Macroeconomic models

Matt Nuttall

30 October 2018

Below are the loadings for the principal components for the economic, commodity price, and producer price variables. The first two principal components for each set of variables account for over 95% of the total variance, and so there is a fairly strong justification for dropping the other PC’s. Below I have only reported the first two PC’s.

The first PC for the economic variables shows the strong positive covariance between all variables, excluding agricultural proportion of GDP (agr\_gdp), which is negatively correlated with all the others. Luc suggested that in this case I could select i) one of the variables with negative loadings to represent the others, in this case probably foreign direct investment (fdi), as it has the strongest loading, and then ii) agricultural proportion of GDP (agr\_gdp) as this will reflect the contrasting relationship. For PC2 I could select industrial proportion of GDP (ind\_gdp) because it has a strong contrasting loading compared to the others, and normally I would then select agr\_gdp to represent the positive loadings but it has already been selected for PC1.

Economic variable principal component loadings

|  |  |  |
| --- | --- | --- |
| econvar | PC1 | PC2 |
| gdp | -0.422 | 0.270 |
| fdi | -0.446 | 0.232 |
| ind\_gdp | -0.415 | -0.522 |
| agr\_gdp | 0.394 | 0.600 |
| dev\_agri | -0.359 | 0.461 |
| dev\_env | -0.407 | 0.170 |

For the commodity price variables, again the first PC shows that all of the variables contribute equally to the PC. Therefore I can just select one of the variables to be used as a proxy. In this case I would select Agricultural Raw Materials Index (armi) as it has the strongest loading. For PC2, two variables contribute positively and 3 contribute negatively. I could select one positive and one negative variable, in this case median rice price (rice\_med) and median sugar price (sug\_med), as they have the strongest loadings.

Commodity variable principal component loadings

|  |  |  |
| --- | --- | --- |
| comvars | PC1 | PC2 |
| armi | 0.461 | -0.260 |
| rice\_med | 0.433 | 0.660 |
| rub\_med | 0.451 | -0.350 |
| corn\_med | 0.443 | 0.430 |
| sug\_med | 0.450 | -0.441 |

For the producer price variables, the first PC could be represented by sugar producer price (prod\_sug) as all but one are positively contributing to the PC and so this variable would be an effective proxy. Cassava producer price (prod\_cass) could be selected to represent PC2 for the same reason.

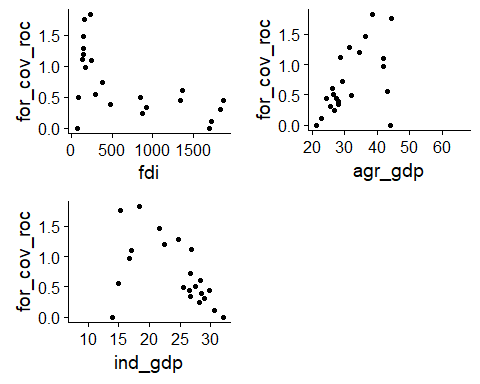
Producer price variable principal component loadings

|  |  |  |
| --- | --- | --- |
| prodvars | PC1 | PC2 |
| prod\_rice | 0.504 | -0.13 |
| prod\_rub | 0.467 | -0.13 |
| prod\_cass | -0.210 | -0.10 |
| prod\_corn | 0.480 | -0.15 |
| prod\_sug | 0.510 | -0.02 |

Below are scatter plots of the selected raw variables with forest cover rate of change

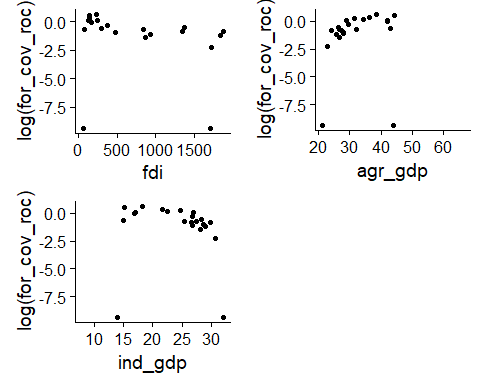
Economic variables:

## Warning: Removed 1 rows containing missing values (geom\_point).  
  
## Warning: Removed 1 rows containing missing values (geom\_point).  
  
## Warning: Removed 1 rows containing missing values (geom\_point).

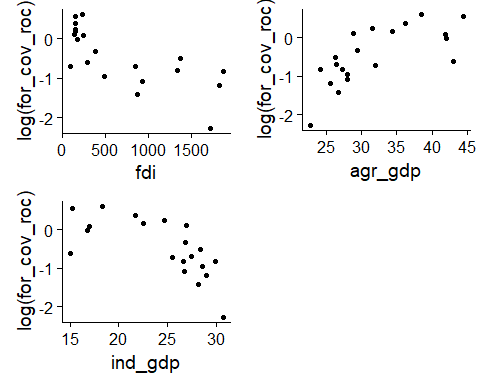


When we log-transform the response variable we get:

## Warning: Removed 1 rows containing missing values (geom\_point).  
  
## Warning: Removed 1 rows containing missing values (geom\_point).  
  
## Warning: Removed 1 rows containing missing values (geom\_point).



As we can see there are two data points causing problems - the forest cover rate of change values from 1994 and 2015. I’m not sure what is going on here, as the values are both identical (which in itself seems odd). If we remove those two outliers we get:



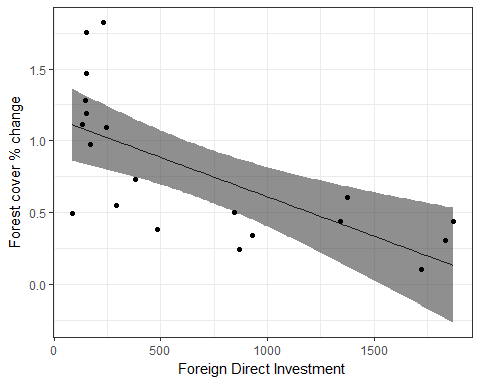
Although I don’t necessarily think that those relationships look linear, Luc thought that a linear model would definitely be worth exploring, and then comparing to a more complicated model.

But his advice is that using principal components in the models is conceptually quite complicated, both for me and for the reader. Furthermore he thinks that making predictions using the models will be much more challenging if we use PC’s, rather than raw predictor variables.

I have done a quick bit of modelling as a working example.

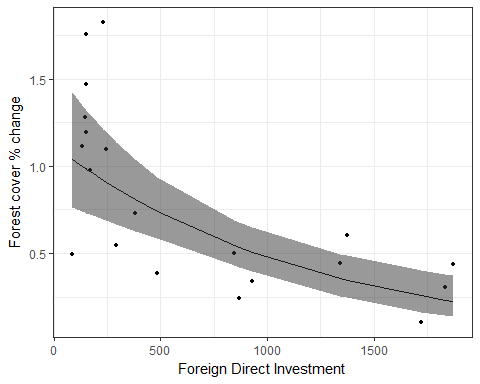
### Linear model of forest cover rate of change ~ foreign direct investment:

##   
## Call:  
## lm(formula = for\_cov\_roc ~ fdi, data = dat\_sub)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.62091 -0.24385 0.02207 0.19727 0.79284   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.1589948 0.1279971 9.055 4.02e-08 \*\*\*  
## fdi -0.0005474 0.0001406 -3.892 0.00107 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3876 on 18 degrees of freedom  
## Multiple R-squared: 0.457, Adjusted R-squared: 0.4268   
## F-statistic: 15.15 on 1 and 18 DF, p-value: 0.001068



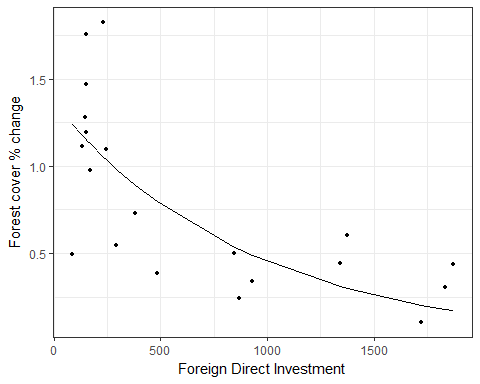
### Linear model with log-transformed response

##   
## Call:  
## lm(formula = log(for\_cov\_roc) ~ fdi, data = dat\_sub)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.9105 -0.4131 0.1477 0.3050 0.6882   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.1111928 0.1705803 0.652 0.522735   
## fdi -0.0008523 0.0001874 -4.548 0.000249 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5166 on 18 degrees of freedom  
## Multiple R-squared: 0.5347, Adjusted R-squared: 0.5088   
## F-statistic: 20.68 on 1 and 18 DF, p-value: 0.0002493



### 2 parameter exponential model

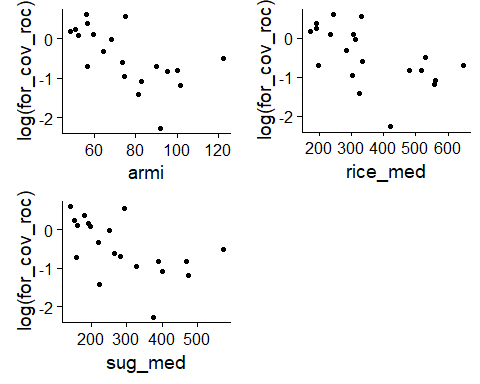
##   
## Formula: for\_cov\_roc ~ a \* exp(-b \* fdi)  
##   
## Parameters:  
## Estimate Std. Error t value Pr(>|t|)   
## a 1.3580036 0.1793857 7.570 5.33e-07 \*\*\*  
## b 0.0011016 0.0003683 2.991 0.00784 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3623 on 18 degrees of freedom  
##   
## Number of iterations to convergence: 9   
## Achieved convergence tolerance: 2.433e-06



The log linear model performs better than the linear model, but the exponential model performs better than the log linear. I used AIC to compare the exponential model and log linear (anova doens’t work as the response variable is different), but I’m not sure if AIC is an appropriate test.

Below are the scatter plots for the commodity price and producer price variables (with the two outliers removed), and log-transformed response.

Commodity price variables:



Producer price variables:

