Investigating the Drivers of Western Monarch Decline Using Partial Least Squares Regression

By Jackie Luu
Virginia Commonwealth University
ENVS 603 Spring 2023

LESSON

This lesson aims to explore the drivers of the western monarch butterfly decline. Students will compare the effects of climate factors and land use factors on the western monarch butterfly population. Students will also be introduced to partial least squares regression (PLSR), and then apply PLSR to address the monarch question.

WESTERN MONARCH BUTTERFLY DECLINE

Students will learn about differences between the western and eastern monarch butterfly populations. Students will explore drivers of the monarch decline based on conclusions from prior literature.

PARTIAL LEAST SQUARES REGRESSION

Using data from a prior study, students will investigate the relationship between climate and land use factors and the western monarch butterfly population. Students will create a PLSR model and biplot and explore the strengths and weaknesses of PLSR.

DRIVERS OF SPECIES DECLINE

Many species worldwide are experiencing a decline in their population. This decline is due to various factors such as habitat degradation, overharvesting, the introduction of invasive species, pollution, and climate change (Hogue & Breon, 2022). It is important for conservationists to understand the factors affecting a species' population, since this information can be used to assess the severity of the threats harming a species. For example, this information is used for listing decisions under the Endangered Species Act (Smith et al., 2018). When the factors endangering a species can be resolved without government intervention, sometimes a species will not be listed even if their population is declining (Crone et al. 2019).

However, it can be difficult to isolate the main factor(s) driving the population decline. Typically, numerous factors affecting a species' population decline occur simultaneously. How can scientists determine the main drivers of a species' decline when global warming, habitat destruction, overhunting, etc. are all happening at the same time? Scientists are particularly concerned about the simultaneous impacts from both climate and land use factors (Brook et al., 2008).

MONARCH BUTTERFLIES

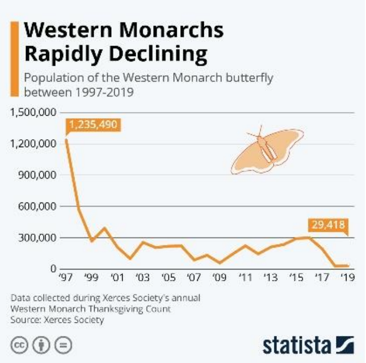
In July 2022, the migratory monarch butterfly (*Danaus plexippus plexippus*) was listed as endangered on the IUCN Red List with climate change and habitat destruction cited as the major threats to this species (*Migratory Monarch Butterfly Now Endangered - IUCN Red List*, n.d.).

The monarch butterfly has a western and eastern migratory population. The western population spends the winter in coastal California and breeds west of the Rocky Mountains (Crone et al., 2019). The eastern population spends the winter in central Mexico and breeds east of the Rocky Mountains (Crone et al., 2019).

> *“Today’s Red List update highlights the fragility of nature’s wonders, such as the unique spectacle of monarch butterflies migrating across thousands of kilometers.”*
> *-Dr. Bruno Oberle, IUCN Director General*

DRIVERS OF MONARCH DECLINE

Previous studies have evaluated the effects of various climate and land use factors on the eastern monarch population (Thogmartin et al., 2017). Studies have noted land use changes as a significant threat to the eastern monarch due to the loss of milkweed (Stenoien et al., 2018).


Western Monarchs Rapidly Declining
Population of the Western Monarch butterfly between 1997-2019

Data collected during Xerces Society's annual Western Monarch Thanksgiving Count
Source: Xerces Society

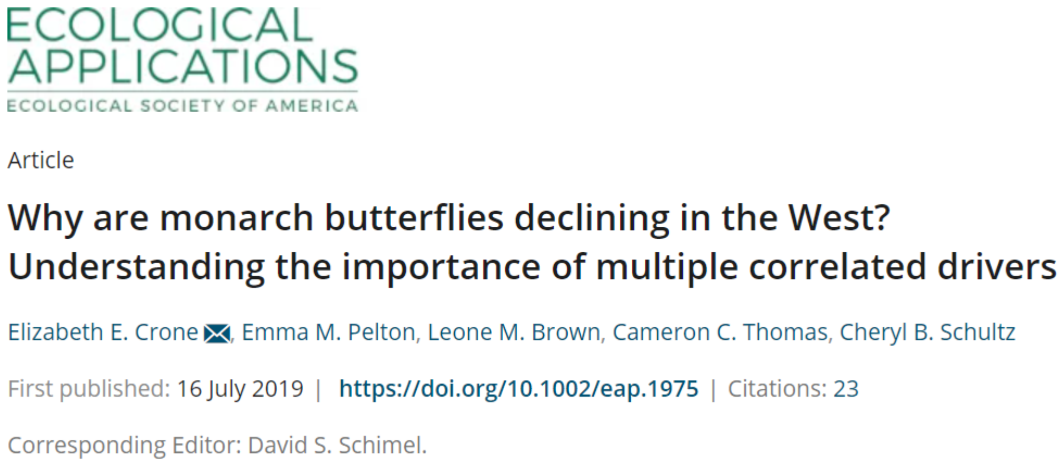
Milkweed is the sole host plant for monarch caterpillars, meaning that monarch caterpillars can only eat milkweed leaves (*Milkweed for Monarchs*, n.d.). Given this fact, milkweed plants are necessary for the survival of monarch populations.

Land use factors, such as the use of herbicides, have been attributed to a decline in milkweed (Stenoien et al., 2018). However, studies prior to the focal paper did not explore land use impacts on the western monarch population decline (Crone et al., 2019).

FOCAL PAPER

> *“Considering the current global biodiversity crisis, it is critical to uncover what is happening with diverse and functionally important species such as [butterflies] before it’s too late.”*
> *-Sophie Ledger, Zoological Society of London Researcher*

Students will be guided through a lesson focused on the study questions, data, and analyses from the Crone et al., 2019 paper.


ECOLOGICAL APPLICATIONS
ECOLOGICAL SOCIETY OF AMERICA

Article

Why are monarch butterflies declining in the West? Understanding the importance of multiple correlated drivers

Elizabeth E. Crone, Emma M. Pelton, Leone M. Brown, Cameron C. Thomas, Cheryl B. Schultz

First published: 16 July 2019 | https://doi.org/10.1002/eap.1975 | Citations: 23

Corresponding Editor: David S. Schimel.

Before proceeding with the rest of this lesson, please read the introduction and methods sections of the focal paper to get familiar with the study objectives and the methods employed.

STUDY OBJECTIVES

The authors explore the western monarch population decline by investigating both climate and land use factors. The authors also emphasize how to analyze data with multiple correlated predictors using partial least squares regression.

THE DATA

The lesson will focus on the data used for the a priori PLSR in the focal paper. The a priori PLSR included eight predictors.

CLIMATE PREDICTORS	LAND USE PREDICTORS	OTHER PREDICTORS
Average winter minimum temperature in overwintering habitat	Proportion of coastal development near overwintering habitat	Monarch abundance in the previous year
Winter precipitation in the overwintering habitat	Glyphosate use	
Average summer maximum temperature in the breeding habitat	Neonicotinoid use	
Summer drought severity in the breeding habitat		

The focal paper data was downloaded from Dryad. The original dataset included 29 environmental predictors for the kitchen sink PLSR. For this lesson, students will only be conducting the a priori PLSR. Therefore, unnecessary columns were removed from the original dataset, and some column names were renamed for simplicity.

PARTIAL LEAST SQUARES REGRESSION

Before working with the monarch data, we will go over the basics of partial least squares regression and work through a simple example.

PLSR is related to ordinary least squares regression which is the standard method used to fit a linear regression. Ordinary least squares regression only works when the assumptions of a linear regression are met. In situations where there are multiple correlated predictors, ordinary least square regression should be replaced with another method such as PLSR (Korstanje, 2021).

Want to help monarchs?

Join the western monarch Thanksgiving and New Year's count: http://www.westernmonarchcount.org/

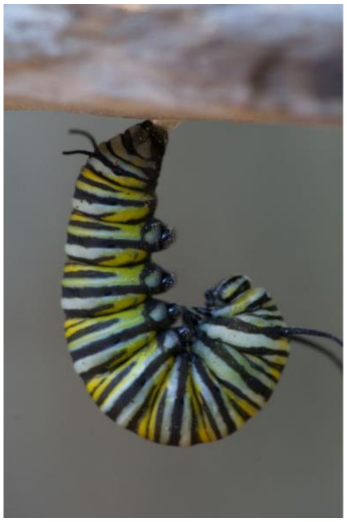
Submit sightings of western monarchs and milkweed to help create maps: http://www.monarchmilkweedmapper.org/

Student tagging a monarch.

PLSR is a multivariate analysis that combines characteristics from multiple regression and principal component analysis. PLSR is useful when there are numerous correlated predictors (Abdi, n.d.). PLSR reduces the dimensionality of correlated variables to form uncorrelated components that represent the shared, underlying information of the variables (Korstanje, 2021).

The number of components is typically decided from the amount of residual variation (Tobias, n.d.). The components are then used for fitting a linear regression. Unlike other techniques, the components explain both the predictor and dependent variables since it is based on covariance, rather than variance (Bobbitt, 2020; *Partial Least Squares Regression (PLS)*, n.d.).

CORRELATED PREDICTORS



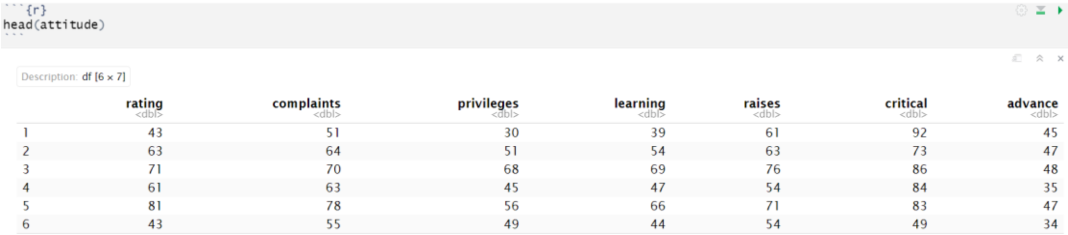
So why should we use PLSR? As noted above, PLSR takes multiple correlated predictors and uses them to form uncorrelated PLS components. One of the main arguments of the focal paper is that PLSR is a better technique than multiple regression when analyzing data with correlated predictors.

To understand why the ability to form uncorrelated components is important, first we must explore why multicollinearity is a problem. A predictor that is highly correlated with other predictors will likely have a large standard error. This large standard error implies that there is a lot of uncertainty in the regression coefficient, and this weakens the statistical significance of the predictors. For example, if two predictors are highly correlated to each other, neither of these variables are likely to be statistically significant, even if they are both highly correlated with the dependent variable (Allen, 1997).

Now that we have a better understanding of why PLSR is used, we can try fitting a PLS model with a simple example.

PLSR EXAMPLE

For this example, we will use the pre-built "attitude" dataset in R. Let’s take a quick look at the dataset.



~~~
```{r}
head(attitude)
```
~~~

Description: df [6 x 7]

	rating <dbl>	complaints <dbl>	privileges <dbl>	learning <dbl>	raises <dbl>	critical <dbl>	advance <dbl>
1	43	51	30	39	61	92	45
2	63	64	51	54	63	73	47
3	71	70	68	69	76	86	48
4	61	63	45	47	54	84	35
5	81	78	56	66	71	83	47
6	43	55	49	44	54	49	34

This dataset shows how employees rated the departments in their organization. Each of the columns lists the proportion of favorable responses.

Here are brief descriptions of the columns:

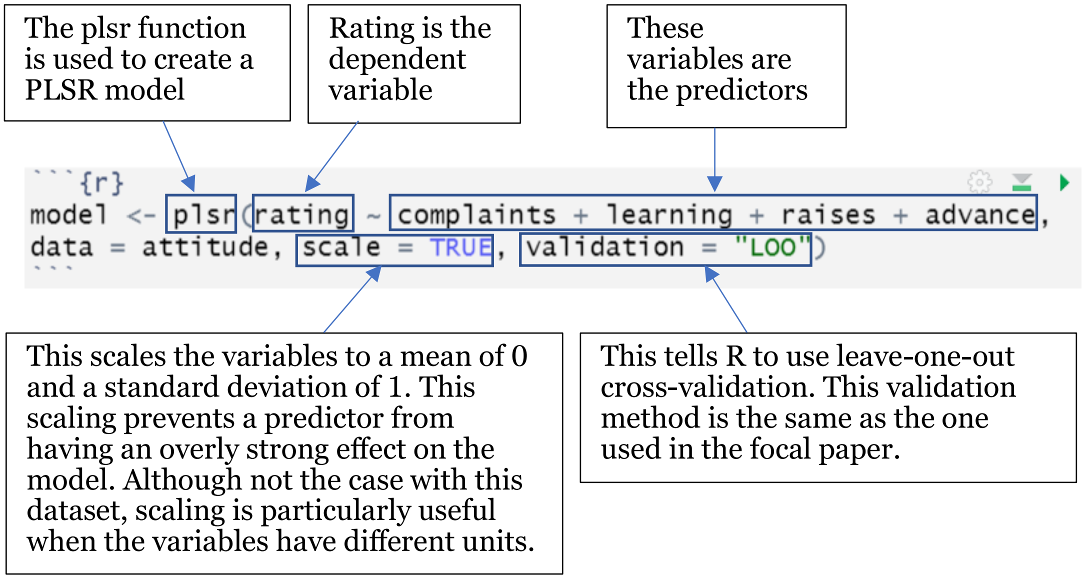
rating: Overall rating
complaints: How the department handled employee complaints
privileges: The department did not allow special privileges
learning: There was an opportunity to learn
raises: Raises were given based on performance
critical: How critical was the department
advance: There were advancement opportunities given by the department

For more information about the attitude dataset, view the R documentation: https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/attitude.html

PLSR MODEL

For this example, we will use rating as the response variable and complaints, learning, raises, and advance as the predictors.

We can create a PLSR model for these variables and view the summary statistics. We will use this information to help determine how many components we want to keep in our final model.



The plsr function is used to create a PLSR model

Rating is the dependent variable

These variables are the predictors

```{r}
model <- plsr(rating ~ complaints + learning + raises + advance,
data = attitude, scale = TRUE, validation = "LOO")
```

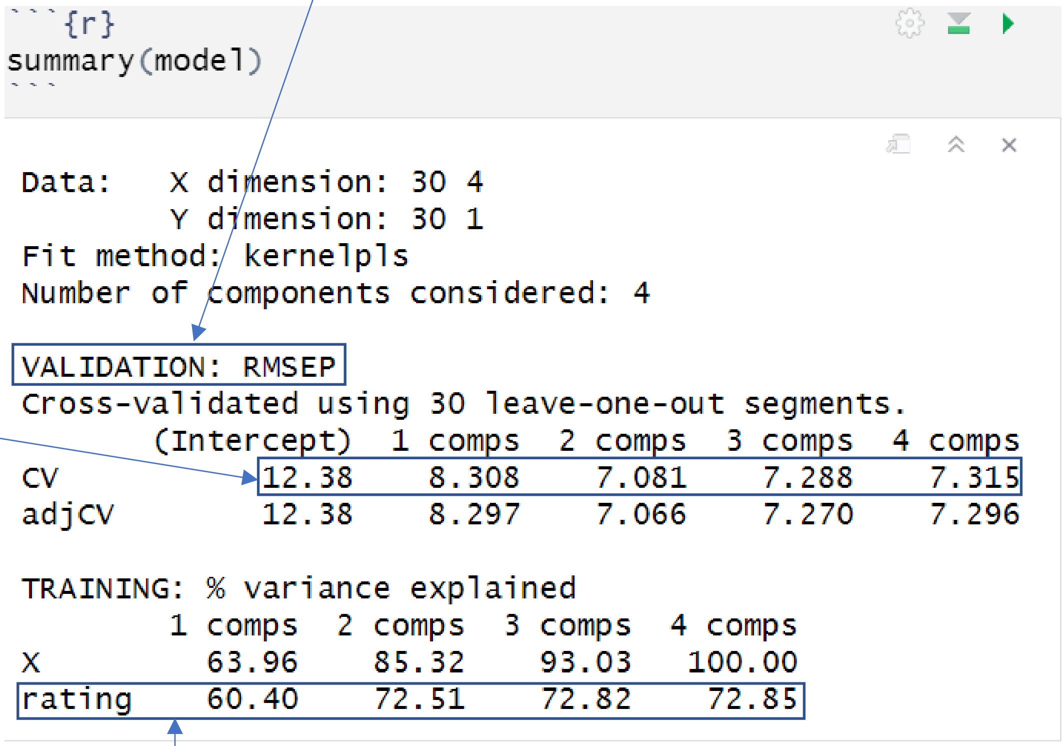
This scales the variables to a mean of 0 and a standard deviation of 1. This scaling prevents a predictor from having an overly strong effect on the model. Although not the case with this dataset, scaling is particularly useful when the variables have different units.

This tells R to use leave-one-out cross-validation. This validation method is the same as the one used in the focal paper.

Now, we can take a look at the summary statistics for the PLSR model.

The Root Mean Square Error of Prediction (RMSEP) measures the average difference between predicted values and the actual values. RMSE is a measure of uncertainty, so a good model generally has a low RMSE.

When the model includes just the intercept term, the RMSE is 12.38. After adding the first and second components, the RMSE decreases. But when we add the third component, the RMSE increases. Since we want the model with a small RMSE, we can determine that we should include less than three components in the model.



~~~
```{r}
summary(model)
```
~~~

```
Data:   X dimension: 30 4 
        Y dimension: 30 1
Fit method: kernelpls
Number of components considered: 4

VALIDATION: RMSEP
Cross-validated using 30 leave-one-out segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps
CV           12.38    8.308    7.081    7.288    7.315
adjCV        12.38    8.297    7.066    7.270    7.296

TRAINING: % variance explained
        1 comps  2 comps  3 comps  4 comps
X         63.96    85.32    93.03   100.00
rating    60.40    72.51    72.82    72.85
```

This line depicts the percent variance in the dependent variable explained by the components. For example, the first component explains 60.4% of the variation in the dependent variable. After adding the second component, the model can explain 72.51% of the variation in the dependent variable. The percent of variation explained will always increase as we add more components, but adding components three and four only slightly increases the percent of variation explained.

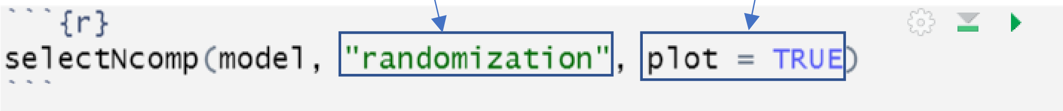
The results from the summary statistics suggest that we want less than three components in our final model. But there is a bit of subjectivity around the exact number of components that should be included in our final model. Some sources suggest selecting the number of components that minimizes the RMSE (Bobbitt, 2020). In this example, the smallest RMSE is when two components are included in the model. Another method is to select the fewest number of components that is still one standard error away from the best model (one-sigma approach). Other sources suggest a permutation approach which uses the fewest number of components possible so long as it does not significantly decrease the performance of the model (Mevik & Wehrens, 2022). The latter option can be automated using the selectNcomp function in R, and this is the method used in the focal paper.

PLS COMPONENT SELECTION

We can run the selectNcomp function and see how the results compare to visually inspecting the summary statistics.

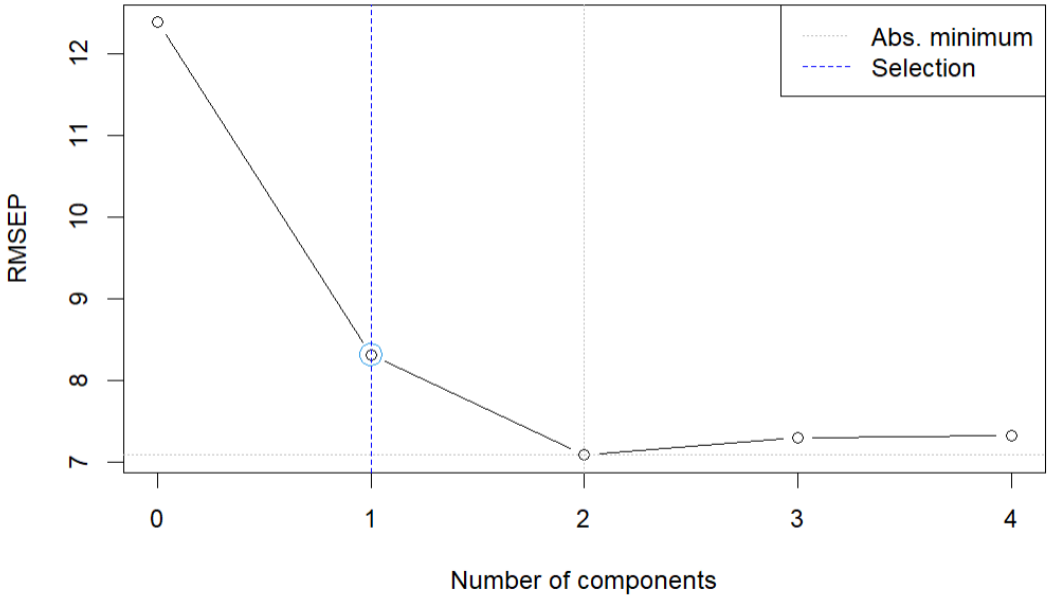
This tells R to use the randomization/permutation method. This can be replaced with "onesigma" to use the one-sigma approach

This creates the cross-validation plot shown below. This plot shows the suggested optimal number of components



```{r}
selectNcomp(model, "randomization", plot = TRUE)
```

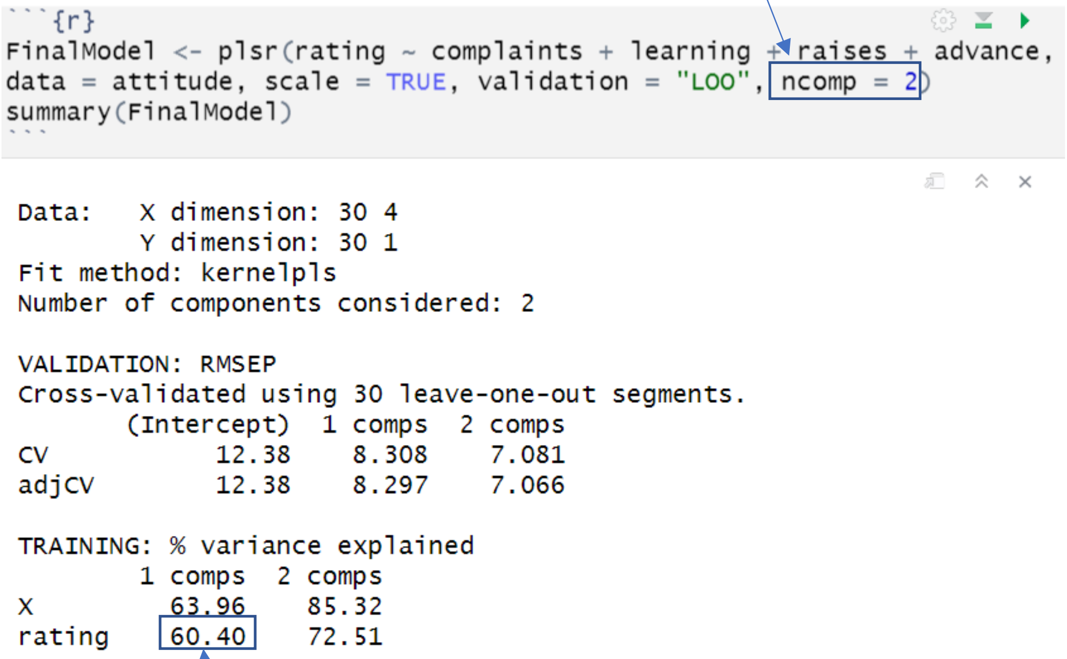
For more information about the selectNcomp function, check out the R documentation: [*https://search.r-project.org/CRAN/refmans/pls/html/selectNcomp.html*](https://search.r-project.org/CRAN/refmans/pls/html/selectNcomp.html)



The function suggests that only one component should be included in the model. Even though there is a lower RMSE with two components, the cross-validation plot suggests that a model with one component still performs well. The focal paper uses this randomization approach with the selectNcomp function. However, because we want to take a look at a scoreplot and biplot in this example, we will use two components in the final model, instead of the one component suggested by the cross-validation plot.

FITTING THE FINAL PLS MODEL

This final model is similar to the last model. The only difference is that we state the number of components that we want to include



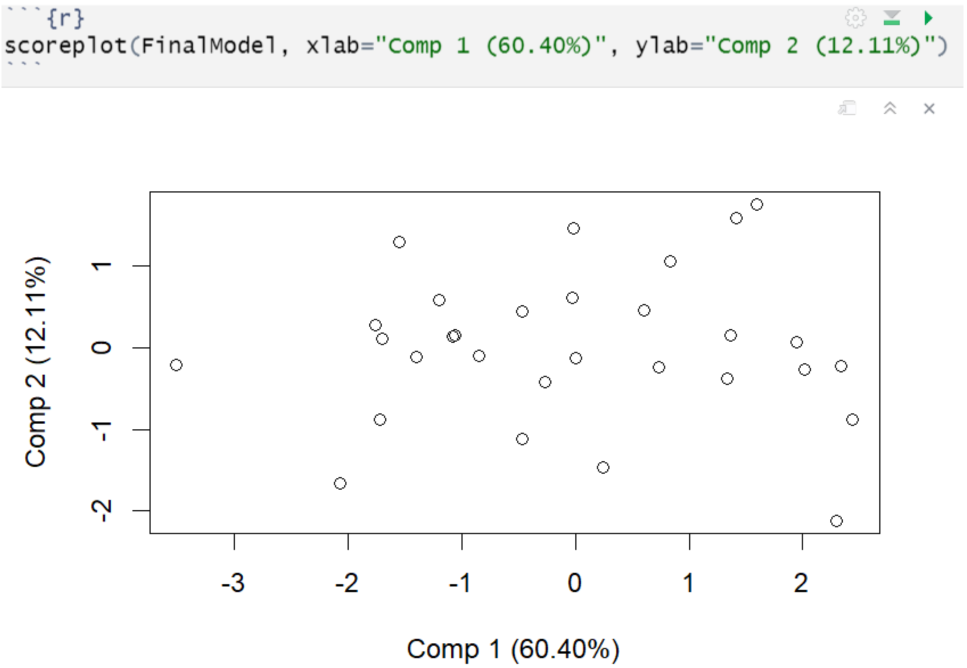
~~~
```{r}
FinalModel <- plsr(rating ~ complaints + learning + raises + advance,
data = attitude, scale = TRUE, validation = "LOO", ncomp = 2)
summary(FinalModel)
```
~~~

```
Data:    X dimension: 30 4
         Y dimension: 30 1
Fit method: kernelpls
Number of components considered: 2

VALIDATION: RMSEP
Cross-validated using 30 leave-one-out segments.
       (Intercept)  1 comps  2 comps
CV           12.38    8.308    7.081
adjCV        12.38    8.297    7.066

TRAINING: % variance explained
        1 comps  2 comps
X         63.96    85.32
rating    60.40    72.51
```

The first component explains about 60% of the variation in the dependent variable (the overall rating given to each department)

CREATING A SCOREPLOT

In a scoreplot, points are clustered based on their similarity. Points close to the average value will appear near the origin of the score plot. Whereas scores further out are more extreme values. Scoreplots are usually used to observe the structure of our data. We can locate similar observations and any outliers using the scoreplot.

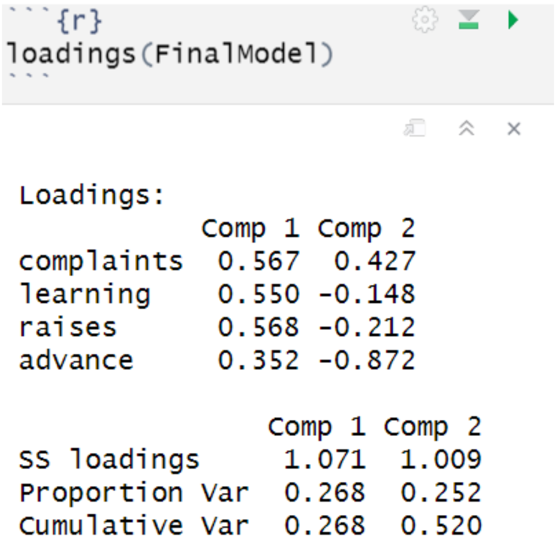


~~~
```{r}
scoreplot(FinalModel, xlab="Comp 1 (60.40%)", ylab="Comp 2 (12.11%)")
```
~~~

In this example, we can see the general structure of the data points representing the overall rating given to the departments.

ASSESSING THE LOADINGS

The loadings represent how much the variable contributes to a component.



~~~
```{r}
loadings(FinalModel)
```
~~~

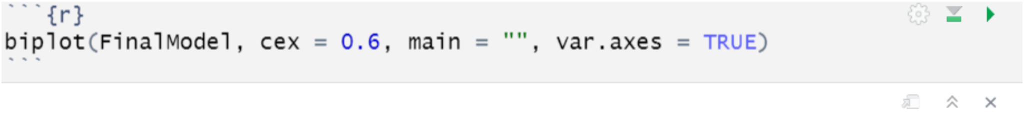
```
Loadings:
           Comp 1 Comp 2
complaints  0.567  0.427
learning    0.550 -0.148
raises      0.568 -0.212
advance     0.352 -0.872

               Comp 1 Comp 2
SS loadings     1.071  1.009
Proportion Var  0.268  0.252
Cumulative Var  0.268  0.520
```

For example, we can see that the complaints variable has a positive correlation with both components but has a stronger correlation with component 1 than component 2. Learning has a positive correlation with component 1 and a weaker negative correlation with component 2.

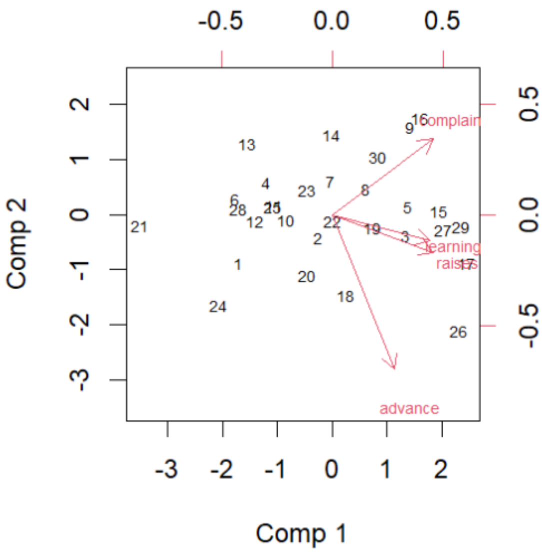
CREATING A BIPLOT

A biplot combines a scoreplot (represented by the numbers) and a loading plot (represented by the arrows). The length of the arrows represents the magnitude of the loadings. A longer arrow has a stronger influence on a component than a shorter arrow.

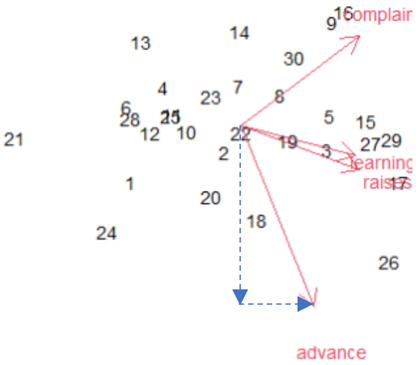
The direction/orientation of the arrow *along* a component axis represents which component the variable influences. The direction of the arrow also indicates whether the relationship with the component is positive or negative.



~~~
```{r}
biplot(FinalModel, cex = 0.6, main = "", var.axes = TRUE)
```
~~~



For example, let’s take a look at the advance variable. Imagine breaking the arrow for this variable up into a triangle with one side being completely vertical and another side being completely horizontal (see the blue dotted arrows below for reference).



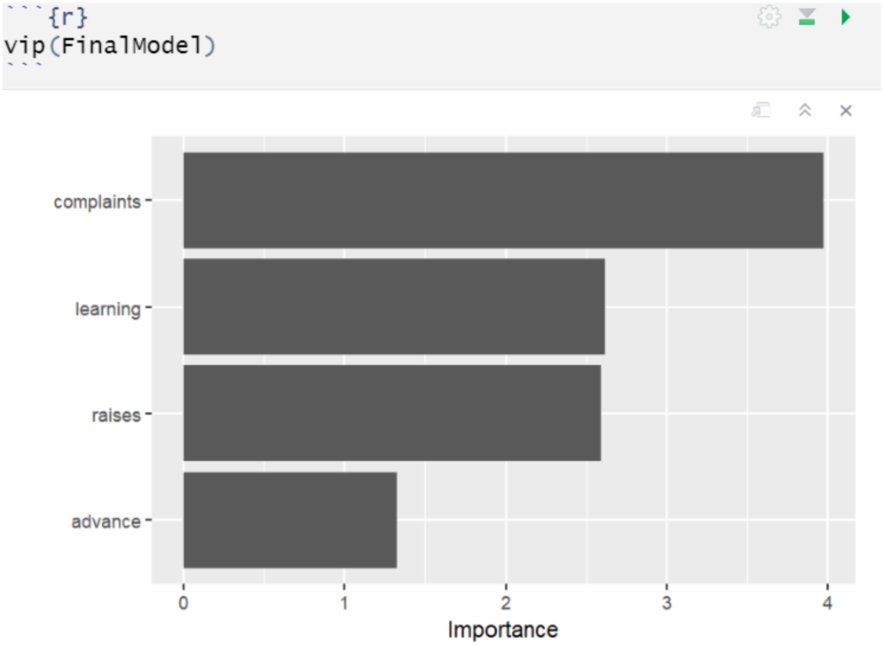
Notice that the vertical length of the arrow is longer than the horizontal length of the arrow. Also, note that the vertical arrow points down and the horizontal arrow points right.

Looking at the biplot above, we see that component 1 is represented by the horizontal x-axis and component 2 is represented by the vertical y-axis.

Combining these pieces of information, we can interpret that the advance variable has a strong negative correlation with component 2 and a weaker positive correlation with component 1.

VARIABLE IMPORTANCE PLOT

We will now create a Variable Importance Plot (VIP) for the PLS model. A VIP shows the relative contribution of each predictor to the PLS components included in the final model. For our example, variables that contribute more to component 1 will likely have higher variable importance, since component 1 explains 60% of the variation in the response variable.



~~~
```{r}
vip(FinalModel)
```
~~~

The VIP shows that the complaints variable is the most important variable in our model. This result is supported by the loadings for the complaints variable which we observed previously. The loadings show that the complaints variable has a relatively high contribution to both of the components.

R ANALYSIS

You now have the skills to conduct create and validate a partial least squares regression model. Proceed to the R markdown document to apply a PLSR to the western monarch abundance and environmental data. In the R analysis section, you will also learn how to create and interpret a variable importance plot and biplot.

KEY TAKEAWAYS

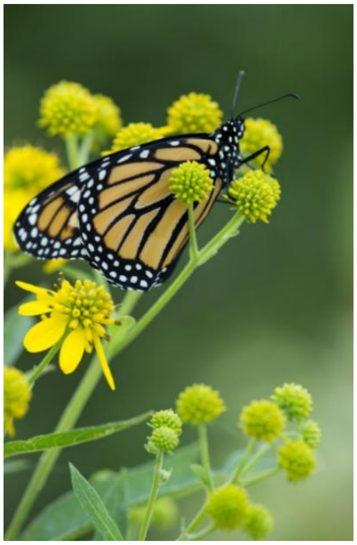
After reading through the lesson and working through the R analysis, you can now answer the following ecological and analysis questions.

Ecological questions:

1. Why is it important to understand what factors are driving a species' decline?
2. What is the difference between the western and eastern monarch populations?
3. Briefly research a threatened or endangered species of your choice. Name one climate factor and one land use factor that might be affecting that species' population.

Study design questions:

4. Is the study from the focal paper an experiment or an observational study?
5. Choose one of the variables used in the study. Where did the data for that variable come from?



Questions from R markdown:

6. After evaluating the RMSE and percent variance explained in the initial PLSR model, how many components do you think should be included in the final PLSR model?
7. Looking at the final model for the monarch data, how much variation in the monarch abundance is explained by the first component? How much variation in the monarch abundance is explained by the second component?
8. Based on the biplot, do land use variables or climate variables appear to have a stronger effect on monarch abundance (i.e. which variables have a strong correlation with component 1)?
9. Based on the Variable Importance Plot, do land use or climate factors contribute more to our model?

Critical thinking questions:

10. One of the goals of the focal paper is to advocate for the use of PLSR over multiple regression and principal component analysis when there are multiple correlated variables. In your opinion, did you find their argument convincing, or do you think the cons of PLSR in its difficult interpretability outweigh its accuracy?
11. Think of a research project you are working on or any topic of interest. Give an example of why PLSR might be well-suited or ill-suited for that data.

References

Abdi, H. (n.d.). *Partial Least Squares (PLS) Regression.*

Allen, M. P. (1997). The problem of multicollinearity. In *Understanding Regression Analysis* (pp. 176–180). Springer US. https://doi.org/10.1007/978-0-585-25657-3_37

Bobbitt, Z. (2020a, November 17). *An Introduction to Partial Least Squares*. Statology. https://www.statology.org/partial-least-squares/

Bobbitt, Z. (2020b, November 17). *Partial Least Squares in R (Step-by-Step)*. Statology. https://www.statology.org/partial-least-squares-in-r/

Brook, B., Sodhi, N., & Bradshaw, C. (2008). Synergies among extinction drivers under global change. *Trends in Ecology & Evolution*, *23*(8), 453–460. https://doi.org/10.1016/j.tree.2008.03.011

Crone, E. E., Pelton, E. M., Brown, L. M., Thomas, C. C., & Schultz, C. B. (2019). Why are monarch butterflies declining in the West? Understanding the importance of multiple correlated drivers. *Ecological Applications*, *29*(7). https://doi.org/10.1002/eap.1975

Hartmann, K., Krois, J., & Waske, B. (2018). *E-Learning Project SOGA: Statistics and Geospatial Data Analysis*. Department of Earth Sciences, Freie Universitaet Berlin. https://www.geo.fu-berlin.de/en/v/soga/Geodata-analysis/Principal-Component-Analysis/principal-components-basics/Interpretation-and-visualization/index.html

Hogue, A. S., & Breon, K. (2022). The greatest threats to species. *Conservation Science and Practice*, *4*(5), e12670. https://doi.org/10.1111/csp2.12670

Korstanje, J. (2021, July 18). *Partial Least Squares*. Medium. https://towardsdatascience.com/partial-least-squares-f4e67144452a

Mevik, B.-H., & Wehrens, R. (2022). *Introduction to the pls Package.*

Migratory monarch butterfly now Endangered—IUCN Red List. (n.d.). IUCN. Retrieved April 13, 2023, from https://www.iucn.org/press-release/202207/migratory-monarch-butterfly-now-endangered-iucn-red-list

Milkweed for Monarchs. (n.d.). National Wildlife Federation. Retrieved April 13, 2023, from https://www.nwf.org/Home/Garden-for-Wildlife/About/Native-Plants/Milkweed

Partial Least Squares regression (PLS). (n.d.). XLSTAT. Retrieved April 15, 2023, from https://www.xlstat.com/en/solutions/features/partial-least-squares-regression

Roper, W. (2020, February 4). *Western Monarchs Rapidly Declining*. Statista. https://www.statista.com/chart/20716/monarch-butterfly-population-decline

SelectNcomp. (n.d.). RDocumentation. Retrieved April 16, 2023, from https://www.rdocumentation.org/packages/pls/versions/2.8-1/topics/selectNcomp

Smith, D. R., Allan, N. L., McGowan, C. P., Szymanski, J. A., Oetker, S. R., & Bell, H. M. (2018). Development of a Species Status Assessment Process for Decisions under the U.S. Endangered Species Act. *Journal of Fish and Wildlife Management*, *9*(1), 302–320. https://doi.org/10.3996/052017-JFWM-041

Stenoien, C., Nail, K. R., Zalucki, J. M., Parry, H., Oberhauser, K. S., & Zalucki, M. P. (2018). Monarchs in decline: A collateral landscape-level effect of modern agriculture: Landscape-level effect of GM crops on monarchs. *Insect Science*, *25*(4), 528–541. https://doi.org/10.1111/1744-7917.12404

Thogmartin, W. E., Wiederholt, R., Oberhauser, K., Drum, R. G., Diffendorfer, J. E., Altizer, S., Taylor, O. R., Pleasants, J., Semmens, D., Semmens, B., Erickson, R., Libby, K., & Lopez-Hoffman, L. (2017). Monarch butterfly population decline in North America: Identifying the threatening processes. *Royal Society Open Science*, *4*(9), 170760. https://doi.org/10.1098/rsos.170760

Tobias, R. D. (n.d.). *An Introduction to Partial Least Squares Regression*.

Picture References

https://digitalmedia.fws.gov/digital/collection/natdiglib/id/18003/rec/2

https://www.statista.com/chart/20716/monarch-butterfly-population-decline/

https://digitalmedia.fws.gov/digital/collection/natdiglib/id/16066/rec/12

https://digitalmedia.fws.gov/digital/collection/natdiglib/id/18238/rec/38

https://digitalmedia.fws.gov/digital/collection/natdiglib/id/18251/rec/49

https://digitalmedia.fws.gov/digital/collection/natdiglib/id/17945/rec/52

https://digitalmedia.fws.gov/digital/collection/natdiglib/id/18273/rec/23