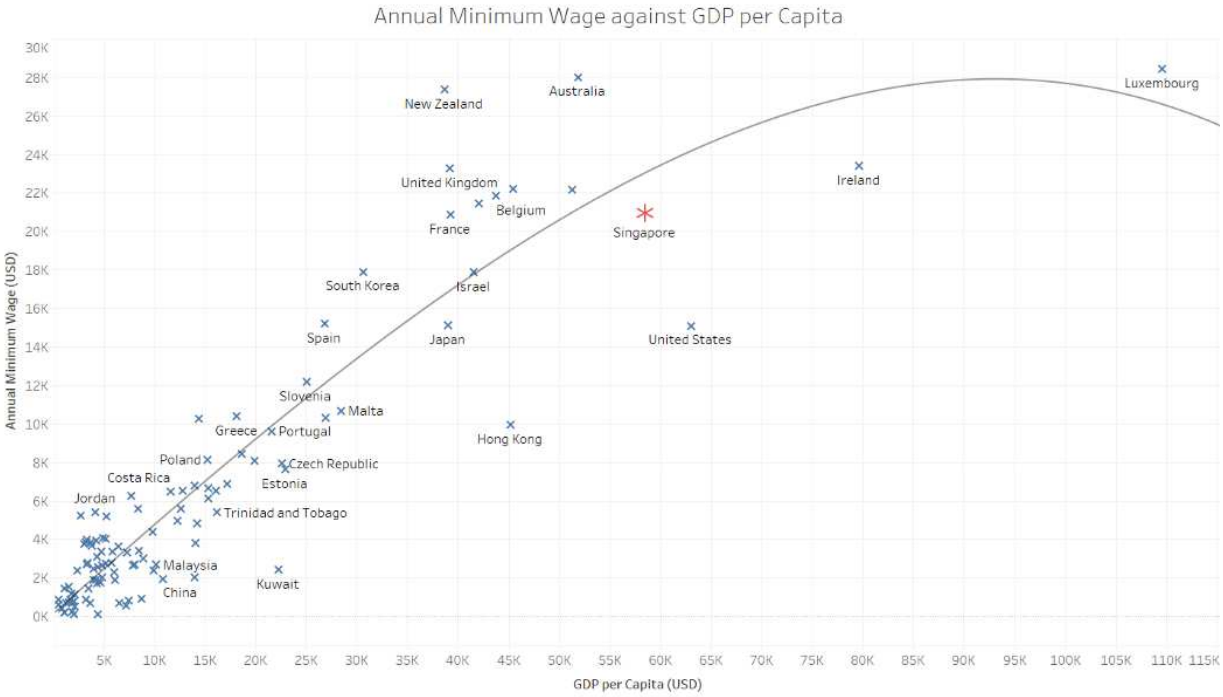
We took a neutral stance acknowledging the validity of both sides of the argument and instead presented a comparison of a prediction of Singapore's minimum wage using certain metrics across different countries. The predicted minimum wage was also contrasted with the wage floors in the Progressive Wage Model (PWM) across certain jobs to spark some discussion about whether the poorest are earning enough.

The Methodology

We used data from Wikipedia and World Data to collect data on minimum wage, cost of living, and quality of life. The quality of life dataset includes scores in a few categories: Stability, Rights, Health, Safety, Climate, Costs, and Popularity.

The scores across the indicators and categories were fed into a linear regression model, which was then used to predict the minimum wage using Singapore's statistics as independent variables. This linear model was coded on Python using sklearn, and more details about the coding can be viewed in our previous article. However, I will also briefly outline the modelling and prediction process in this article as well.

The predicted annual minimum wage was US$20,927.50 for Singapore. A brief comparison can be seen in this graph below.



Annual Minimum Wage for Singapore was about $20,000 USD (Image from Author)

Our professor encouraged us to use stepwise regression to better understand our variables. From this iteration, we incorporated stepwise regression to assist us in dimensionality reduction not only to produce a simpler and more effective model, but to derive insights in our data.

Stepwise Regression

So what exactly is stepwise regression? In any phenomenon, there will be certain factors that play a bigger role in determining an outcome. In simple terms, stepwise regression is a process that helps determine which factors are important and which are not. Certain variables have a rather high p-value and were not meaningfully contributing to the accuracy of our prediction. From there, only important factors are kept to ensure that the linear model does its prediction based on factors that can help it produce the most accurate result.

In this article, I will outline the use of a stepwise regression that uses a backwards elimination approach. This is where all variables are initially included, and in each step, the most statistically insignificant variable is dropped. In other words, the most ‘useless’ variable is kicked. This is repeated until all variables left over are statistically significant.

The Coding Bits

Before proceeding to analyse the regression models, we first modified the data to reflect a monthly wage instead of annual wage. This was because we recognised that most people tend to view their wages in months rather than across the entire year. Expressing our data as such would allow our audience to better understand our data. However, it is also worth noting that this change in scale would not affect the modelling process or the outcomes.

Looking at our previous model, we produced the statistics to test the accuracy of the model. But before that, we would first have to specify the relevant X and Y columns, and obtain that information from the datafile.

```
## getting column names
x_columns = ["Workweek (hours)", "GDP per capita", "Cost of Living
Index", "Stability", "Rights", "Health", "Safety", "Climate",
"Costs", "Popularity"]
y = data["Monthly Nominal (USD)"]
```

Next, to gather the model statistics, we would have to use the statmodels.api library. Here, a function is created which grabs the columns of interest from a list, and then fits an ordinary least squares linear model to it. The statistics summary can then be very easily printed out.

```
## creating function to get model statistics
import numpy as np
import statsmodels.api as sm

def get_stats():
    x = data[x_columns]
    results = sm.OLS(y, x).fit()
    print(results.summary())

get_stats()
```

```
                            OLS Regression Results
==============================================================================================
Dep. Variable:     Monthly Nominal (USD)   R-squared (uncentered):                   0.925
Model:                               OLS   Adj. R-squared (uncentered):              0.917
Method:                    Least Squares   F-statistic:                              120.0
Date:                   Tue, 09 Mar 2021   Prob (F-statistic):                    5.01e-50
Time:                           17:00:35   Log-Likelihood:                         -724.27
No. Observations:                    107   AIC:                                      1469.
Df Residuals:                         97   BIC:                                      1495.
Df Model:                             10
Covariance Type:               nonrobust
==============================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Workweek (hours)          -5.7263      4.335     -1.321      0.190     -14.329       2.877
GDP per capita             0.0148      0.002      5.912      0.000       0.010       0.020
Cost of Living Index       8.0372      2.331      3.448      0.001       3.410      12.664
Stability                 -3.1099      2.132     -1.459      0.148      -7.341       1.122
Rights                     5.6809      2.161      2.629      0.010       1.392       9.970
Health                     1.1002      1.485      0.741      0.460      -1.846       4.047
Safety                    -0.0602      1.486     -0.041      0.968      -3.009       2.888
Climate                   -1.2375      1.328     -0.932      0.354      -3.874       1.399
Costs                     -1.9052      2.078     -0.917      0.361      -6.029       2.218
Popularity                 4.3583      1.733      2.515      0.014       0.919       7.798
==============================================================================
Omnibus:                       14.866   Durbin-Watson:                   2.162
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               40.070
Skew:                           0.372   Prob(JB):                     1.99e-09
Kurtosis:                       5.904   Cond. No.                     5.03e+03
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 5.03e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Original Regression Statistics (Image from Author)

Here we are concerned about the column "P > |t|". Quoting some technical explanations from the UCLA Institute for Digital Research and Education, this column gives the 2-tailed p-value used in testing the null hypothesis.

> *"Coefficients having p-values less than alpha are statistically significant. For example, if you chose alpha to be 0.05, coefficients having a p-value of 0.05 or less would be statistically significant (i.e., you can reject the null hypothesis and say that the coefficient is significantly different from 0)."*

In other words, we would generally want to drop variables with a p-value greater than 0.05. As seen from the initial summary above, the least statistically significant variable is "Safety" with a p-value of 0.968. Hence, we would want to drop "Safety" as a variable as shown below. The new summary is shown below as well.

```
x_columns.remove("Safety")
get_stats()
```

OLS Regression Results

Dep. Variable:	Monthly Nominal (USD)	R-squared (uncentered):	0.925
Model:	OLS	Adj. R-squared (uncentered):	0.918
Method:	Least Squares	F-statistic:	134.7
Date:	Tue, 09 Mar 2021	Prob (F-statistic):	4.18e-51
Time:	17:00:35	Log-Likelihood:	-724.27
No. Observations:	107	AIC:	1467.
Df Residuals:	98	BIC:	1491.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Workweek (hours)	-5.7713	4.169	-1.384	0.169	-14.044	2.502
GDP per capita	0.0148	0.002	5.950	0.000	0.010	0.020
Cost of Living Index	8.0602	2.249	3.583	0.001	3.596	12.524
Stability	-3.1298	2.064	-1.516	0.133	-7.226	0.967
Rights	5.6658	2.118	2.676	0.009	1.464	9.868
Health	1.0793	1.385	0.779	0.438	-1.669	3.828
Climate	-1.2384	1.321	-0.937	0.351	-3.860	1.384
Costs	-1.9067	2.067	-0.923	0.358	-6.008	2.194
Popularity	4.3506	1.714	2.539	0.013	0.950	7.751

Omnibus:	14.831	Durbin-Watson:	2.164
Prob(Omnibus):	0.001	Jarque-Bera (JB):	40.015
Skew:	0.370	Prob(JB):	2.05e-09
Kurtosis:	5.903	Cond. No.	4.89e+03

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 4.89e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

Regression Statistics after only removing “Safety” (Image from Author)

This time, the new least statistically significant variable is “Health”. Similarly, we would want to remove this variable.

```
x_columns.remove("Health")
get_stats()
```

OLS Regression Results

Dep. Variable:	Monthly Nominal (USD)	R-squared (uncentered):	0.925
Model:	OLS	Adj. R-squared (uncentered):	0.919
Method:	Least Squares	F-statistic:	152.1
Date:	Tue, 09 Mar 2021	Prob (F-statistic):	4.43e-52
Time:	17:00:35	Log-Likelihood:	-724.60
No. Observations:	107	AIC:	1465.
Df Residuals:	99	BIC:	1487.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Workweek (hours)	-5.9825	4.152	-1.441	0.153	-14.221	2.255
GDP per capita	0.0147	0.002	5.956	0.000	0.010	0.020
Cost of Living Index	8.3726	2.209	3.790	0.000	3.990	12.756
Stability	-2.9028	2.040	-1.423	0.158	-6.950	1.144
Rights	6.0625	2.051	2.955	0.004	1.992	10.133
Climate	-1.3762	1.307	-1.053	0.295	-3.969	1.217
Costs	-1.3918	1.954	-0.712	0.478	-5.269	2.486
Popularity	4.5995	1.680	2.737	0.007	1.265	7.934

Omnibus:	13.567	Durbin-Watson:	2.176
Prob(Omnibus):	0.001	Jarque-Bera (JB):	34.196
Skew:	0.341	Prob(JB):	3.75e-08
Kurtosis:	5.684	Cond. No.	4.85e+03

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 4.85e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

Regression Statistics after removing “Safety” and “Health” (Image from Author)

We continue this process until all p-values are below 0.05.

```
x_columns.remove("Costs")
x_columns.remove("Climate")
x_columns.remove("Stability")
```

```
                            OLS Regression Results
==============================================================================================
Dep. Variable:     Monthly Nominal (USD)   R-squared (uncentered):                   0.922
Model:                               OLS   Adj. R-squared (uncentered):              0.918
Method:                    Least Squares   F-statistic:                              241.8
Date:                   Tue, 09 Mar 2021   Prob (F-statistic):                    6.98e-55
Time:                           17:00:35   Log-Likelihood:                         -726.39
No. Observations:                    107   AIC:                                      1463.
Df Residuals:                        102   BIC:                                      1476.
Df Model:                              5
Covariance Type:               nonrobust
==============================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Workweek (hours)      -11.0565      2.107     -5.246      0.000     -15.237      -6.876
GDP per capita          0.0154      0.002      6.501      0.000       0.011       0.020
Cost of Living Index    7.6078      2.103      3.618      0.000       3.437      11.779
Rights                  4.6458      1.436      3.236      0.002       1.798       7.494
Popularity              4.6057      1.631      2.824      0.006       1.371       7.841
==============================================================================
Omnibus:                       11.471   Durbin-Watson:                   2.097
Prob(Omnibus):                  0.003   Jarque-Bera (JB):               25.234
Skew:                           0.301   Prob(JB):                     3.32e-06
Kurtosis:                       5.302   Cond. No.                     2.93e+03
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 2.93e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Finally, we find that there are 5 variables left, namely Workweek, GDP per Capita, Cost of Living Index, Rights, and Popularity. Since each of the p-values are below 0.05, all of these variables are said to be statistically significant.

We can now produce a linear model based on this new set of variables. We can also use this to predict Singapore's minimum wage. As seen, the predicted monthly minimum wage is about $1774 USD.

```
## creating a linear model and prediction
x = data[x_columns]
linear_model = LinearRegression()
linear_model.fit(x, y)
sg_data = pd.read_csv('testing.csv')
x_test = sg_data[x_columns]
y_pred = linear_model.predict(x_test)
print("Prediction for Singapore is ", y_pred)

>> Prediction for Singapore is  [1774.45875071]
```

Finding Meaning in the Data

This is the most important part of the process. Carly Fiorina, former CEO of Hewlett-Packard, once said: "The goal is to turn data into information, and information into insight." This is exactly what we aim to achieve.

> *"The goal is to turn data into information, and information into insight."*
> *~ Carly Fiorina, former CEO of Hewlett-Packard*

From just looking at the variables, we would have easily predicted which were statistically significant. For example, the GDP per Capita and Cost of Living Index would logically be good indicators of the minimum wage in a country. Even the number of hours in a workweek would make sense as an indicator.

However, we noticed that "Rights" was still included in the linear model. This spurred us to first look at the relationship between Rights and Minimum Wage. Upon plotting the graph, we found this aesthetically pleasing relationship.

Monthly Minimum Wage against Rights (Image by Author)

Initially, we wouldn't have considered Rights to be correlated to Minimum Wage since the more obvious candidates of GDP and Cost of Living stood out more as contributors to the minimum wage level. This made us reconsider how we understood minimum wage and compelled us to dig deeper.

From World Data, "Rights" involved civil rights, and revolved mainly around people's participation in politics and corruption. We found that the Civil Rights Index includes democratic participation by the population and measures to combat corruption. This index also involves public perception of the government including data from Transparency.org.

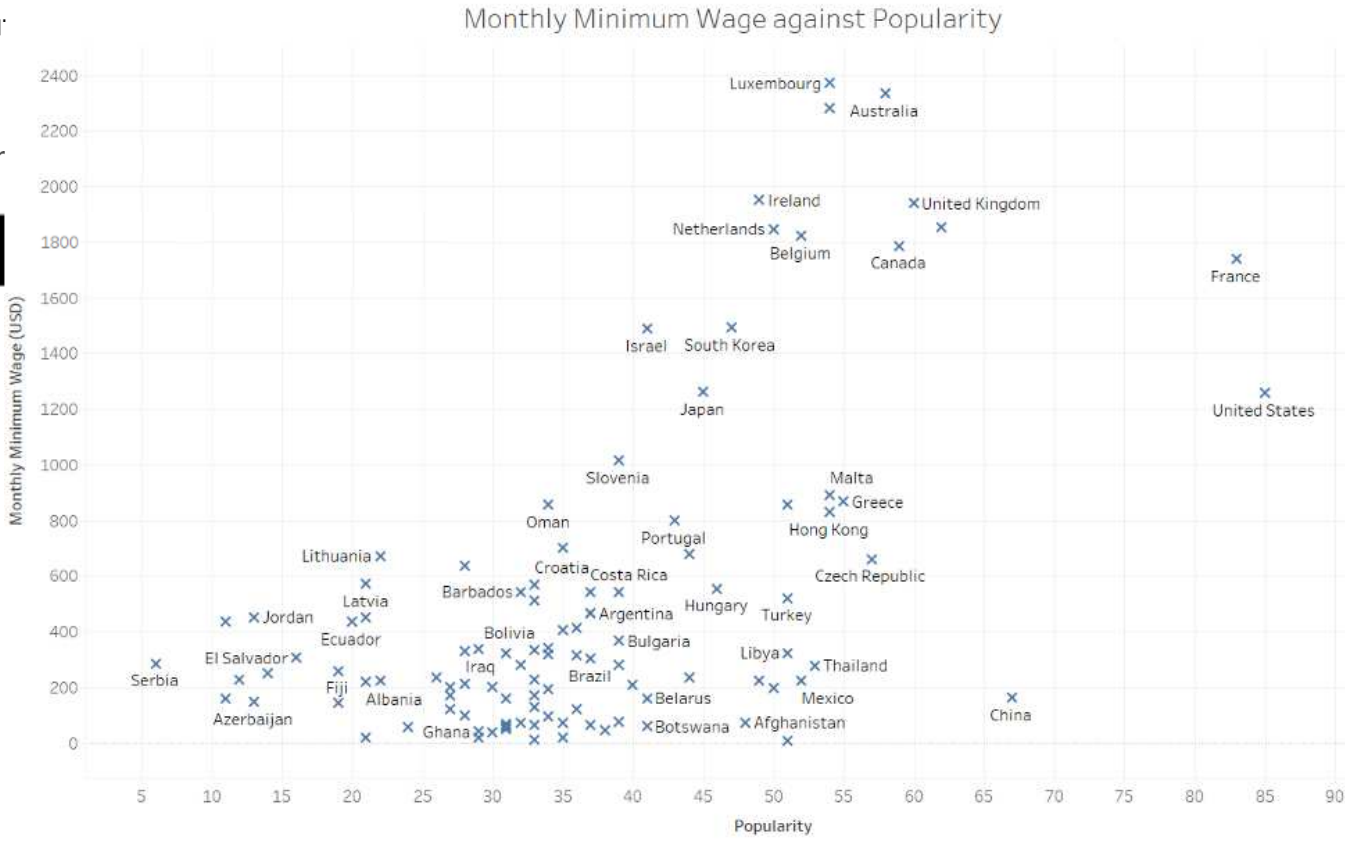
> *"In addition, other factors include democratic participation by the population and (with less emphasis) measures to combat corruption. In order to assess not only the measures against corruption, but also its perception by the population, the corruption index based on Transparency.org was also taken into account."*

This forced us to consider the correlation between Civil Rights and minimum wage. Knowing this information, we did further research and found several articles that might

hts, relea
ger mini
w-payin
lso likely
over tin

from Wo

> *"Th[illegible]al [illegible] and the [illegible] of foreign [illegible] before evaluated as indicators [illegible] country's popularity. A lower rating was also used to compare the refugee situation in the respective c[illegible] A higher number of foreign refugees results in higher popularity, while a high number of fleeing refug[illegible] reduces popularity."*



Monthly Minimum Wage against Popularity (Image by Author)

At first glance, it seems like there is no correlation. However, if we consider China, France, USA, a Spain as outliers, the majority of the data points seem to better fit an exponential graph. This rais questions. Firstly, why is there a relationship between Popularity and Minimum Wage? Secondly, these four countries outliers?

To be very honest, this stumped us. We simply could not see any way where popularity could be correlated to a minimum wage. Nevertheless, there was an important takeaway: that popularity i somehow statistically significant in predicting a minimum wage of a country. While we might not people to discover that relationship, this gives insight into our otherwise less meaningful data.

Conclusion

It is important to bring back the quote from Carly Fiorina, “The goal is to turn data into informatic information into insight.” We as humans require tools and methods to convert data into informati experience/knowledge to convert that information into insight.

We first used Python as a tool and executed stepwise regression to make sense of the raw data. Th discover not only information that we had predicted, but also new information that we did not ini consider. It is easy to guess that Workweek, GDP, and Cost of Living would be strong indicators of minimum wage. However, it is only through regression that we discovered that Civil Rights and Popularity are also statistically significant.

In this case, there were research online that we found that could possibly explain this informatior resulted in new insight that minimum wage is actually seen as a human right, and an increase in democratic participation can possibly result in more conversations about a minimum wage and he increasing it.

However, it is not always possible to find meaning in data that easily. Unfortunately, we, as univer students, may not be the best people to offer probable explanations to our information. This is see attempts to explain the relationship between Popularity and Minimum Wage. However, it is withi capacity to take this information and spread it to the world, leaving it as an open ended question f discussions to flourish.

That is how we can add value to the world using data.

- Written in collaboration with Erika Medina