Measurement, Scaling, & Dimensional Analysis
Prof. Adam Enders
ICPSR, 2019

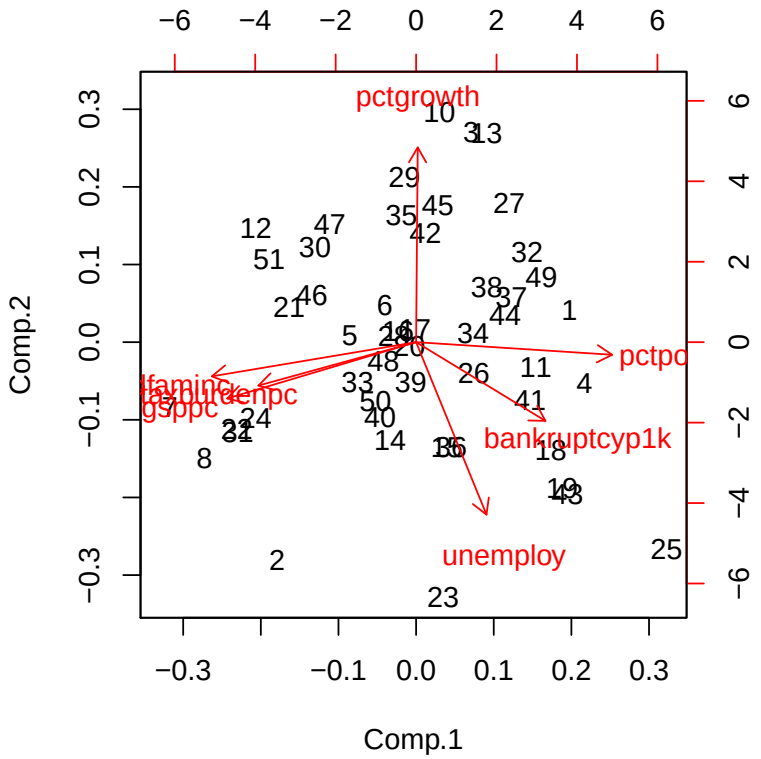
Assignment 3:
Biplots and Principal Components Analysis

First, I subset the dataset since we don't require all of the variables, some of which are identifiers, others of which we won't need until we run the regression models. Then, I standardized the variables to a mean of 0 and variance of 1 (actually, I do both in one step).

```
> statecon2 <- scale(statecon[-9,4:10])
```

Next, I performed a principal components analysis. I could have done a singular value decomposition on the standardized data and then constructed a biplot "by hand." Instead, I opted to use the "biplot()" function to create it for me. The resultant biplot is depicted below.

```
> biplot(princomp(statecon2))
```



The variable vectors can be interpreted by considering their relative length and direction. A longer vector is one for which the R^2 between the fitted and observed values of the variable are high. That is, the length of the variable vectors is proportional to the R^2 between the fitted and observed values. It appears that `pcgtrowth` and `unemploy` are the longer vectors, but not by much.

There are, however, clear differences in the way the variable vectors relate to each other and the principal components (dimensions). We see that `medfaminc`, `taxburdenpc`, and `gsppc` are all similar (highly correlated) measures of state economy. They are also very highly related to the first dimension. We can also see that `pctpov` and `bankruptcyp1k` are fairly similar (although not as similar as the previous three mentioned), and strongly negatively correlated with `medfaminc`, `taxburdenpc`, and `gsppc`, but positively correlated with the first dimension.

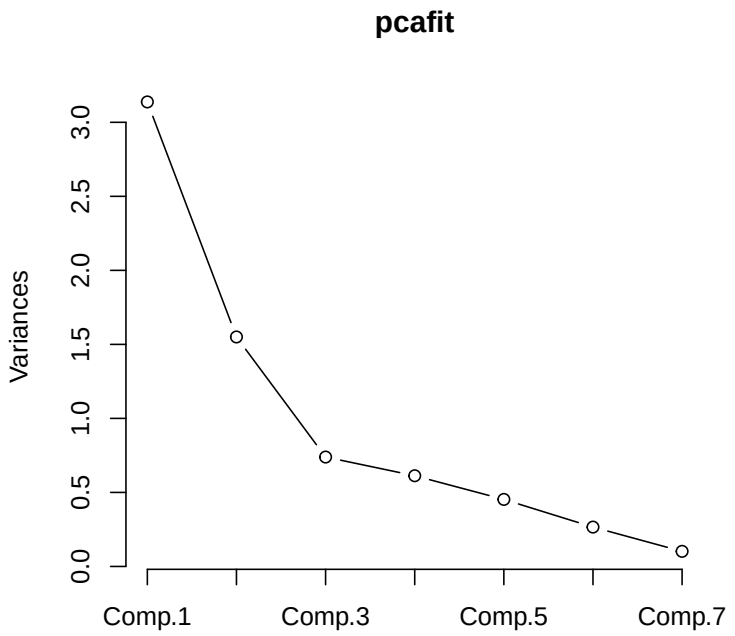
Finally, we see that `pcgtrowth` and `unemply` are more strongly correlated with the second dimension, though in opposite directions. This simply means that `pcgtrowth` will have a strong positive loading, and `unemply` will have a strong negative loading. This also suggests that an increase in economic growth corresponds to a decrease in unemployment – exactly as we'd expect.

Next, I attempted to "assess" dimensionality. First, I looked at the proportion of variance explained by each dimension. The principal components analysis reveals that the first component captures approximately 46% of the variance in the seven economy variables. The second component captures an additional 23% of the variation (for a cumulative total of 68%). This could be better, but the increase in variance explained from component 2 to 3 is fairly small (10%), leading me to think that a two dimensional solution might be "best."

```
> pcafit <- princomp(statecon2)
> summary(pcafit, loadings = TRUE, cutoff = 0)

Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
Standard deviation     1.7713625 1.2450462 0.8596355 0.7827561 0.6728469
Proportion of Variance 0.4573944 0.2259679 0.1077220 0.0893159 0.0659946
Cumulative Proportion  0.4573944 0.6833623 0.7910843 0.8804002 0.9463948
                           Comp.6     Comp.7
Standard deviation     0.51555951 0.31926460
Proportion of Variance 0.03874659 0.01485858
Cumulative Proportion  0.98514142 1.00000000
```

Second, I examined a scree plot of the variances of the components against the associated components. I see a reasonably clear "elbow" at the third component, which suggests that I should proceed with a two-dimensional solution, or the first two principal components. This comports with my inspection of the proportions of variance explained by each



component, so I feel comfortable in substantively interpreting the biplot and using the first two components in subsequent statistical analyses in lieu of the original dataset. Examination of the component loadings provides information very similar to that contained in the biplot above. The `pctgrowth` and `unemploy` variables are extremely weakly correlated with the first dimension (component), though the others have non-trivial loadings. Just the opposite is the case for the second component, where the `pctgrowth` and `unemploy` variables load very highly, where the others are only weakly related to the second component (next highest loading is -0.278 for `bankruptcyp1k`).

```
Loadings:
              Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
bankruptcyp1k  0.321 -0.278  0.714 -0.393  0.363 -0.156 -0.015
pctgrowth      0.004  0.688  0.078  0.301  0.629 -0.162  0.090
gsppc         -0.469 -0.199  0.163  0.066  0.343  0.756  0.145
medfaminc     -0.507 -0.121  0.228  0.245  0.021 -0.285 -0.731
pctpov         0.487 -0.044 -0.387 -0.097  0.319  0.355 -0.612
taxburdenpc   -0.391 -0.153 -0.470 -0.568  0.399 -0.333  0.102
unemploy       0.175 -0.609 -0.190  0.599  0.305 -0.245  0.225
```

Finally, I turned toward using the principal components in a multiple regression model predicting presidential approval. First, I estimated a "naive" model with all of the original variables ("Model 1"). I observed statistically significant effects for `gsppc`, `medfamic`, and

`taxburdenpc`, all of which were highly correlated with each other and the first principal component according to the biplot and the results of the PCA. This might suggest that the dimension those variables are related to has something to do with presidential approval. Yet, this is a messy model with too many economic indicators, many of which we probably don't have a particular theoretical expectation for (at least, not one that is unique to each distinct variable).

Next, I estimated a model with all of the components, instead of the original variables. The results of this regression appear in the second column (called "Model 2") of the table. In practice, we would never do this, since the purpose of PCA is to reduce dimensionality. By including all of the components, we're including the same number of variables as in the first "naive" model, and explaining exactly the same amount of variance (see the R^2 values for Model 1 and Model 2). Regardless, only the first, sixth, and seventh components have a statistically significant effect on presidential approval. Since the sixth and seventh components capture very little of the variance (about 5%) in the seven state economy variables, I decided to drop them (and the other statistically insignificant components) from the final model ("Model 3" in the table).

Although the R^2 decreased by about 0.11 from Model 2 to Model 3 (and the change in the adjusted R^2 is a bit less at 0.08), the second model is much more parsimonious – a goal of all scientific research. Furthermore, when I attempted to add the second and third components to the model, nested model tests (model-comparison F-tests) showed that they really didn't help explain a significant amount of the variation in presidential approval. In terms of substantive interpretation, we can say that partisanship and one measure of the variance in the set of the seven state economy variables (the first principal component) accounts for approximately 66% of the variance in presidential approval.

	Model 1	Model 2	Model 3
(Intercept)	24.07*	59.66*	59.30*
	(10.68)	(1.46)	(1.56)
as.factor(partisan)LD	1.06	1.06	−0.31
	(2.04)	(2.04)	(2.11)
as.factor(partisan)LR	0.36	0.36	0.34
	(4.58)	(4.58)	(4.51)
as.factor(partisan)SD	3.88*	3.88*	4.93*
	(1.75)	(1.75)	(1.81)
as.factor(partisan)SR	−11.28*	−11.28*	−11.70*
	(2.43)	(2.43)	(2.69)
bankruptcyp1k	−0.02		
	(3.54)		
pctgrowth	0.35		
	(0.32)		
gsppc	−0.00*		
	(0.00)		
medfaminc	0.00*		
	(0.00)		
pctpov	0.30		
	(0.37)		
taxburdenpc	0.00*		
	(0.00)		
unemploy	0.46		
	(0.72)		
Comp.1		−1.65*	−1.47*
		(0.35)	(0.39)
Comp.2		−0.30	
		(0.45)	
Comp.3		−0.53	
		(0.69)	
Comp.4		0.43	
		(0.73)	
Comp.5		1.07	
		(0.79)	
Comp.6		−3.48*	
		(1.08)	
Comp.7		−4.09*	
		(1.80)	
R^2	0.77	0.77	0.66
adj. R^2	0.71	0.71	0.63
n	50	50	50

Standard errors in parentheses, * indicates $p < 0.05$