PCA and PCR

Module 8

DATA 5600
Introduction to Regression and Machine Learning for Analytics
Marc Dotson

Module Overview

- High Dimensional Data
- Principal Component Analysis (PCA)
- Principal Component Regression (PCR)
  - Example with Continuous Response
  - Example with Binary Response

Introduction & High Dimensional Data

Recall From Module 4

Ways to address multicollinearity:

1. Remove some variables from the model – choose a subset of predictor variables

2. Apply a shrinkage method (ridge regression, LASSO, elastic net, etc.)

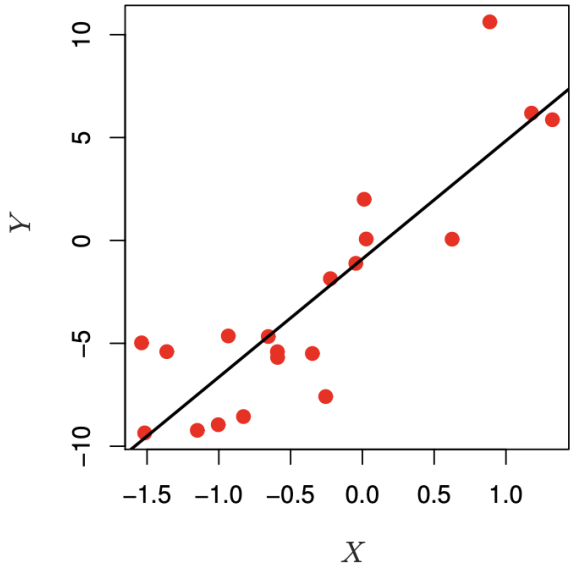
3. Combine the correlated variables somehow, like with **principal component regression (PCR)**

Ultimately, we want model *parsimony*: we want the simplest model that will get the job done

High Dimensional Data

- Recall, the rough rule of thumb of needing 6 to 10 observations per variable included in a regression model for stable/reliable results
- When the number of predictors is close to or greater than the sample size ($p \approx n$ or $p > n$), we have "high dimensional" data
  - High dimensional data has become increasingly more common (SNPs of DNA, search terms, etc.)
- With high dimensional data, traditional regression overfits to the data (see next slide)

High Dimensional Data

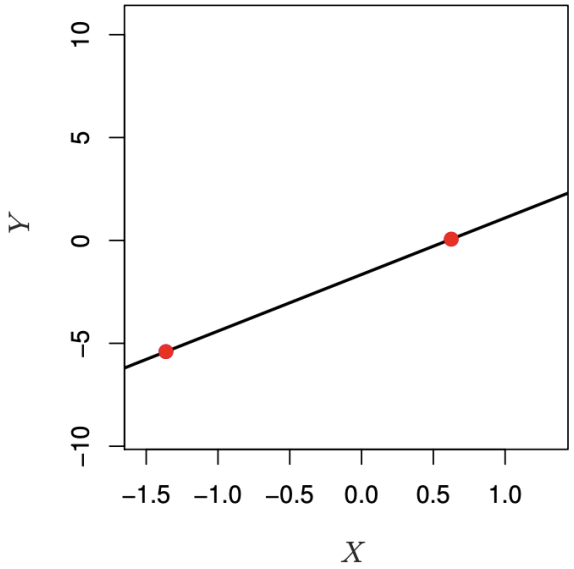


Low dimensional data

- $n = 20, p = 1$
- Regression works well

20:1

a ratio of 6-10:1 is generally considered ideal



High dimensional data

2:1

- $n = 2, p = 1$ $(p \approx n)$
- Regression breaks down
  - Model is perfectly fit to the training data, but will likely perform dismally on testing/new data
- Cannot get a stable estimate of the variability of the model, so things like the AIC and BIC fail
- Note: LASSO and elastic net can work in this situation

Principal Component Analysis (PCA)

PCA

- Principal Component Analysis (PCA) is a technique designed to reduce dimensionality
  - Ex: PCA could let us go from a 10-variable model to a 2-variable model while still retaining the important information from all 10 variables
  - The goal is to find a low dimensional representation of the data set that captures the important information from the original high dimensional data
- Evidence suggests most data sets exist on a lower dimensional manifold (ex: balled up piece of paper)
  - While all n observations exist in a p-dimensional space, not every dimension is interesting

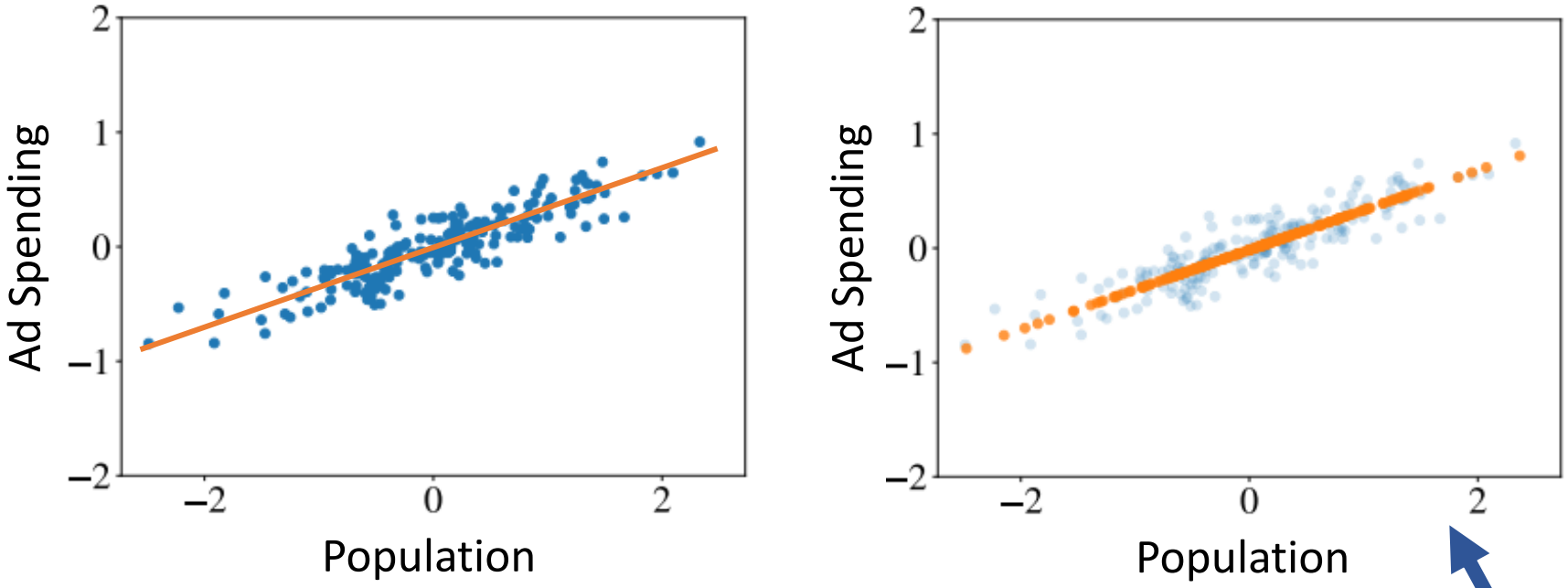
PCA

- **PCA identifies new dimensions that are linear combinations of all the predictor variables** Ex: $0.2X_1 + 0.8X_2$ or $0.7X_1 + 0.3X_2$
- PCA can transform highly correlated variables into a few main, **uncorrelated** components
  - Multicollinearity is no longer an issue
- Note: Since linear algebra is not a pre-requisite for this class, we will discuss only the big idea here
  - Under the hood, eigenvectors and eigenvalues are key to fully understanding PCA

PCA

Data set with two predictor variables (2-dimensional data):

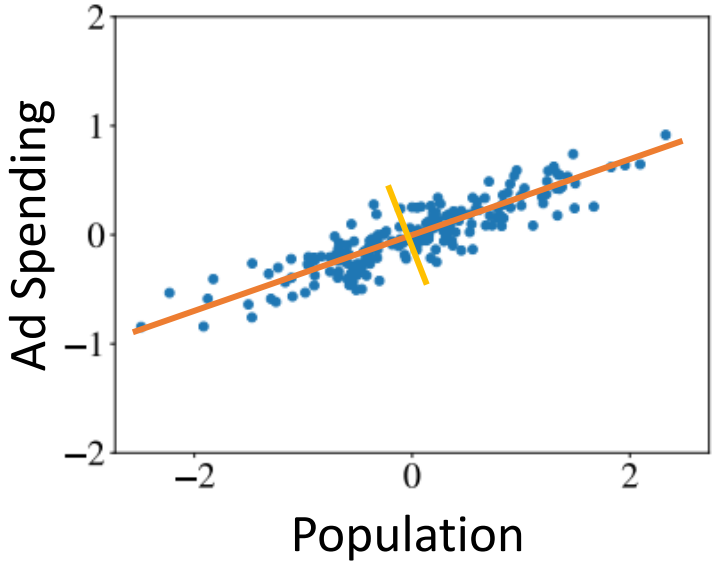
X_1: Ad Spending X_2: Population



- X_1 and X_2 are highly correlated
- Most of the information between X_1 and X_2 lies along the orange line
  - If we moved all the points to lie along the orange line, we'd capture a lot of the information from X_1 and X_2 (and decrease the dimensionality by 1)
  - The orange line is the **first principal component**

Figures modified from: http://ema.cri-info.cm/wp-content/uploads/2019/07/2019BurkovTheHundred-pageMachineLearning.pdf

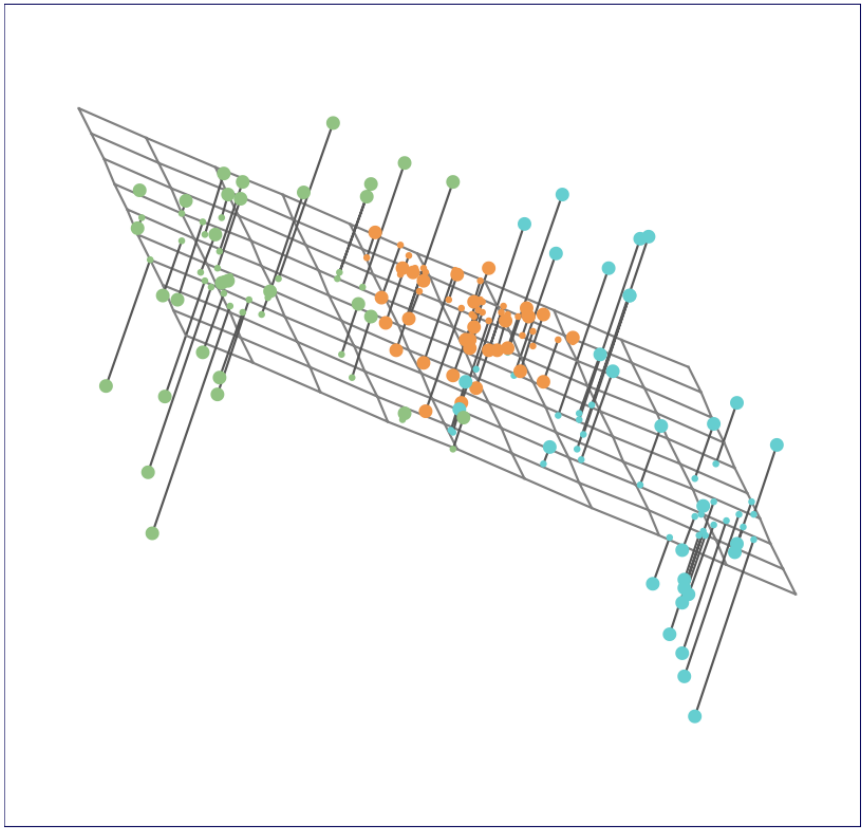
PCA



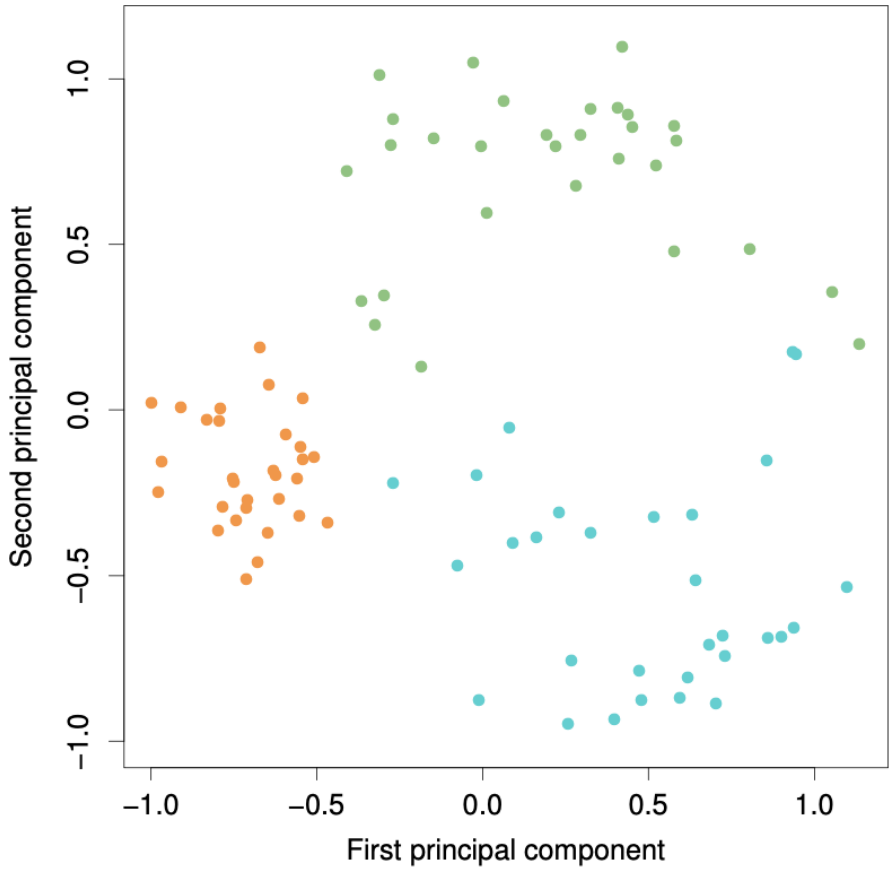
- The first principal component (PC_1) goes in the direction where the observations *vary the most*
- The second principal component (PC_2) is orthogonal to the first component and goes in the direction of the second largest variance of the observations
- If we had more than two variables, the third principal component (PC_3) would be orthogonal to both the first and second components and go in the direction of the third largest variance of the observations

PCA

Original data
(3 predictors, 3-dimensional)



Principal Components of the Original Data (2 components, 2-dimensional)



PCA

PC_1 is a **normalized linear combination** of the variables in our data:

$$\mathrm{PC}_1 = \phi_{1,1}X_1 + \phi_{2,1}X_2 + \cdots + \phi_{p-1,1}X_{p-1}$$

where $\sum_{j=1}^{p-1} \phi_{j,1}^2 = 1$ (this is what we mean by "normalized")

Notes:

- p represents the number of βs in the model (including the intercept), so $p - 1$ represents the number of variables
- ϕs are the "loadings" of PC_1
- $\phi_{j,1}$ represents the jth loading, corresponding to the jth variable, of the 1st principal component
- The X variables are all standardized (to have a mean of 0 and standard deviation of 1) before preforming PCA

PCA

PC_1 is a **normalized linear combination** of the variables in our data:

$$\mathrm{PC}_1 = \phi_{1,1}X_1 + \phi_{2,1}X_2 + \cdots + \phi_{p-1,1}X_{p-1}$$

where $\sum_{j=1}^{p-1} \phi_{j,1}^2 = 1$ (this is what we mean by "normalized")

Notes:

$$\mathrm{PC}_2 = \phi_{1,2}x_1 + \phi_{2,2}x_2 + \cdots + \phi_{p-1,2}x_{p-1}$$

- p represents the number of βs in the model (including the intercept), so $p - 1$ represents the number of variables
- ϕs are the "loadings" of PC_1
- $\phi_{j,1}$ represents the j[th] loading, corresponding to the j[th] variable, of the 1[st] principal component
- The X variables are all standardized (to have a mean of 0 and standard deviation of 1) before preforming PCA

Otherwise, the variable with the largest scale would have the largest variance in values and would automatically be the dimension with the largest variance

PCA

We find the best loadings based on this optimization problem:

$$\max_{\phi_{1,1},\ldots,\phi_{p-1,1}} \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p-1}\phi_{j,1}x_{ij}\right)^2$$

such that $\sum_{j=1}^{p-1}\phi_{j,1}^2 = 1$

- This can be solved via eigen decomposition, but we'll let Python do it for us
- Once the first principal component is found, we can calculate the next linear combination of variables that are uncorrelated (perpendicular, orthogonal) to PC_1
- There are at most p, principal components

PCA Example

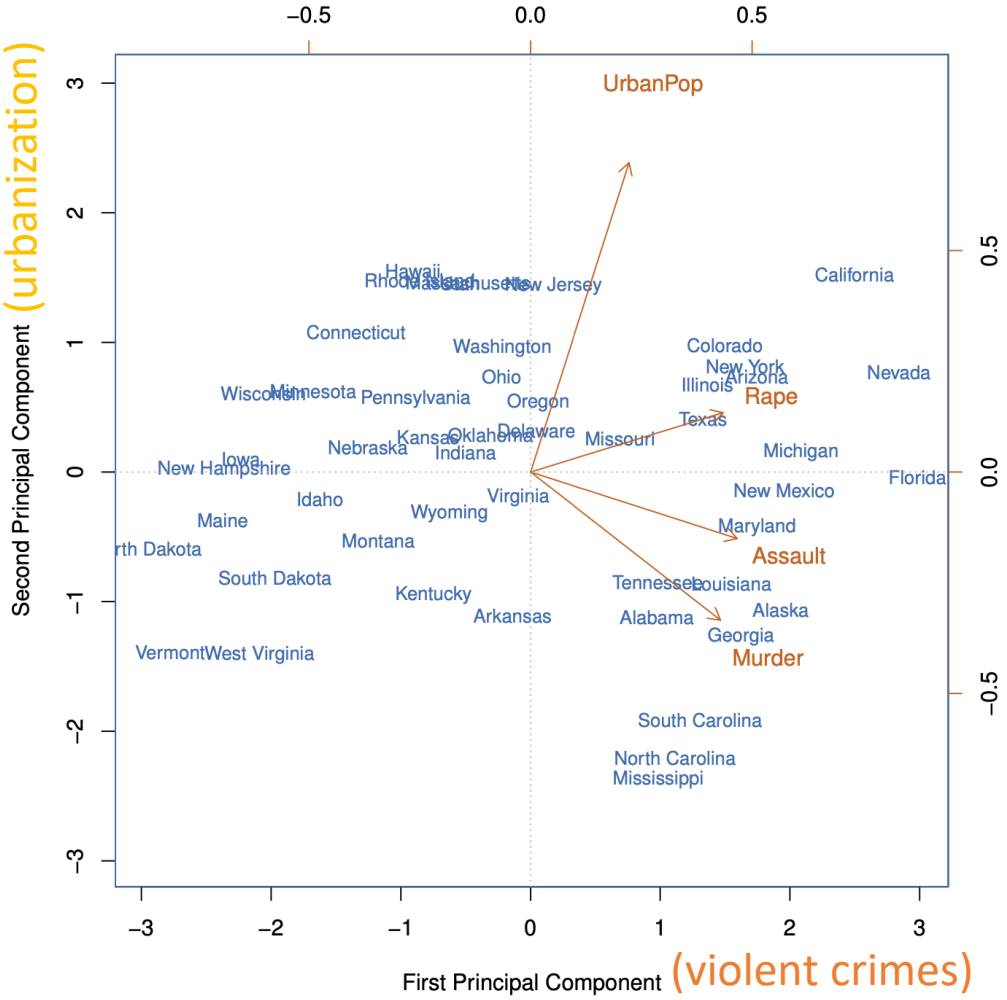
Variable	Description
Murder	Number of arrests for murder per 100,000 residents
Assault	Number of arrests for assault per 100,000 residents
Rape	Number of arrests for rape per 100,000 residents
UrbanPop	Percent of the population living in urban areas

- Data consists of 50 rows (one per each state) for the year of 1973
- No response variable
  - This is unsupervised data
  - You've learned about clustering in past classes, and we can also use PCA to better understand the data

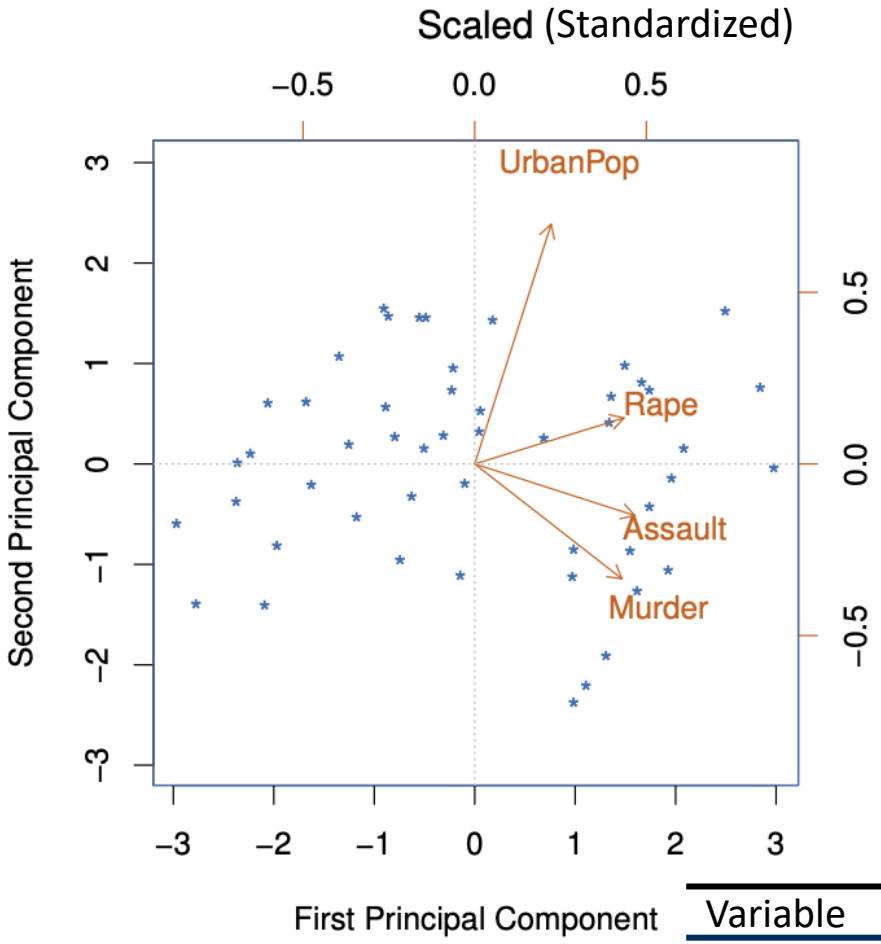
PCA Example

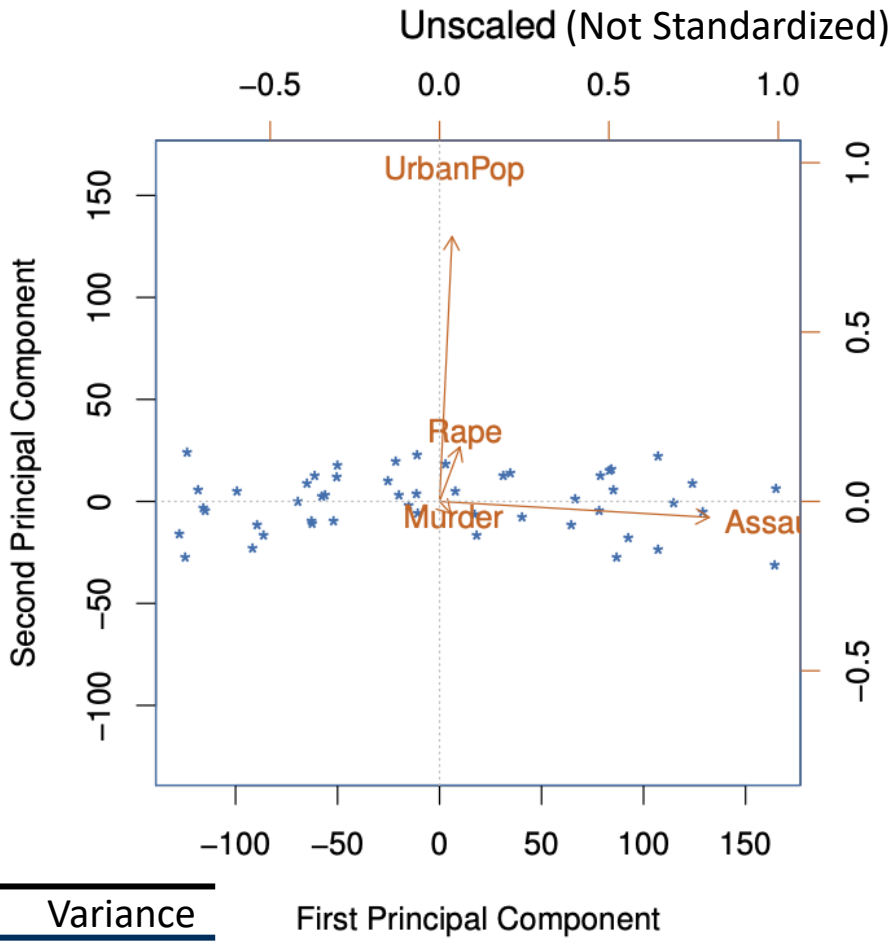
- PCA was performed after standardizing each variable
- PC_1 places almost equal weight on all three crimes
- PC_2 seems to focus mostly on the level of urbanization of the state
- We can interpret this as saying there are really 2 dimensions in this data (violent crimes and urbanization), compared to the original 4-dimensional data set

	PC1	PC2
Murder	0.5358995	−0.4181809
Assault	0.5831836	−0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186



The Importance of Standardizing



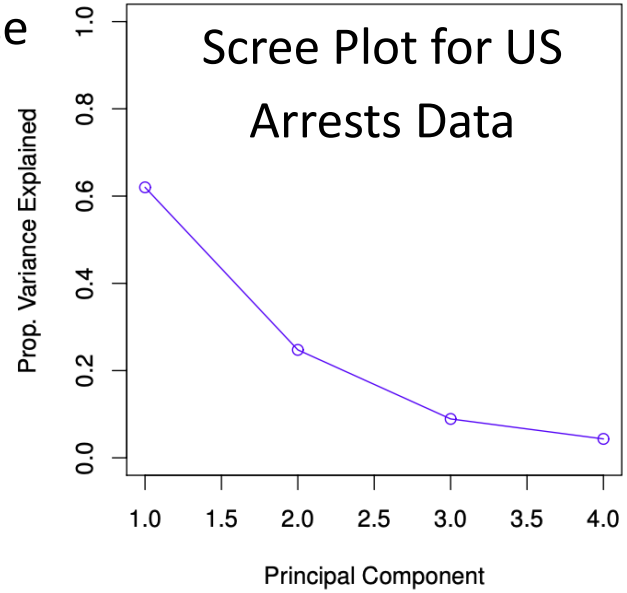


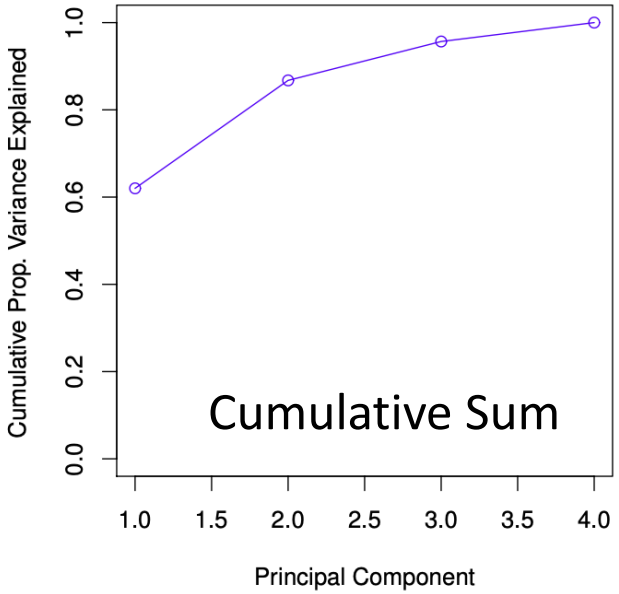
Variable	Variance
Murder	19.0
Assault	6945.2
Rape	87.7
UrbanPop	209.5

Since Assault has the largest variance, it will have the largest component unless scaled (we don't want this since units are playing a factor – we want to know which is most important regardless of the units in which it was measured)

Scree Plot

- A scree plot can be used to help us select the number of principal components to use
  - Shows the number of principal components (x-axis) vs the proportion of variance explained (y-axis)
  - The proportion of variance explained is the variance explained by the m^{th} principal component divided by the total variance in the data (predictors)
  - We look for the "elbow" in the plot to decide how many principal components to use





Principal Component Regression (PCR)

PCR

- Once you have selected a subset of the principal components $(\mathrm{PC}_1, \mathrm{PC}_2, \ldots \mathrm{PC}_M)$, they can be used to perform regression as usual – they become the new predictor variables
- By using the principal components, a smaller number of predictors from a high-dimensional dataset can be used
  - Helps avoid overfitting and, by construction, multicollinearity
  - What will the VIFs of the principal components be? 1, by design
- **The number of principal components used should be selected via cross-validation (preferred over the scree plot)**

Since we have supervised data (Y), we can do CV

- PCR does not perform feature selection since each principal component is a linear combination of **all** predictor variables
  - PCR is often very difficult to interpret since each principal component is itself a linear function

PCR Example 1: Continuous Response

Environmental Impact Data Set

- Recall from Module 5

We only have enough observations to support including at most 10 variables in the model

- To what extent do environmental conditions affect human mortality?

Different cities in the US

- $n = 60$

These three are what the researchers are interested in

Variable	Description
AnnPrecip	Mean annual precipitation
MeanJanTemp	Average January temperature (in degrees Fahrenheit)
MeanJulyTemp	Average July temperature (in degrees Fahrenheit)
PctGT65	Percent of population greater than 65 years old
PopPerHouse	Population per household
School	Median school years completed
PctSound	Percent of housing units that are "sound"
PopPerSqMile	Population per square mile
PctNonWhite	Percent of population that is nonwhite
PctWhiteCollar	Percent of employment in white-collar jobs
PctU20000	Percent of families with income under $20,000
log(Hydrocarbons)	Relative pollution potential of hydrocarbons
log(Nitrogen)	Relative pollution potential of oxides in nitrogen
log(SO2)	Relative pollution potential of oxides in sulfur dioxide
RelHumid	Annual average relative humidity
AAMort	Age-adjusted mortality

hypothermia

hyperthermia

Response

Regression Model with All Variables

	coef	std err	t	P>|t|	[0.025	0.975]
const	2089.5002	437.227	4.779	0.000	1208.327	2970.674
AnnPrecip	2.9056	0.851	3.415	0.001	1.191	4.620
MeanJanTemp	-3.1943	1.041	-3.067	0.004	-5.293	-1.095
MeanJulyTemp	-3.8649	2.064	-1.872	0.068	-8.025	0.295
PctGT65	-15.0148	7.767	-1.933	0.060	-30.667	0.638
PopPerHouse	-159.3303	66.434	-2.398	0.021	-293.219	-25.441
School	-18.8108	10.382	-1.812	0.077	-39.734	2.113
PctSound	-0.6080	1.606	-0.379	0.707	-3.845	2.629
PopPerSqMile	0.0036	0.004	0.931	0.357	-0.004	0.011
PctNonWhite	3.9988	1.279	3.126	0.003	1.421	6.577
PctWhiteCollar	0.0076	1.476	0.005	0.996	-2.967	2.982
PctU20000	0.8490	2.944	0.288	0.774	-5.084	6.782
log.Hydro	-35.4428	15.288	-2.318	0.025	-66.255	-4.631
log.Nit	53.9858	15.142	3.565	0.001	23.469	84.503
log.SO2	-8.1383	7.017	-1.160	0.252	-22.281	6.004
RelHumid	-0.1774	1.024	-0.173	0.863	-2.241	1.886

Metric	OLS All Variables
Adjusted R^2 (↑)	73%
BIC (↓)	635
AIC (↓)	601
MSE (↓)	1051
Assumptions Met	NO

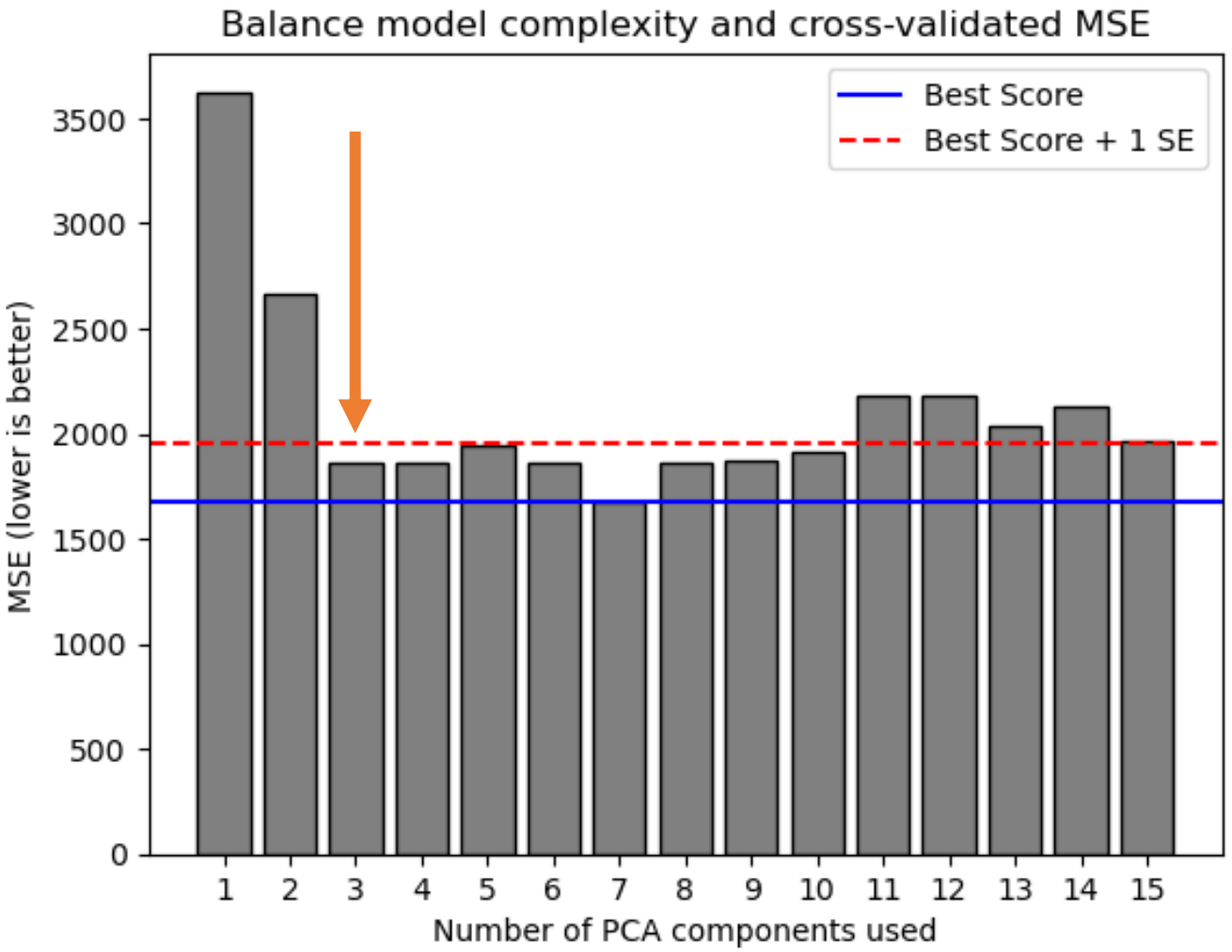
Regression Model From Module 5 PCR

	coef	std err	t	P>|t|	[0.025	0.975]
const	1002.2638	87.705	11.428	0.000	826.427	1178.101
AnnPrecip	2.2616	0.624	3.626	0.001	1.011	3.512
MeanJanTemp	-2.0340	0.487	-4.174	0.000	-3.011	-1.057
School	-13.9333	6.097	-2.285	0.026	-26.157	-1.710
PctNonWhite	3.7287	0.638	5.844	0.000	2.450	5.008
log.Nit	19.4866	4.356	4.474	0.000	10.754	28.219

	OLS All Variables	OLS Subset
Adjusted R^2 (↑)	73%	72%
BIC (↓)	635	608
AIC (↓)	601	595
MSE (↓)	1051	1081
Assumptions Met	NO	YES

PCR Model – Choosing the Number of Principal Components

- Similarly to LASSO and elastic net, we typically choose the smallest number of components that is within 1 standard error of the model with the "best" MSE



PCR Model with 3 Components

		coef	std err	t	P>|t|	[0.025	0.975]
	const	940.3585	5.329	176.472	0.000	929.684	951.033
PC_1	x1	-11.9099	2.526	-4.715	0.000	-16.970	-6.850
PC_2	x2	22.0243	3.214	6.853	0.000	15.587	28.462
PC_3	x3	-9.9552	3.355	-2.968	0.004	-16.675	-3.235

	OLS All Variables	OLS Subset	PCR 3 Components
Adjusted R^2 (↑)	73%	72%	56%
BIC (↓)	635	608	629
AIC (↓)	601	595	621
MSE (↓)	1051	1081	1704
Assumptions Met	NO	YES	YES

PCR Model with 7 Components

	coef	std err	t	P>|t|	[0.025	0.975]
const	940.3585	4.813	195.377	0.000	930.700	950.017
x1	-11.9099	2.281	-5.221	0.000	-16.488	-7.332
x2	22.0243	2.903	7.588	0.000	16.200	27.849
x3	-9.9552	3.030	-3.286	0.002	-16.035	-3.875
x4	2.5194	4.136	0.609	0.545	-5.780	10.819
x5	-0.9430	4.309	-0.219	0.828	-9.589	7.703
x6	1.3309	5.689	0.234	0.816	-10.085	12.747
x7	-23.4155	5.824	-4.021	0.000	-35.101	-11.730

	OLS All Variables	OLS Subset	PCR 3 Components	PCR 7 Components
Adjusted R^2 (↑)	73%	72%	56%	64%
BIC (↓)	635	608	629	629
AIC (↓)	601	595	621	612
MSE (↓)	1051	1081	1704	1390
Assumptions Met	NO	YES	YES	YES

PCR Model with 15 Components

	coef	std err	t	P>|t|	[0.025	0.975]
const	940.3585	4.186	224.669	0.000	931.923	948.794
x1	-11.9099	1.984	-6.003	0.000	-15.908	-7.912
x2	22.0243	2.524	8.725	0.000	16.937	27.112
x3	-9.9552	2.635	-3.778	0.000	-15.266	-4.645
x4	2.5194	3.597	0.700	0.487	-4.729	9.768
x5	-0.9430	3.747	-0.252	0.802	-8.494	6.608
x6	1.3309	4.947	0.269	0.789	-8.640	11.301
x7	-23.4155	5.064	-4.624	0.000	-33.622	-13.209
x8	-9.2624	6.373	-1.453	0.153	-22.107	3.582
x9	14.8663	8.905	1.669	0.102	-3.081	32.813
x10	-7.4036	9.409	-0.787	0.436	-26.366	11.559
x11	-8.9608	11.470	-0.781	0.439	-32.077	14.156
x12	-12.4351	12.299	-1.011	0.318	-37.222	12.352
x13	31.5809	13.351	2.365	0.022	4.674	58.488
x14	4.8121	18.591	0.259	0.797	-32.655	42.279
x15	-87.4877	25.310	-3.457	0.001	-138.498	-36.478

	OLS All Variables	OLS Subset	PCR 3 Components	PCR 7 Components	PCR 15 Components
Adjusted R^2 (↑)	73%	72%	56%	64%	73%
BIC (↓)	635	608	629	629	635
AIC (↓)	601	595	621	612	601
MSE (↓)	1051	1081	1704	1390	1051
Assumptions Met	NO	YES	YES	YES	YES

Code!

PCR Example 2: Binary Response

Breast Cancer Data Set

- Breast cancer is the most frequent cancer among women, impacting about 2.1 million women each year
- The goal is to predict breast cancer survival using clinical data and gene expression profiles
- n = 1904 and p = 311 after data cleaning

	patient_id	age_at_diagnosis	type_of_breast_surgery	cancer_type	cancer_type_detailed	er_status_measured_by_ihc	...	mtap_mut	ppp2cb_mut	smarcd1_mut	nras_mut
0	0	75.65	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	Positve	...	0	0	0	0
1	2	43.19	BREAST CONSERVING	Breast Cancer	Breast Invasive Ductal Carcinoma	Positve	...	0	0	0	0
2	5	48.87	MASTECTOMY	Breast Cancer	Breast Invasive Ductal Carcinoma	Positve	...	0	0	0	0
3	6	47.68	MASTECTOMY	Breast Cancer	Breast Mixed Ductal and Lobular Carcinoma	Positve	...	0	0	0	0
4	8	76.97	MASTECTOMY	Breast Cancer	Breast Mixed Ductal and Lobular Carcinoma	Positve	...	0	0	0	0

- Data was split into train/test sets

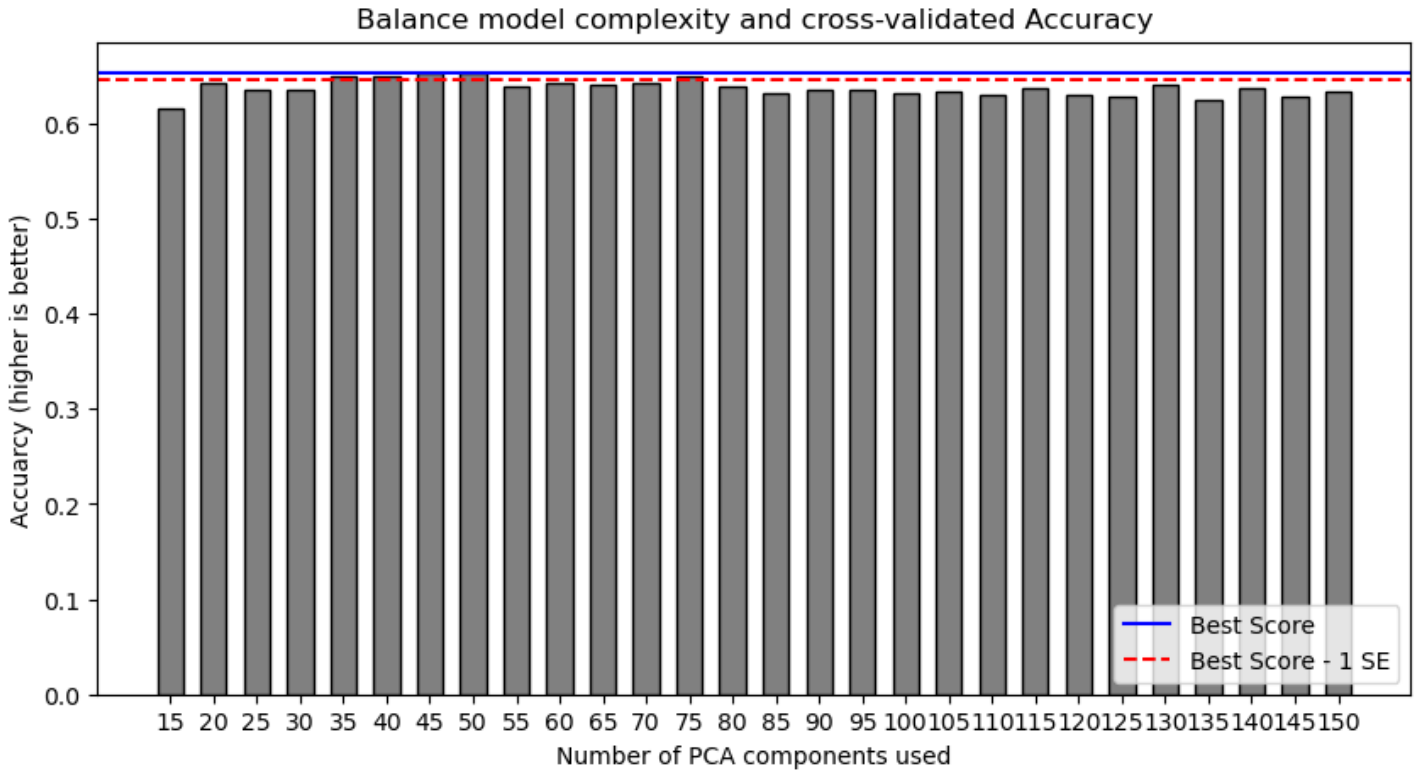
- Majority classifier/baseline accuracy rate on the test set: 59.58%

Logistic Regression Models

Model	Test Set Accuracy (↑)
Baseline	59.58%
Logistic LASSO	59.84%
Logistic Ridge	58.79%

- "Logistic LASSO" is logistic regression with LASSO (L1 penalty) applied
  - Of the 311 predictors, this model kept 283 predictors
  - A cutoff probability of 0.46 (obtained from the training set) was used to create a confusion matrix for the test set
- "Logistic Ridge" is logistic regression with ridge (L2 penalty) applied
  - A cutoff probability of 0.47 (obtained from the training set) was used to create a confusion matrix for the test set

PCR Model – Choosing the Number of Principal Components



- I only looked at 15 to 150 components in steps of 5 to save on computation time
- How many principal components would you choose?

PCR (Logistic Regression) Model

- "PCR 35 Components" is logistic regression with 35 variables that are principal components
  - A cutoff probability of 0.47 (obtained from the training set) was used to create a confusion matrix for the test set

Model	Test Set Accuracy (↑)
Baseline	59.58%
Logistic LASSO	59.84%
Logistic Ridge	58.79%
PCR 35 Components	64.83%

Code!