Using Stepwise Regression and K Fold Cross Validation to Build a Model that Predicts Unemployment Level of NJ

Abstract

The current project sought to identify a model that best predicted the effect of seven economic factors (Income, Minimum Wage, GDP, Population, Degrees Earned, Labor Force, and Rental Vacancies) on unemployment in New Jersey using data sourced through *FRED*. Analyses yielded two different models: one with (modelnew2) and one without the Population the factor (modelnew1). K fold Cross validation was used to compare the models. Results showed that modelnew1 (Y~GDP+Min.wage+Labor.force) was the best model for predicting unemployment because it had the lowest RMSE (0.384609) and MAE (0.3733439).

Introduction

The current project sought to identify a model that best predicted the effect of economic factors on unemployment in New Jersey. These factors included: Income, Minimum Wage, GDP, Population, Degrees Earned, Labor Force, and Rental Vacancies. When we first started to build the model, we had no specific hypotheses about what we were going to find. Thus, this project was exploratory in nature and is intended to describe a process.

Data from a larger dataset were assessed using scatterplots and regression analyses. Analyses yielded two different models (one with the factor population and one without), which were then compared using K fold cross validation.

Methods

All of the data used in the project come from https://fred.stlouisfed.org/, which is a reliable economic source, which included data collected over an eleven-year period (Jan 2009 to Jan 2019). Specific data collection methods and materials used to collect data are unknown. But for the purposes of this paper, it is assumed that all standards have been met.

Data were cleaned and organized to report changes on the first of each year and scatterplots were then created to visually inspect data. Alpha = .05 was used for the entirety of this project. Scatterplots indicated that unemployment was negatively correlated with education and GDP, respectively. In contrast, unemployment was shown to be positively correlated with rental vacancies. See Figure 1.

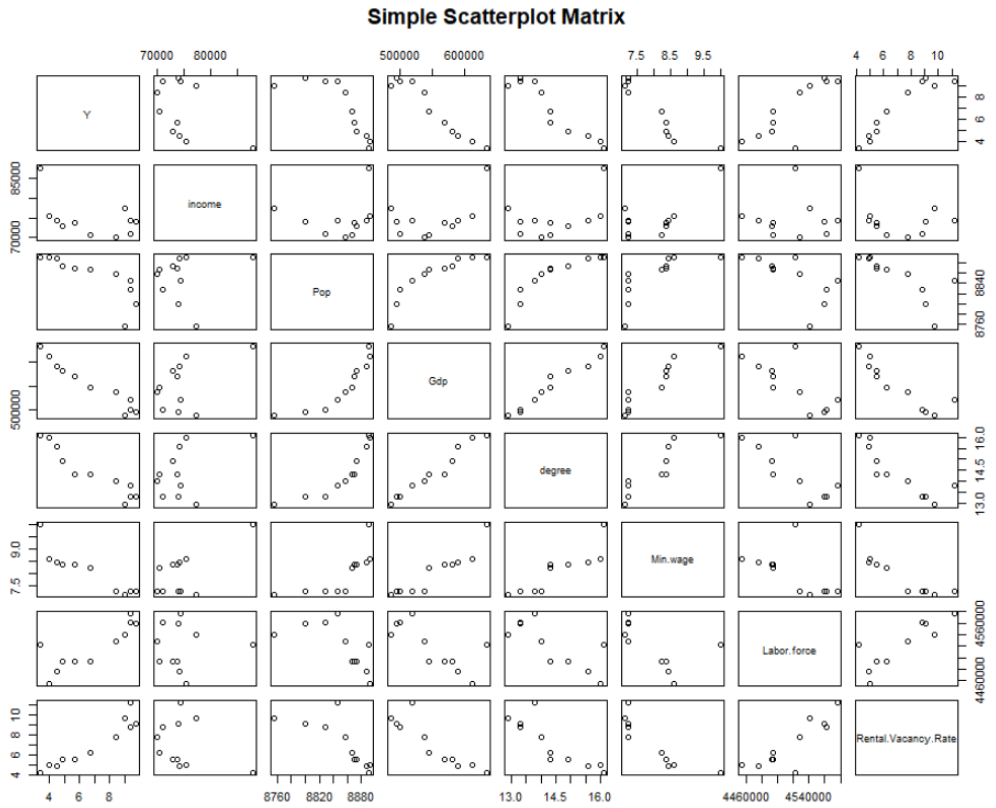

Figure 1: Scatterplot

Additionally, a base model was used to assess all factors together. An F test was then performed on the base model, which showed a significant difference between our model and the intercept-only model. However, there were no significant factors identified from t-tests. The ANOVA of the base model showed that Income, Population, and GDP significantly contributed to unemployment. Furthermore, all possible regressions, both and forward stepwise, and best subset were conducted in order to improve upon the base model, as well as to remove some of the noise created by extraneous factors.

These subsequent findings led us to two models. Modelnew1 was chosen by all four processes and modelnew2 was included for its higher R squared adjusted. Both appeared to be a better fit than the full model. Thus, in order to assess the primary concern of this paper, K fold cross validation was used to examine which of the resulting models better predicted unemployment. Results from the K fold cross validation are presented in the next section.

Results

From the results of the K fold cross validation showed that modelnew1 (Y~GDP+Min.wage+Labor.force) was the best model for predicting unemployment because it had the lowest RMSE (0.5179042) and MAE (0.4808447). *See appendix for output.*

Discussion

Scatterplots revealed that unemployment was negatively correlated with economic conditions that are generally considered “good”, such as level of education. For example, as the number of degrees earned, as well as GDP increased, Unemployment generally trended down. (See Figure 1)

The residuals were not satisfactory because they violated the assumption of normality, but for the purpose of this project they were ignored. In a real-life scenario, we would investigate transformations or try to fit higher order models.

Because analyses were limited to a small dataset, K fold cross validation was appropriate because there was not enough data to split and create a training set and test set. The use of K fold cross validation may also have helped to avoid overfitting. Thus, while modnew2 had a higher R sq adjusted value compared to modnew1, the simpler model is the preferred model. This is supported by findings from Nested f test, which showed that no extra factors were significant beyond the three (GDP, minimum wage, labor force) included in modnew1.

Since only some of the factors included in the dataset go beyond 2019, the economic conditions included in this project do not reflect the potential economic impacts caused by COVID-19. As such, I am curious about how this model would be different in the coming years. Will 2020 and 2021 be outliers or will we learn about a new important factor for predicting unemployment in NJ? It would also be interesting to see how well this model could be used to predict unemployment in other states.

Appendix A

Output

```
> print(modelnew1)
Linear Regression

11 samples
 3 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 9, 9, 8, 9, 9
Resampling results:

  RMSE       Rsquared   MAE
  0.5179042  0.9941842  0.4808447

Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
> print(modelnew2)
Linear Regression

11 samples
 4 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 9, 9, 8, 9, 9
Resampling results:

  RMSE      Rsquared   MAE
  0.625951  0.9934372  0.5285747

Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
Call:
lm.default(formula = Y ~ income + Pop + Gdp + degree + Min.wage +
    Labor.force + Rental.Vacancy.Rate, data = data)

Residuals:
       1        2        3        4        5        6        7        8        9       10
-0.12311  0.28807 -0.19107 -0.13250  0.31459  0.22038 -0.09093 -0.23733 -0.41799  0.27253
      11
 0.09736

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -2.203e+02  1.839e+02  -1.198    0.317
income               1.192e-04  2.165e-04   0.550    0.620
Pop                  2.327e-02  2.842e-02   0.819    0.473
Gdp                 -4.304e-05  3.204e-05  -1.343    0.272
degree              -1.333e-01  7.580e-01  -0.176    0.872
Min.wage            -1.170e+00  1.115e+00  -1.049    0.371
Labor.force          1.051e-05  1.689e-05   0.622    0.578
Rental.Vacancy.Rate -2.457e-02  2.921e-01  -0.084    0.938

Residual standard error: 0.4556 on 3 degrees of freedom
Multiple R-squared:  0.9894,    Adjusted R-squared:  0.9647
F-statistic: 40.06 on 7 and 3 DF,  p-value: 0.005822
```

```
> summary(modnew1lm)

Call:
lm.default(formula = Y ~ Gdp + Min.wage + Labor.force, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.48751 -0.19848  0.03526  0.18627  0.48182

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.727e+01  2.212e+01  -3.041  0.01881 *
Gdp         -1.883e-05  7.368e-06  -2.556  0.03778 *
Min.wage    -1.000e+00  3.505e-01  -2.854  0.02453 *
Labor.force  2.047e-05  4.544e-06   4.506  0.00278 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3594 on 7 degrees of freedom
Multiple R-squared:  0.9846,    Adjusted R-squared:  0.9781
F-statistic: 149.6 on 3 and 7 DF,  p-value: 1.04e-06

> summary(modnew2lm)

Call:
lm.default(formula = Y ~ Gdp + Min.wage + Labor.force + Pop,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.46566 -0.14948 -0.02656  0.23306  0.30905

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.289e+02  5.103e+01  -2.527  0.04488 *
Gdp         -2.830e-05  1.000e-05  -2.830  0.02995 *
Min.wage    -7.945e-01  3.673e-01  -2.163  0.07378 .
Labor.force  1.986e-05  4.341e-06   4.574  0.00379 **
Pop          7.685e-03  5.796e-03   1.326  0.23310
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3413 on 6 degrees of freedom
Multiple R-squared:  0.9881,    Adjusted R-squared:  0.9802
F-statistic: 124.8 on 4 and 6 DF,  p-value: 6.649e-06

> anova(modnew1lm,modnew2lm)
Analysis of Variance Table

Model 1: Y ~ Gdp + Min.wage + Labor.force
Model 2: Y ~ Gdp + Min.wage + Labor.force + Pop
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1      7 0.90397
2      6 0.69911  1   0.20486 1.7582 0.2331
```

```
> ols_step_best_subset(model,print_plot=TRUE)
                        Best Subsets Regression
--------------------------------------------------------------------------------
Model Index    Predictors
--------------------------------------------------------------------------------
     1         Gdp
     2         Min.wage Labor.force
     3         Gdp Min.wage Labor.force
     4         Pop Gdp Min.wage Labor.force
     5         income Pop Gdp Min.wage Labor.force
     6         income Pop Gdp degree Min.wage Labor.force
     7         income Pop Gdp degree Min.wage Labor.force Rental.Vacancy.Rate
--------------------------------------------------------------------------------

                                                Subsets Regression Summary
-----------------------------------------------------------------------------------------------------------------------------------------
Model    R-Square    Adj.        Pred
                     R-Square    R-Square     C(p)        AIC        SBIC        SBC        MSEP       FPE        HSP        APC
-----------------------------------------------------------------------------------------------------------------------------------------
  1        0.9349      0.9277      0.9033    11.4473    25.6103     -7.5471    26.8040     4.7000     0.5029     0.0532     0.0940
  2        0.9703      0.9629      0.9551     3.4178    18.9800    -10.5843    20.5716     2.4511     0.2780     0.0312     0.0520
  3        0.9846      0.9781      0.9638     1.3544    13.7292     -9.9364    15.7187     1.4792     0.1761     0.0215     0.0329
  4        0.9881      0.9802      0.9068     2.3676    12.9023     -5.9234    15.2897     1.3728     0.1695     0.0233     0.0317
  5        0.9893      0.9786       0.884     4.0378    13.7689      0.2649    16.5542     1.5480     0.1949     0.0315     0.0364
  6        0.9894      0.9735      0.8651     6.0071    15.6569      7.5224    18.8400     2.0431     0.2554     0.0520     0.0477
  7        0.9894      0.9647      0.6172     8.0000    17.6310     14.8588    21.2120     3.0574     0.3586     0.1038     0.0670
-----------------------------------------------------------------------------------------------------------------------------------------
AIC: Akaike Information Criteria
 SBIC: Sawa's Bayesian Information Criteria
 SBC: Schwarz Bayesian Criteria
 MSEP: Estimated error of prediction, assuming multivariate normality
 FPE: Final Prediction Error
 HSP: Hocking's Sp
 APC: Amemiya Prediction Criteria
```

20	13	2	Gdp Min.wage	0.9400709	0.9250886	11.986211
19	14	2	Gdp degree	0.9354541	0.9193176	13.294791
9	15	2	income Gdp	0.9353091	0.9191364	13.335887
24	16	2	degree Labor.force	0.9303522	0.9129403	14.740849
27	17	2	Min.wage Rental.Vacancy.Rate	0.9296414	0.9120518	14.942321
23	18	2	degree Min.wage	0.9241995	0.9052494	16.484755
13	19	2	income Rental.Vacancy.Rate	0.9206388	0.9007985	17.494010
12	20	2	income Labor.force	0.9071390	0.8839238	21.320360
18	21	2	Pop Rental.Vacancy.Rate	0.8997015	0.8746269	23.428434
28	22	2	Labor.force Rental.Vacancy.Rate	0.8923030	0.8653787	25.525471
11	23	2	income Min.wage	0.8908495	0.8635619	25.937436
15	24	2	Pop degree	0.8898935	0.8623669	26.208408
10	25	2	income degree	0.8886060	0.8607575	26.573327
16	26	2	Pop Min.wage	0.8794760	0.8493450	29.161119
17	27	2	Pop Labor.force	0.8032712	0.7540890	50.760486
8	28	2	income Pop	0.7731621	0.7164526	59.294571
57	29	3	Gdp Min.wage Labor.force	0.9846373	0.9780534	1.354365
60	30	3	degree Min.wage Labor.force	0.9807096	0.9724424	2.467626
36	31	3	income Gdp Labor.force	0.9801840	0.9716915	2.616603
46	32	3	Pop Gdp Labor.force	0.9788548	0.9697926	2.993357
47	33	3	Pop Gdp Rental.Vacancy.Rate	0.9763638	0.9662340	3.699409
32	34	3	income Pop Labor.force	0.9759826	0.9656894	3.807459
59	35	3	Gdp Labor.force Rental.Vacancy.Rate	0.9736383	0.9623404	4.471930
51	36	3	Pop Min.wage Labor.force	0.9722576	0.9603681	4.863247
41	37	3	income Min.wage Labor.force	0.9717412	0.9596303	5.009615
63	38	3	Min.wage Labor.force Rental.Vacancy.Rate	0.9704622	0.9578032	5.372141
55	39	3	Gdp degree Labor.force	0.9685665	0.9550950	5.909450
29	40	3	income Pop Gdp	0.9682134	0.9545905	6.009553
37	41	3	income Gdp Rental.Vacancy.Rate	0.9682106	0.9545866	6.010329
58	42	3	Gdp Min.wage Rental.Vacancy.Rate	0.9670580	0.9529401	6.337015
56	43	3	Gdp degree Rental.Vacancy.Rate	0.9666672	0.9523817	6.447803
50	44	3	Pop degree Rental.Vacancy.Rate	0.9642117	0.9488739	7.143768
40	45	3	income degree Rental.Vacancy.Rate	0.9638427	0.9483467	7.248368
61	46	3	degree Min.wage Rental.Vacancy.Rate	0.9632299	0.9474713	7.422049
62	47	3	degree Labor.force Rental.Vacancy.Rate	0.9604696	0.9435280	8.204425
39	48	3	income degree Labor.force	0.9604202	0.9434574	8.218432
43	49	3	income Labor.force Rental.Vacancy.Rate	0.9523708	0.9319583	10.499938
35	50	3	income Gdp Min.wage	0.9474318	0.9249026	11.899845
45	51	3	Pop Gdp Min.wage	0.9466888	0.9238412	12.110434
44	52	3	Pop Gdp degree	0.9458257	0.9226082	12.355061
33	53	3	income Pop Rental.Vacancy.Rate	0.9409557	0.9156510	13.735427
54	54	3	Gdp degree Min.wage	0.9400744	0.9143920	13.985222
38	55	3	income degree Min.wage	0.9386685	0.9123835	14.383707
34	56	3	income Gdp degree	0.9359456	0.9084937	15.155484
52	57	3	Pop Min.wage Rental.Vacancy.Rate	0.9342172	0.9060246	15.645366
49	58	3	Pop degree Labor.force	0.9332310	0.9046158	15.924885
42	59	3	income Min.wage Rental.Vacancy.Rate	0.9297297	0.8996138	16.917305
48	60	3	Pop degree Min.wage	0.9244700	0.8921001	18.408088
53	61	3	Pop Labor.force Rental.Vacancy.Rate	0.9030574	0.8615106	24.477246
31	62	3	income Pop Min.wage	0.8922700	0.8461000	27.524806

Final Model Output

Model Summary

R	0.992	RMSE	0.359
R-Squared	0.985	Coef. Var	5.264
Adj. R-Squared	0.978	MSE	0.129
Pred R-Squared	0.964	MAE	0.231

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	57.938	3	19.313	149.55	0.0000
Residual	0.904	7	0.129		
Total	58.842	10			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-67.268	22.117		-3.041	0.019	-119.566	-14.969
Gdp	0.000	0.000	-0.388	-2.556	0.038	0.000	0.000
Labor.force	0.000	0.000	0.328	4.506	0.003	0.000	0.000
Min.wage	-1.000	0.350	-0.366	-2.854	0.025	-1.829	-0.172

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Gdp	addition	0.935	0.928	11.4470	25.6103	0.6523
2	Labor.force	addition	0.967	0.958	4.4230	20.2204	0.4945
3	Min.wage	addition	0.985	0.978	1.3540	13.7292	0.3594

> ols_step_forward_p(model)

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Gdp	0.9349	0.9277	11.4473	25.6103	0.6523
2	Labor.force	0.9668	0.9584	4.4225	20.2204	0.4945
3	Min.wage	0.9846	0.9781	1.3544	13.7292	0.3594
4	Pop	0.9881	0.9802	2.3676	12.9023	0.3413