Investigation 4: Forecasting Nonfarm Employment

Angel Sarmiento

3/24/2020

Introduction

This investigation is a continuation of the previous one focused on model selection given multiple criteria. In this investigation, forecasting of nonfarm employment will be done after training the model of subsets of the data and finding the most important variables. The four models from the last investigation are used here for this purpose and compared with their respective out-of-sample RMSEs. After comparing these models, the predictions for actual nonfarm employment will be found, to demonstrate the predictive powers of the models.

After model selection, naturally the next step is to generate point and interval forecasts of future data that is not found in the original data.

Model Selection

From the last investigation, four ARDL models were created with the intention of having the best possible fit to the data. Now, they will be repurposed for prediction. The four different models are as follows:

$$\Delta y_t = \beta_0 + \sum_{(a,l)=0}^{12} \beta_a L_l \Delta y_{t-1} + \sum_{b,k}^{12} \beta_b L_k \Delta X_{lf,t} + \sum_{c,k}^{12} \beta_c L_k \Delta X_{bp,t} + \sum_{d,k}^{12} \beta_d L_k \Delta X_{epr,t} + \beta_e X_m + DATE + \varepsilon_t \quad (1)$$

Where $DATE$ is a time trend, $k = 0, 1, 2, 3, ...12$, $l = 1, 2, 3, ...12$, m is the month from $1, 2, 3, ...12$, and L is the lag.

$$\Delta y_t = \beta_0 + \sum_{(a,l)=0}^{12} \beta_a L_l \Delta y_{t-1} + \sum_{b,k}^{2} \beta_b L_k \Delta X_{lf,t} + \sum_{c,k}^{2} \beta_c L_k \Delta X_{bp,t} + \sum_{d,k}^{2} \beta_d L_k \Delta X_{epr,t} + \beta_e X_m + DATE + \varepsilon_t \quad (2)$$

Where $DATE$ is a time trend, $k = 0, 1, 2$, $l = 1, 2, 3, ...12$, m is the month from $1, 2, 3, ...12$, and L_k is the lag at value k or l.

$$\Delta y_t = \beta_0 + \sum_{(a,l)=0}^{12} \beta_a L_l \Delta y_{t-1} + \sum_{b,k}^{2,12} \beta_b L_k \Delta X_{lf,t} + \sum_{c,k}^{2,12} \beta_c L_k \Delta X_{bp,t} + \sum_{d,k}^{2,12} \beta_d L_k \Delta X_{epr,t} + \beta_e X_m + DATE + \varepsilon_t \quad (3)$$

Where $DATE$ is a time trend, $k = 0, 1, 2 \; or \; 12$, $l = 1, 2, 3, ...12$, m is the month from $1, 2, 3, ...12$, and L_k is the lag at value k or l.

$$\Delta y_t = \beta_0 + \sum_{(a,l)=0}^{12,24} \beta_a L_l \Delta y_{t-1} + \sum_{b,k}^{2,12,24} \beta_b L_k \Delta X_{lf,t} + \sum_{c,k}^{2,12,24} \beta_c L_k \Delta X_{bp,t} + \sum_{d,k}^{2,12,24} \beta_d L_k \Delta X_{epr,t} + \beta_e X_m + \varepsilon_t \quad (4)$$

	RMSE	Rsquared	MAE	AIC	BIC	k-fold
Model 1	0.0048399	0.7703876	0.0036540	-2901.269	-2830.777	0.0043540
Model 2	0.0048402	0.7703781	0.0036289	-2902.259	-2831.768	0.0041781
Model 3	0.0048428	0.7701763	0.0036383	-2901.216	-2818.976	0.0042702
Model 4	0.0048877	0.7685151	0.0036598	-2800.082	-2706.882	0.0042208

Table 1. Model Comparison for Nonfarm employment using LOOCV

From these results, it was concluded that model 2 was the best model due to its relative parsimony and good performance in comparison with the other models. Model 2 explained a great amount of the variance while having a low RMSE as well as AIC and BIC. The plots of these models' performances were then shown as below.

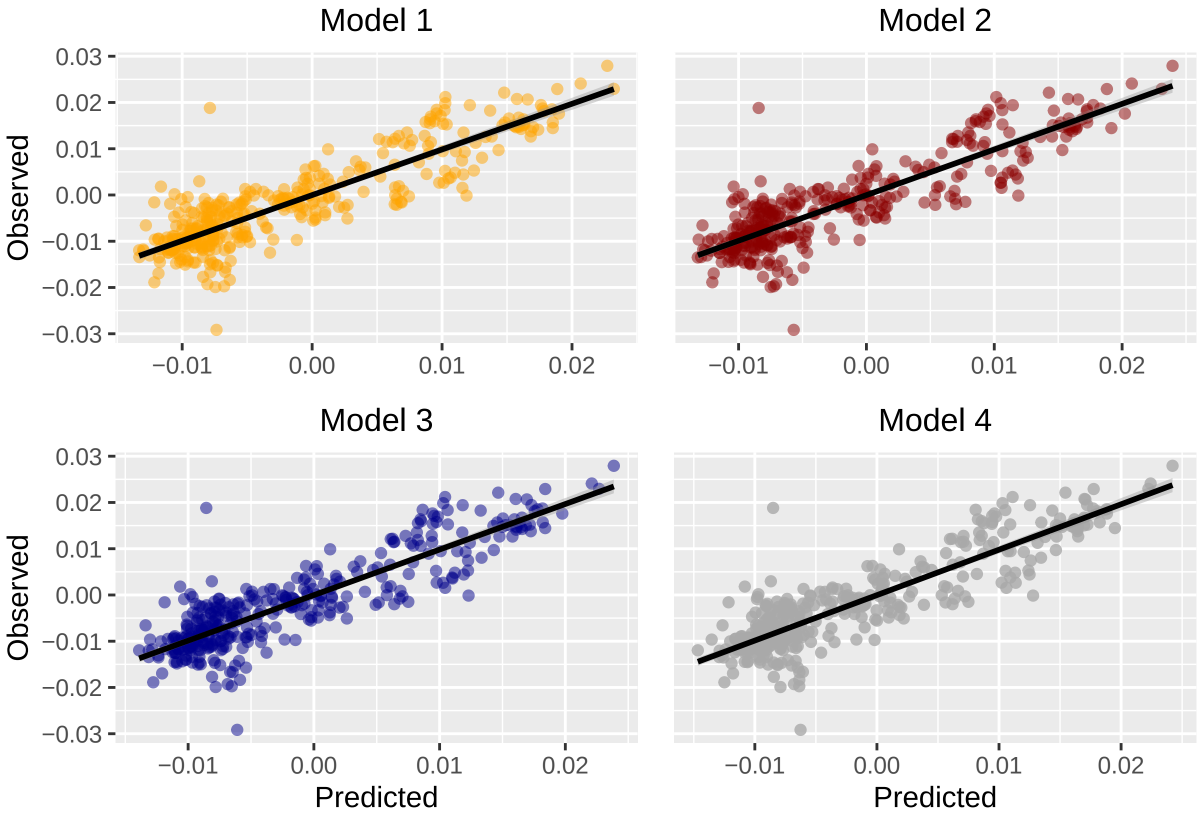

Figure 4. All four models plotted in comparison with one another.

Predicting Nonfarm Employment in 2019

In order to see the true predictive power of the models, they are going to be evaluated on data that they are not trained on. This approach is known as the train-validation set approach. By creating this data partition, only the values up to the last year (2019) will be included to then be evaluated on the test data from 2019. All of the predictors in these models are centered on a mean of zero and scaled. This

	RMSE	Rsquared	AIC	BIC	k-fold	OOS RMSE	num of vars
Model 1	0.0048399	0.7703876	-2901.269	-2830.777	0.0043540	9.101140	56
Model 2	0.0048402	0.7703781	-2902.259	-2831.768	0.0041781	9.100579	26
Model 3	0.0048428	0.7701763	-2901.216	-2818.976	0.0042702	9.100818	29
Model 4	0.0048877	0.7685151	-2800.082	-2706.882	0.0042208	9.101292	24

With a better score in every single metric, it looks like model 2 is still the best performing model of the bunch. It is also relatively parsimonious while explaining most of the variance in nonfarm employment. To develop this model further, transformations will be performed to show the actual level of nonfarm employment predictions. Note that for this model's approximations, normality will be approximately assumed.