Linear Regression and its Cousins

Group 4 Members:
Paul Britton
Anthony Munoz
Luisa Velasco
Javern Wilson

Introduction

Models to be discussed:

- Ordinary Linear Regression
- Partial Least Square (PLS)
- Penalized Models: Lasso, Ridge and Elastic Net

Introduction

Each model can be written in the form directly or indirectly:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_P x_{iP} + e_i,$$

Objective: Each of these models seeks to find estimates of the parameters so that the sum of the squared errors or a function of the sum of the squared errors is minimized. The estimates fall along the spectrum of the bias-variance trade-off.

- Ordinary linear regression, finds parameter estimates that have minimum bias
- Ridge, lasso, and the elastic net find estimates that have lower variance.

Introduction - Things to keep in Mind

Some things to keep in mind when linear regression is applied to time series specific use cases

- Technique is familiar, context is slightly different
- **Time** is an independent variable here (otherwise it's not a "time-series")
- Time series data often not IID.
- Order of the data matters as the goal is to predict values in sequence

Introduction

- **Overall Advantages of models:**
  - Highly interpretable
  - Relationships among predictors can be further interpreted through the estimated coefficients.
  - Enables computation of standard errors of the coefficients.

- **Overall Disadvantages on models:**
  - Relationship between the predictor and response are expected to fall along a straight line (1 - predictor example) or a flat hyperplane for multiple predictors
  - Augmentation or predictors may be required in the case on non-linear relationships and in some cases, linear models may not not be adequate/suitable

What is Ordinary Linear Regression?

- A type of Predictive Analysis
  - Main Players
    - Predictor Variables
    - Dependent / Response Variable

- Objective:
  - Finding the plane that minimizes the sum-of-squared errors (SSE) between the observed and predicted response
  - Why? Reduces bias

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

What is Ordinary Linear Regression?

- Preprocessing techniques:
  - Addressing Multicollinearity
    - Use of the variance inflation factor (VIF)
  - Outlier Treatment
    - Remove or use Transformation techniques
  - Missing Values
    - Imputation or removal

What is Ordinary Linear Regression?

If too many predictors remain after preprocessing steps...

- Danger: Overfitting - Lacks degree of freedom

Alternatives:

- Principal Component Analysis (PCA)
  - Aims at reducing a large set of variables to a small set that explains most the variance.
- PLS
  - Simultaneous dimension reduction
- Ridge, Lasso and Elastic Net
  - Shrinking parameter estimates

What is Ordinary Linear Regression?

- Limitations:
  - Does not take into consideration if the data have curvature or nonlinear structure.
    - Use diagnostic plots for visual
    - Adding Polynomials - Quadratic (squared), cubic (cubed) terms turns a linear regression model into a curve. This makes it a straightforward way to model curves without having to model complicated non-linear models.
  - Focusing on outliers
    - Very sensitive to outliers

What is Ordinary Linear Regression?

- Model Evaluation / Goodness of Fit - Coefficient of Determination (R2)
  - 1. The squared correlation coefficient
    - Measures the percentage of variable behavior explained by the model
  - 2. Residual Standard Error
    - Can be compared to sample mean or standard deviation of y for insight into model accuracy

1. $$R^2 = \frac{\sum(\hat{y}_t - \bar{y})^2}{\sum(y_t - \bar{y})^2},$$

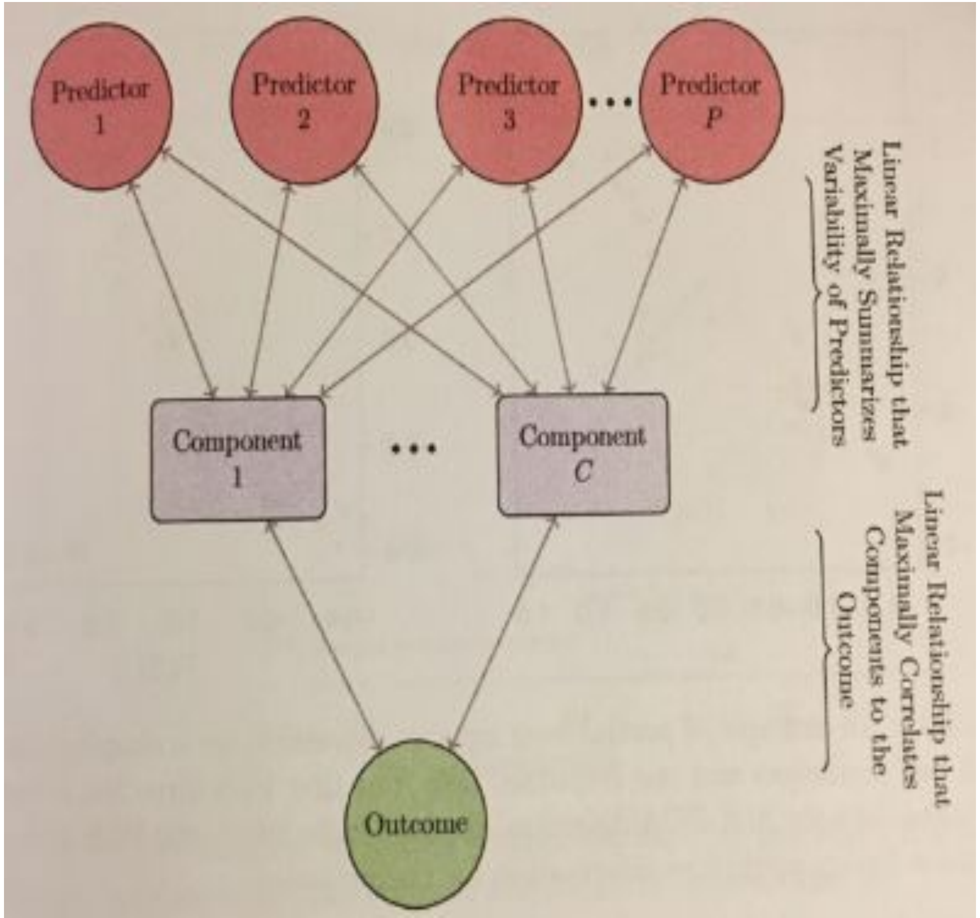
2. $$\hat{\sigma}_e = \sqrt{\frac{1}{T-k-1}\sum_{t=1}^{T} e_t^2},$$

Cousins of Linear Regression

- Partial Least Square (PLS)

PLS

- Predictors are greater than the # of observation
- Highly Correlation between predictors variables.



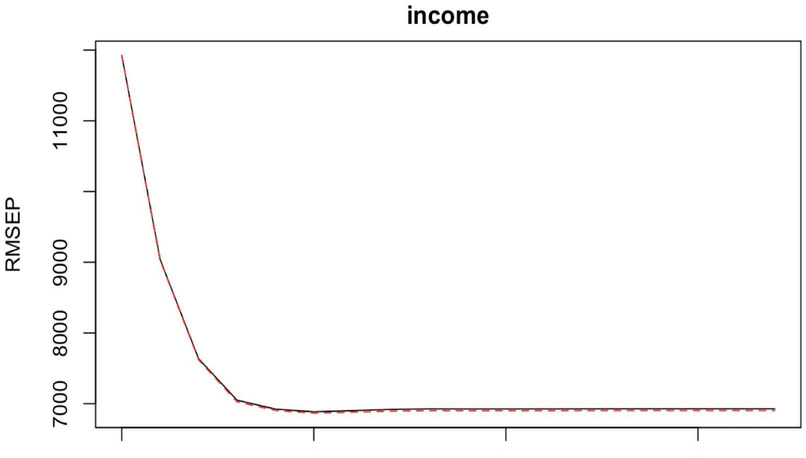
MROZ - LABOR SUPPLY DATA

```
'data.frame':   753 obs. of  18 variables:
 $ work      : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2
 $ hoursw    : int  1610 1656 1980 456 1568 2032 1440 1020 1458 1
 $ child6    : int  1 0 1 0 1 0 0 0 0 0 ...
 $ child618  : int  0 2 3 3 2 0 2 0 2 2 ...
 $ agew      : int  32 30 35 34 31 54 37 54 48 39 ...
 $ educw     : int  12 12 12 12 14 12 16 12 12 12 ...
 $ hearnw    : num  3.35 1.39 4.55 1.1 4.59 ...
 $ wagew     : num  2.65 2.65 4.04 3.25 3.6 4.7 5.95 9.98 0 4.15
 $ hoursh    : int  2708 2310 3072 1920 2000 1040 2670 4120 1995
 $ ageh      : int  34 30 40 53 32 57 37 53 52 43 ...
 $ educh     : int  12 9 12 10 12 11 12 8 4 12 ...
 $ wageh     : num  4.03 8.44 3.58 3.54 10 ...
 $ income    : int  16310 21800 21040 7300 27300 19495 21152 1890
 $ educwm    : int  12 7 12 7 12 14 14 3 7 7 ...
 $ educwf    : int  7 7 7 7 14 7 7 3 7 7 ...
 $ unemprate : num  5 11 5 5 9.5 7.5 5 5 3 5 ...
 $ city      : Factor w/ 2 levels "no","yes": 1 2 1 1 2 2 1 1 1 1
 $ experience: int  14 5 15 6 7 33 11 35 24 21 ...
```

```
Data:   X dimension: 366 17
        Y dimension: 366 1
Fit method: kernelpls
Number of components considered: 17

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
CV           11925     9039     7637     7050     6924     6886     6901     6920     6928
adjCV        11925     9030     7622     7026     6906     6865     6879     6896     6903
       9 comps  10 comps  11 comps  12 comps  13 comps  14 comps  15 comps  16 comps  17 comps
CV        6926      6927      6927      6928      6928      6928      6928      6928      6928
adjCV     6902      6902      6903      6903      6904      6904      6904      6904      6904

TRAINING: % variance explained
        1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps
X         18.09    28.61    37.99    50.35    59.01    62.74    65.45    69.14    74.10
income    44.66    61.79    68.55    69.72    70.28    70.35    70.38    70.39    70.39
        10 comps  11 comps  12 comps  13 comps  14 comps  15 comps  16 comps  17 comps
X          79.00     83.43     86.47     89.40     91.96     95.61     97.71    100.00
income     70.39     70.39     70.39     70.39     70.39     70.39     70.39     70.39
```



Cousins of Linear Regression

Penalized Methods (*a.k.a. Regularization, Shrinkage Methods)*

- Helps to reduce OLS high variance due to overfitting and/or multicollinearity
- Adds penalty to OLS Sum of Squared Residuals when estimates become too large
- Trades-off a little Bias for substantial drop in Variance

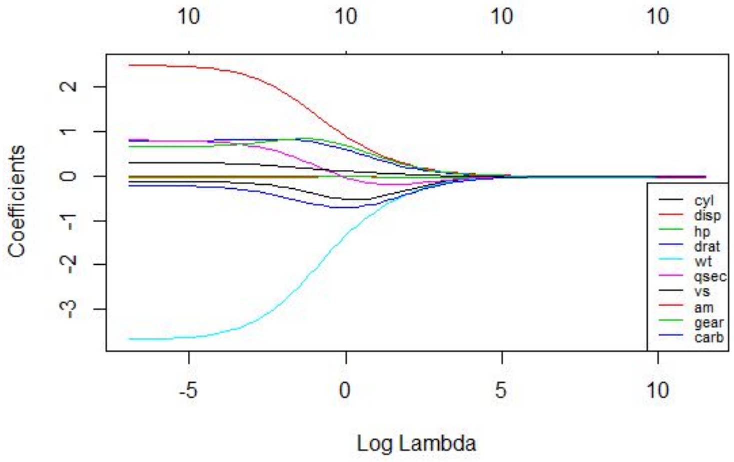
Regression Models: Ridge, Lasso, Elastic Net

R Implementation: cv.glmnet, glmnet
λ - parameter that controls shrinkage ($\lambda \geq 0$)
α - parameter that identifies the penalty model ($0 \leq \alpha \leq 1$)

Cousins of Linear Regression

- **Ridge Regression** $L = \sum(\hat{Y}i - Yi)^2 + \lambda\sum\beta^2$

- penalizes sum of squared values (L2 penalty)
- can shrink the coefficients towards 0 as penalty increases (but does not equal 0)



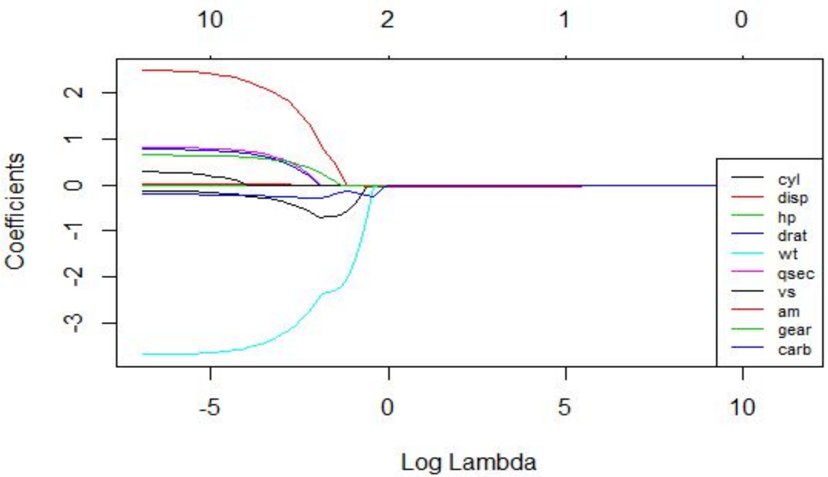
cyl	-0.374250969
disp	-0.005317757
hp	-0.011519902
drat	1.055565066
wt	-1.207393974
qsec	0.160397178
vs	0.784538697
am	1.594311112
gear	0.541902172
carb	-0.536015354

Cousins of Linear Regression

- **Lasso Regression**

$$L = \sum(\hat{Y}i - Yi)^2 + \lambda\sum|\beta|$$

- penalizes sum of absolute values (L1 penalty)
- can shrink the coefficients to absolute 0 (= feature selection)

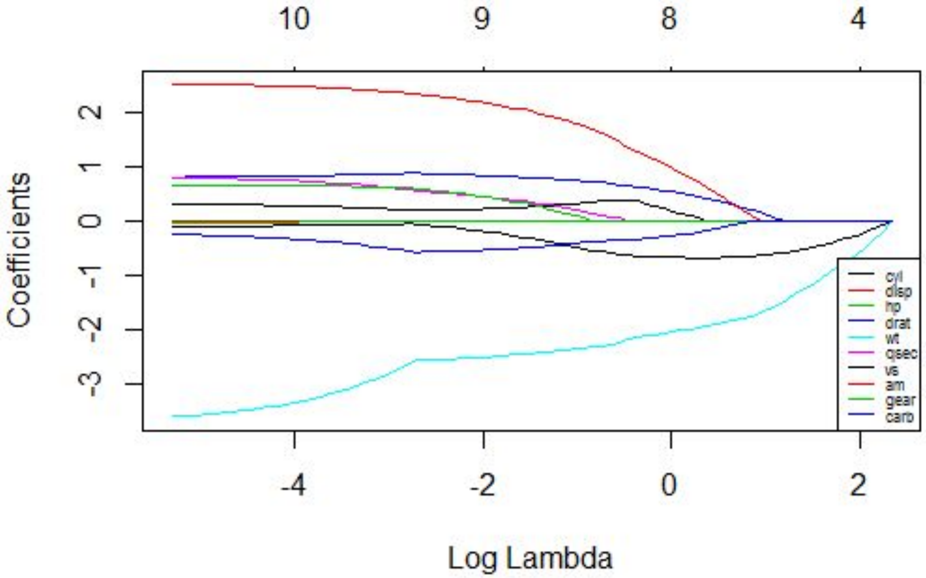



```
cyl          -0.88521909
disp           .
hp           -0.01160163
drat           .
wt           -2.70237673
qsec           .
vs             .
am             .
gear           .
carb           .
```

Cousins of Linear Regression

- **Elastic Net Regression**
  - Dynamic blending of Ridge & Lasso

$$L = \sum(\hat{Y}i - Yi)^2 + \lambda \sum \beta^2 + \lambda \sum |\beta|$$



cyl	-0.65755612
disp	-0.00423864
hp	-0.01144720
drat	0.25445124
wt	-1.77867668
qsec	.
vs	.
am	0.21478442
gear	.
carb	-0.04189935

Case Study - Predicting Solubility Using Chemical Structures

Researchers set out to predict solubility of numerous chemical compounds

1267 compounds examined

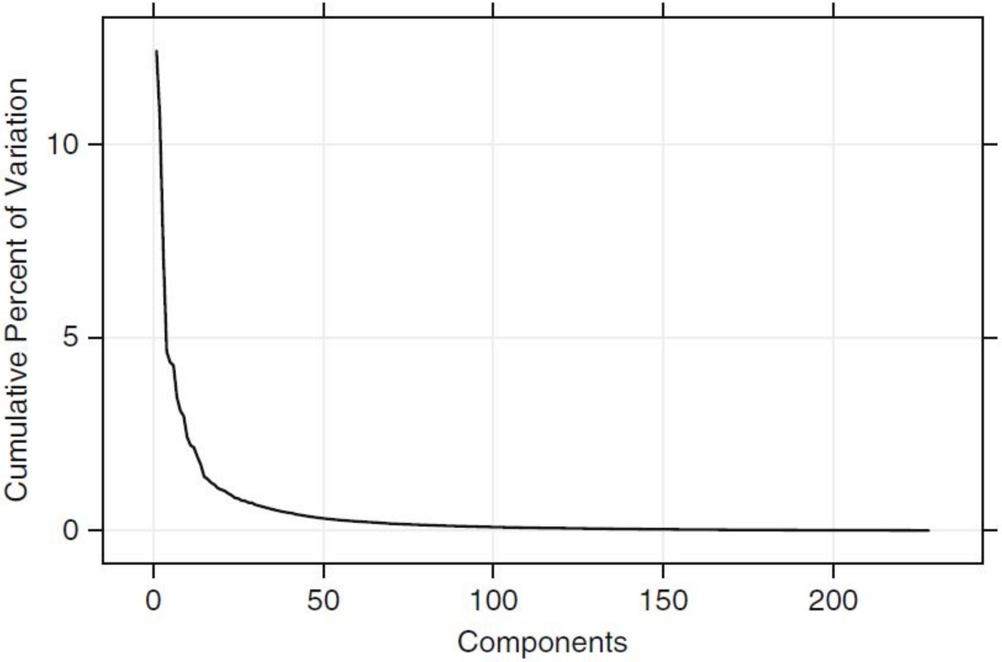
In addition ~230 more intuitive descriptor variables from 3 distinct groups - uncorrelated on average, but many strong pairwise relationships (208 binary, 16 discreet descriptors, 4 continuous)

Case Study - Processing

-Researchers note that based on the nature of the variables there is likely room for reduction

-They apply PCA and determine that the data structure can be summarized in a much smaller space than the original dimensions

-Suggestive of a potential problem with multicollinearity - an issue for linear regression ...something to watch out for



Ref: Applied Predictive Modeling. Kuhn & Johnson. 2013

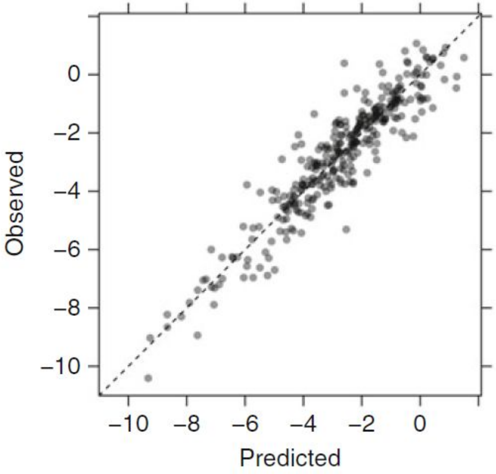
Case Study - Linear Regression

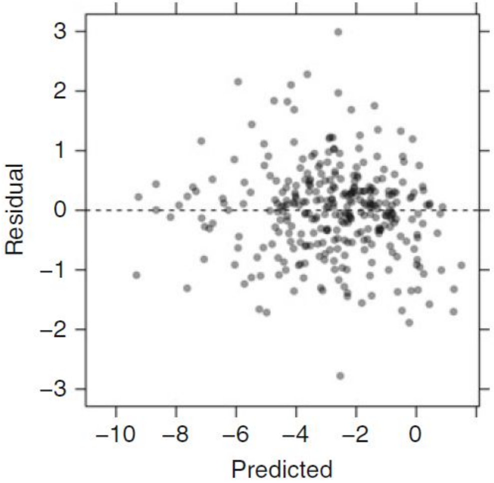
-They started by removing all predictors with pair-wise correlations > 0.9

-Trained a linear model and validated using k-fold (k=10) cross validation

-R^2 and RMSE were 0.88 and 0.71 respective

-Model appears to fit quite well.





Ref: Applied Predictive Modeling. Kuhn & Johnson. 2013

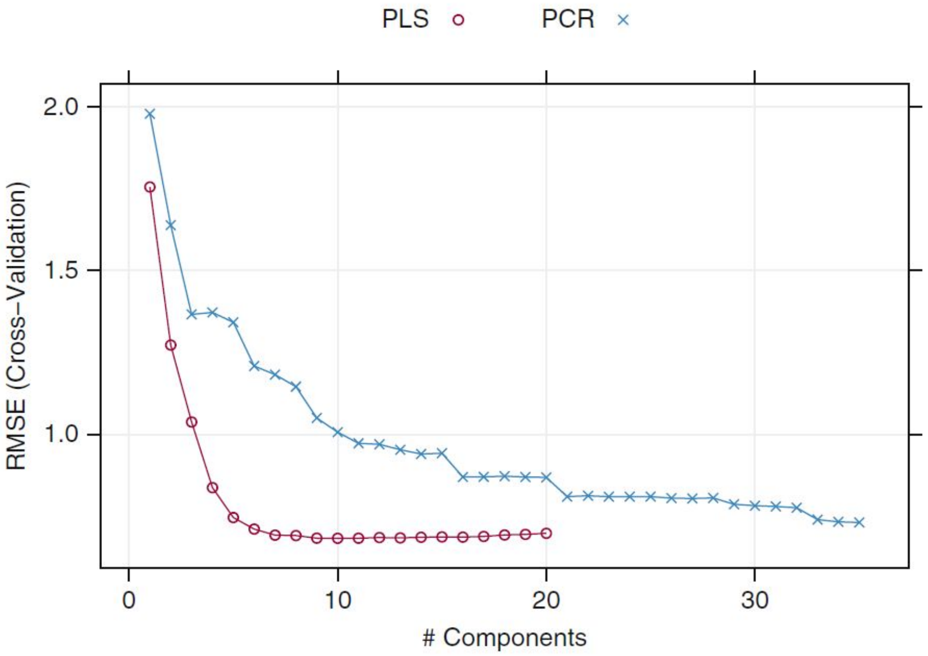
Case Study - Partial Least Squares

-Demonstrated for when a LR-type solution is desired and the predictors are correlated (as opposed to PCR)

-They ran both PCR and PLS to highlight differences

-By incorporating info about both variability AND correlation to response, PLS improves on PCA + Regrssion

-PLS and PCR produce RMSE of 0.982 and 0.731 respectively here

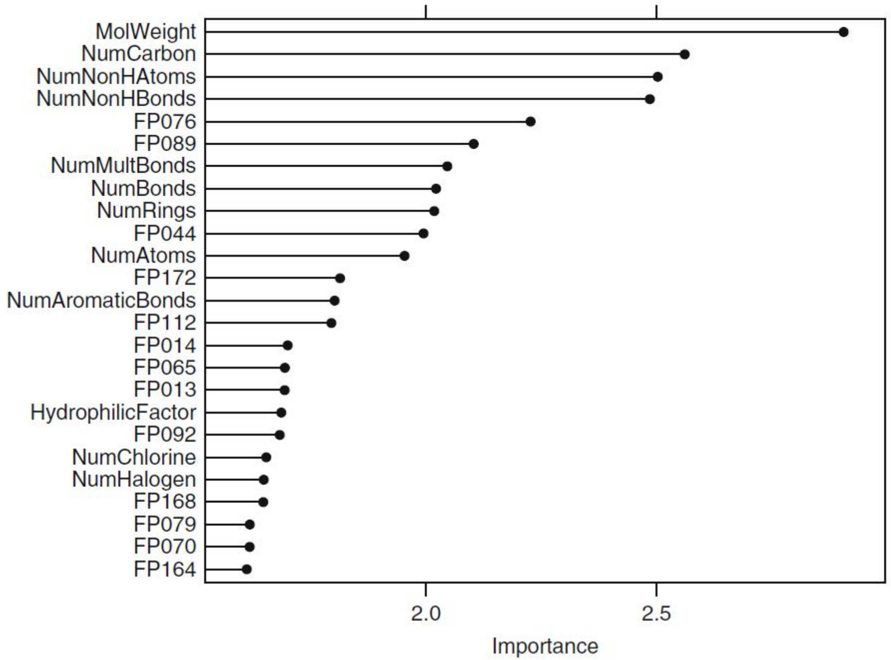


Ref: Applied Predictive Modeling. Kuhn & Johnson. 2013

Case Study - Partial Least Squares

- Similar to PCA, often hard to interpret contribution of individual variables

-Use VIP (variable importance in the projection) score to assist with gaining intuition about which variables matter, rather than trying to interpret “latent” variables



Ref: Applied Predictive Modeling. Kuhn & Johnson. 2013