Ridge Regression

- **OLS** finds regression coefficients that **minimize** the SSE:

$$SSE = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - etc.)^2$$

- **Ridge** regression finds coefficients that **minimize**:

$$SSE(R) = SSE + \text{shrinkage penalty} = SSE + \lambda\,(\beta_1^2 + \beta_2^2 + \beta_3^2 + etc.)$$

- This seems like a complicated idea but the **concept** is simple:
  - ➢ **Ridge** regression fits a line that **minimizes SSE(R)**
  - ➢ That is, **Ridge** minimizes **SSE plus** a **penalty**
  - ➢ We can vary the penalty **λ** thus **controlling** the **shrinkage**
  - ➢ If we set **λ = 0**, Ridge minimizes **SSE** → same as **OLS**
  - ➢ If we set **λ very large**, then the resulting **β's** have to be **very small** → i.e., we **shrink** the coefficients
  - ➢ So if **λ = ∞** Ridge yields the **null model y = β_0**
  - ➢ The goal is to **select** the **λ** that **minimizes** the **Test MSE**

LASSO Regression

(Least Absolute Shrinkage and Selection Operator)

- Again, **OLS** finds regression coefficients that **minimize** the SSE:

$$SSE = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - etc.)^2$$

- **LASSO** regression finds coefficients that **minimize**:

$$SSE(L) = SSE + \textbf{shrinkage penalty} = SSE + \lambda\,(|\beta_1| + |\beta_2| + |\beta_3| + etc.)$$

- That is, the penalty **λ** is applied over the sum of the **absolute values** of the coefficients, **rather** than over the sum of their **squared values**
- The effect is **similar to Ridge** regression:
  - If we set **λ = 0**, LASSO minimizes **SSE** → same as **OLS**
  - If we set **λ = ∞**, LASSO yields the **null model y = β_0**
  - Again, the goal is to **select** the **λ** that **minimizes** the **Test MSE**
- One important and interesting **difference**: mathematically, the **Ridge coefficients** can **never** be shrunk to **0** (except when λ = ∞), but some **LASSO coefficients** do become exactly **0** eventually as λ increases
  → **LASSO** falls **in between Subset Selection** and **Ridge**.

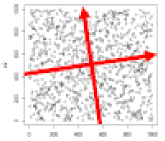
Dimension Reduction Methods

- The basic idea is if we have **P** somewhat correlated **predictors** it is possible to transform these into **M linear combinations**, such that **P > M**, thus **reducing** the number of **variables** in a model.
- **Dimension reduction =** reduce the estimation of **P+1** coefficients **(β_0, β_1, β_2,… β_P)** to estimating **M+1** coefficients **(α_0, α_1, α_2,… α_M)**
- ***Example:*** *if we suspect that a vehicle's volume, horsepower, and weight affect the vehicle's gas mileage, but these 3 variables are highly correlated, we could combine them into a new variable called something like "size" composed of some percentage of volume, plus some of horsepower, plus some of weight, reducing the model variables from 3 to 1.*
- Naturally, we also **lose** some **interpretability**, so it is a **tradeoff**
- Two popular dimension reduction methods are **Principal Components Analysis** (PCA) and **Partial Least Squares** (PLS), both of which use the **correlation matrix** of P predictors to find M (<P) linear combinations of the P predictors
- These methods are may **increase bias** but substantially **reduce variance** of the coefficients, particularly when **P** is **large relative to N**

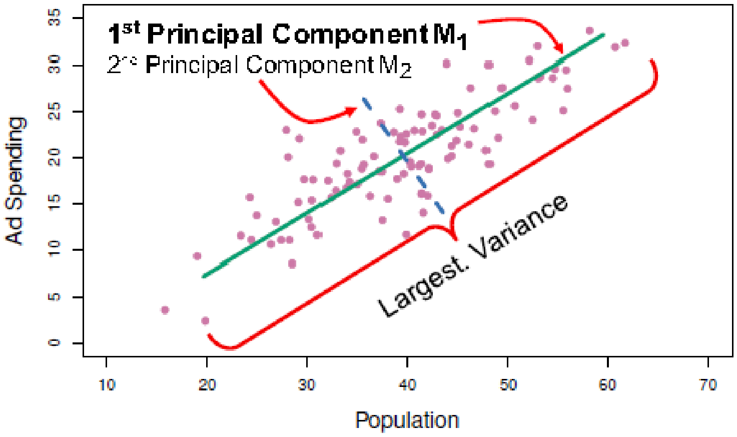
Principal Components: Illustration

This can be **illustrated** with **2 variables** like population size and advertising expenditures (as predictors of sales), resulting in **2 Principal Components**

PC's not helpful here



Think of Principal component as **rotating** the **axes** into the **highest variance** direction and then moving the **origin** to the mean (i.e., **centering**) of the variables involved.



It is clear from this plot that **Ad Spending** and **Population** are highly **correlated**. But Principal Components **M_1** and **M_2** are **not**

Variables, Components & Scores

- For a set of **P variables** → $X_1, X_2, \ldots\ldots X_P$
- There are **P "orthogonal" PC's**:

$$
\begin{aligned}
M_1 &= l_{11}X_1 + l_{12}X_2 + \cdots + l_{1P}X_P \\
M_2 &= l_{21}X_1 + l_{22}X_2 + \cdots + l_{2P}X_P \\
&\ldots\ldots \\
M_P &= l_{P1}X_1 + l_{P2}X_2 + \cdots + l_{PP}X_P
\end{aligned}
$$

- Each with a **sum** of **squared** factor **loadings** = 1:

$$
\begin{aligned}
l_{11}^2 + l_{12}^2 + \cdots + l_{1P}^2 &= 1 \\
l_{21}^2 + l_{22}^2 + \cdots + l_{2P}^2 &= 1 \\
&\ldots\ldots \\
l_{P1}^2 + l_{P2}^2 + \cdots + l_{PP}^2 &= 1
\end{aligned}
$$

- For a data point $X_{i1}, X_{i2}, \ldots\ldots X_{ip}$ the corresponding value m_{i1} is called the **PC "score"** for that data point

Principal Components Regression (PCR)

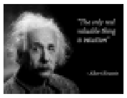
- The main purpose of **PCR** is to **reduce** the **dimensionality** (i.e., number of variables) of a model **without removing variables**.
- Since the PC's are sorted from highest to lowest variance, it is very likely that the **first few PC's** are **sufficient** to represent the variance in the data.
- So, for a given **OLS** regression:

$$Y = \beta_0 + \beta_1(X_1) + \beta_2(X_2) + \ldots\ldots + \beta_P(X_P) + \varepsilon$$

- We can construct **another OLS** model on M PC's:

$$Y = \alpha_0 + \alpha_1(M_1) + \alpha_2(M_2) + \ldots\ldots + \alpha_M(M_M) + \varepsilon$$

- **M** can be anywhere between **1** and **P** (all PC's → same as OLS)
- Since each PC is a linear transformation of all variables using factor loadings, **all variables** are **represented**
- The key is to find the **optimal** number of **M PC's** (i.e., a **"tuning"** parameter) to include in the model
- As with regular OLS, as **M ↑** → **Bias ↓** and **Variance ↑**



Partial Least Squares (PLS): Intuition

- With **PCR**, the X_1, X_2, …… X_p variables are transformed into M_1, M_2, …… M_p components in an **"unsupervised"** way
- That is, the independent variable dimensions are **rotated** to find directions in which the data exhibits highest variance.
- This is an **"unsupervised"** method because the **outcome** variable Y is not taken into account when doing PCA
- While PCR does well in general, there is **no guarantee** the first few M **components** will be the **best directions** to predict **Y**
- In contrast, PLS is a "supervised" method
- Like PCR, **PLS** is a **dimension reduction** method, but unlike PCR, PLS does **further rotation** of the dimensions to maximize the **correlation** with **Y**
- In sum, PLS attempts to find directions that **not only explain** the **predictors**, but also the **outcome** variable
- It does this by placing stronger weight on variables that are more strongly correlated with Y