UNIVERSITY OF ARKANSAS

DASC 4113 Machine Learning

Lecture 6

Ukash Nakarmi

Linear Model Selection and Regularization

Learning Objectives

In this class, we will learn how to:

- Improve the linear models by enforcing some constraints on the parameters(β)

Preface

In Linear Regression, our learning model was of the form:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

And: We learn parameter β by least square criteria between true (Y) and predictions ($\hat{Y}$).

- Does it always have to be least squares?
- Can we improve the prediction and interpretability of model by some alternate criterion?

Note: We take linear regression as an example, but our discussion will apply to other linear models as well.

Prediction Accuracy and Model Interpretability

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

p : Number of Inputs

n : Number of examples (Training data size)

Relation between Prediction Accuracy, Data Size and Number of Inputs:

Case 1: n >>p (n is lot larger than p, training data size is very large)

- Not much variability in fitted model using least squares. Predictions are good enough.

Case 2: n >p (n is not much larger than p, training data size is large)

- Variability in fitted model using least squares. Predictions are not accurate.

Case 3: n < p (n is smaller than p, training data size is very small)

- Solution is not unique. Prediction are incorrect. Variability is infinite

Prediction Accuracy and Model Interpretability

Relation between Model Interpretability and Numbers of Inputs (p):

- As p increases, the interpretability starts getting complex.

- The number of inputs (p) has role on both accuracy and interpretability of the model.
- But we know not all inputs have relation same degree of relation with the response (Y).

(i.e. Not all predictors are equally relevant to response).

So,
If we could find a way to select the inputs (Feature Selection/Variable Selection) that are more relevant to the response:

=> We can improve Model Accuracy and Model Interpretability

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

Find a Way ~ Find a Way to make some constraints on β

Feature Selection Techniques

Three Approaches:

- Subset Selection : Uses subset of p predictors that we believe to be more relevant, and model is fitted on reduced p.

- **Shrinkage:** Coefficients (β) of each input are shrunken towards zero. (Regularization)

- **Dimension Reduction:** Involves projecting p predictors into M-dimension where $M < p$. Then use M projections as predictors instead of original p predictors.

Shrinkage Methods

- Ridge Regression
- LASSO

Ridge Regression

In Linear Regression, our criterion for estimation coefficients (β) was to minimize:

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

In Ridge Regression, our criterion is to minimize:

Tuning parameter. Controls the relative impact

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \lambda \underbrace{\sum_{j=1}^{p} \beta_j^2}$$

Shrinkage Penalty . Not applied to j =0 (intercept)

Ridge Regression

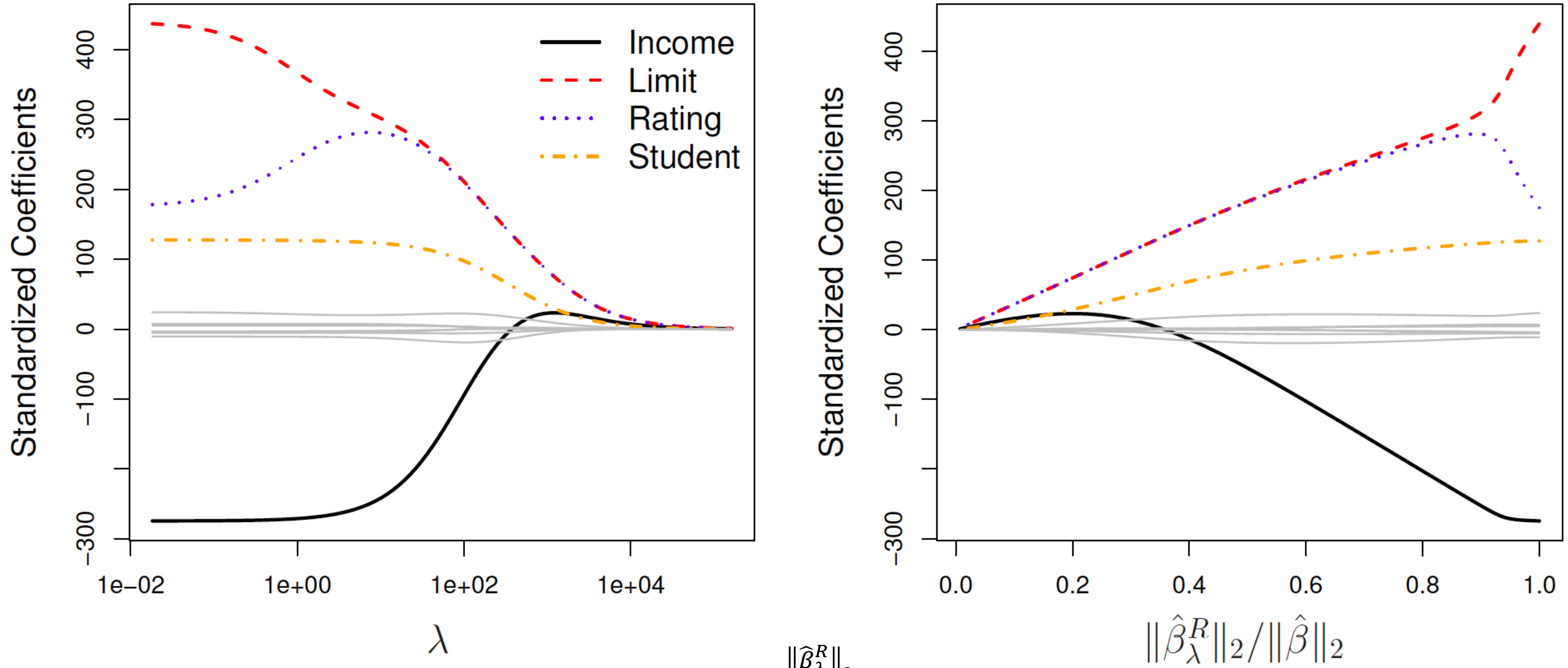


Fig: 6.4 Characteristics of change in coeffs w.r.t λ and $\frac{\|\hat{\beta}_\lambda^R\|_2}{\|\hat{\beta}\|_2}$

$\hat{\beta}$: *Coeff using Least square Criterion*

$\hat{\beta}_\lambda^R$: Coeff using Ridge Regression with tuning parameter value λ

$$\|\beta\|_2 = \sqrt{\sum_{j=1}^{p} {\beta_j}^2}$$

l2 norm: Measures the distance of a vector from origin

Practical Consideration for Ridge Regression

To use Ridge Regression:
We shall standardize each data point (predictors) because the coeff. of Ridge regression are no longer scale-invariant.

$$\tilde{x}_{ij} = \frac{x_{ij}}{\underbrace{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}}_{\text{Estimated Standard deviation of predictors}}}$$

The Lasso

LASSO Regression:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

$$\|\beta\|_1 = \sum|\beta_j|$$

l1 Norm ~ Sum of absolute values of elements of a vector.
May enforce some coeffs to be absolutely equal to 0.
(can serve as a surrogate for l0 norm.)

The Lasso

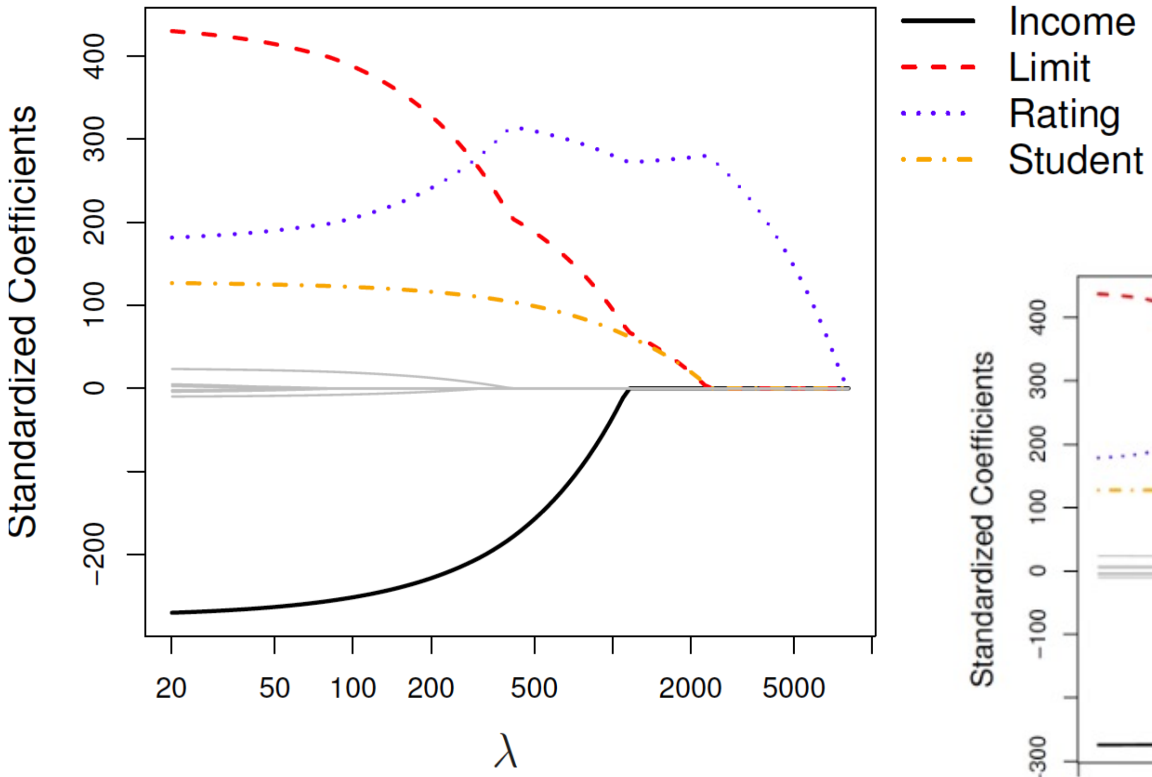
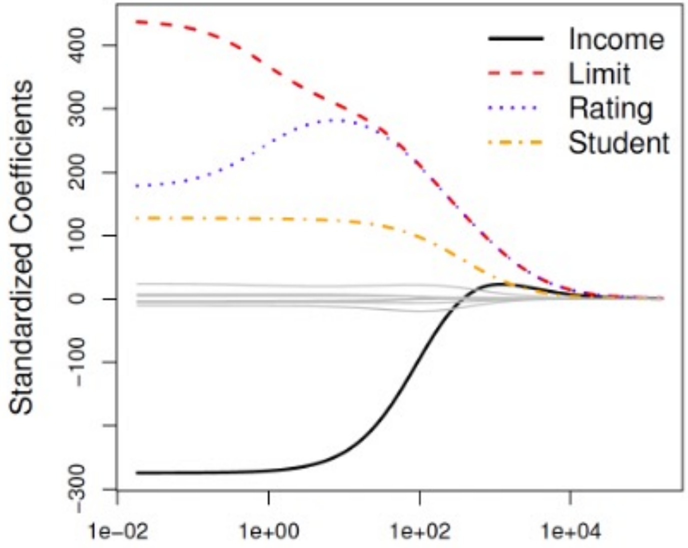


Fig: 6.6 LASSO Regression on credit data



Compare the decay of non relevant predictors (grey)in LASSO and Ridge

The Ridge (l2), Lasso (l1) and Subset selection (l0)

Ridge

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s,$$

Lasso

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

Subset Selection

Governs the total numbers of non-zero coeffs

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} I(\beta_j \neq 0) \leq s.$$

$I(\beta_j \neq 0)$: Indicator Variable,

= 1 if, $\beta_j \neq 0$

= 0; otherwise

The Ridge (l2), Lasso (l1) and Subset selection (l0)

Why l1 is a better approximation of l0 than l2 ? : Geometric Intuition

We will take an example when p = 2

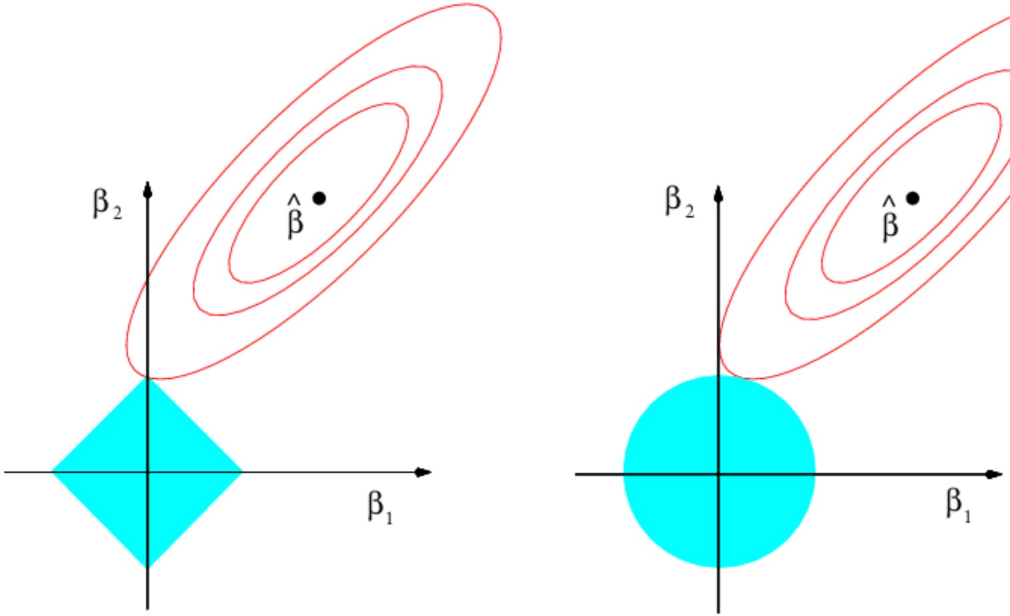


Fig: 6.7 Error contour and norm balls for l1 and l2 norm

Selection of tuning parameter (λ)

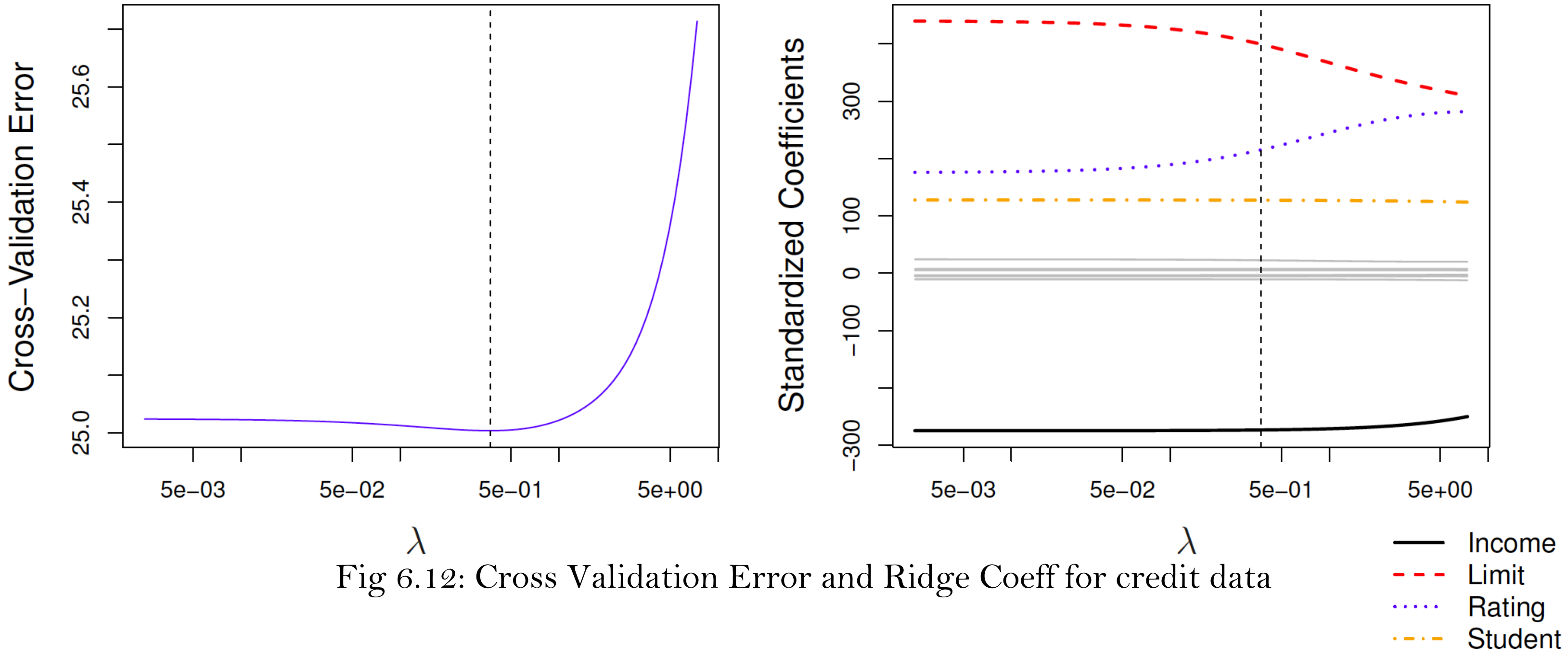


Fig 6.12: Cross Validation Error and Ridge Coeff for credit data

Choose λ that gives best cross validation error.

Feature Selection Techniques

Three Approaches:

- Subset Selection : Uses subset of p predictors that we believe to be more relevant, and model is fitted on reduced p.

- **Shrinkage:** Coefficients (β) of each input are shrunken towards zero. (Regularization)

- **Dimension Reduction:** Involves projecting p predictors into M-dimension where M < p. Then use M projections as predictors instead of original p predictors.

Dimension Reduction

- Represents p input variables as a linear combination of M new variables, $M<p$

Not necessarily always linear

Let $Z_1, Z_2 \ldots Z_M$ represent M linear combination of p variables such that:

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

Dimension Reduction

- Represents p input variables as a $\underbrace{\text{linear combination}}_{\text{Not necessarily always linear}}$ of M new variables, M<p

Let $Z_1, Z_2 \ldots Z_M$ represent M linear combination of p variables such that:

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

For now, lets assume we know how to find these transformation Φ
We will look at two ways to find the transformation.

Dimension Reduction

- Now the original linear regression problem can be viewed as :

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, \quad i = 1, \ldots, n,$$

Find Parameter $\boldsymbol{\theta}$ $i.e.$ $(\theta_0, \theta_1 \ldots \theta_m)$ such that Squared error between y and prediction of $\hat{y}$ is minimized.

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{jm} x_{ij} = \sum_{j=1}^{p} \underbrace{\sum_{m=1}^{M} \theta_m \phi_{jm}}_{} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij}$$

Dimension Reduction

- Now the question is given n training examples, how do we find the transformation :

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

One way to do so is by Principal Components

- Popular way to reduce the dimension of $n\times p$ data matrix.
- Unsupervised approach.
- Captures the direction in which data has higher variation.

Principal Component

Example:

Population vs advertisement data

100 data points (n = 100, p =2)

Note:

- Both Ad spending and Population are input variables in this example.
- The green line is not a linear regression line.
- Goal is to represent two-dimensional data (pop, ad spending) (x,y) using 1 D.
- In other words, find the direction of maximum variance or line closest to the data.

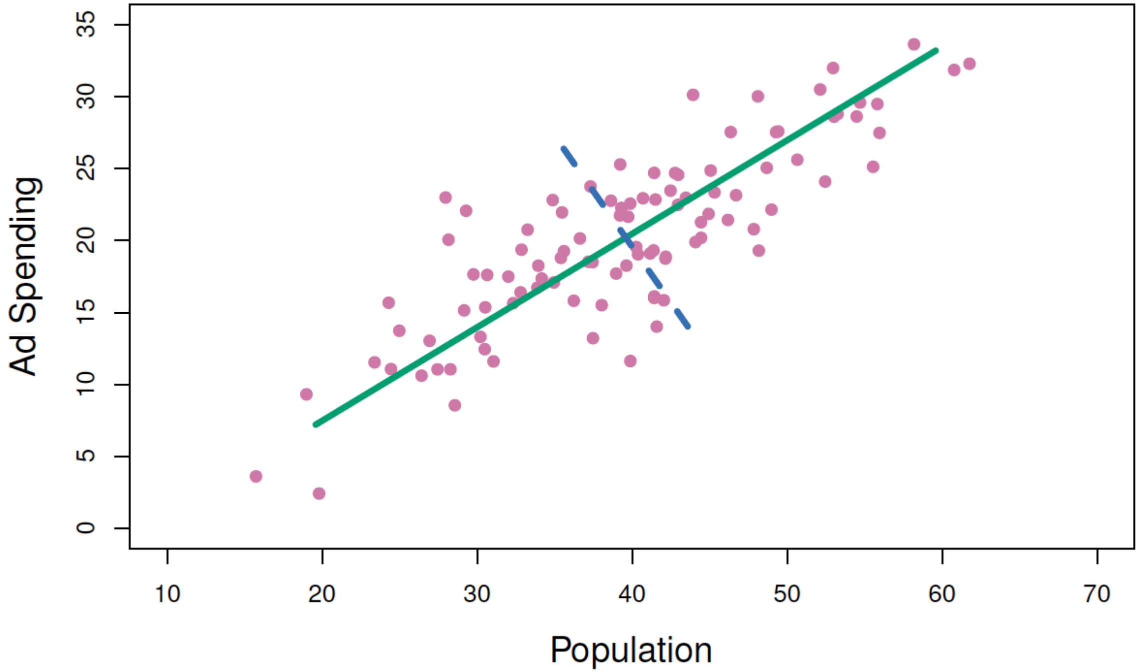
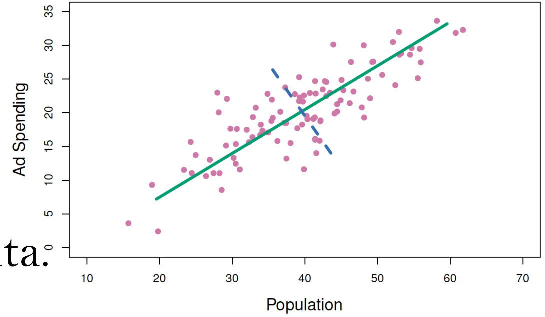


Fig: 6.14, Population and Advertisement spending data on some market

Principal Component

Find the direction of maximum variance

- Project the data into new direction.
- Coeff of Projection is a new representation of the data.



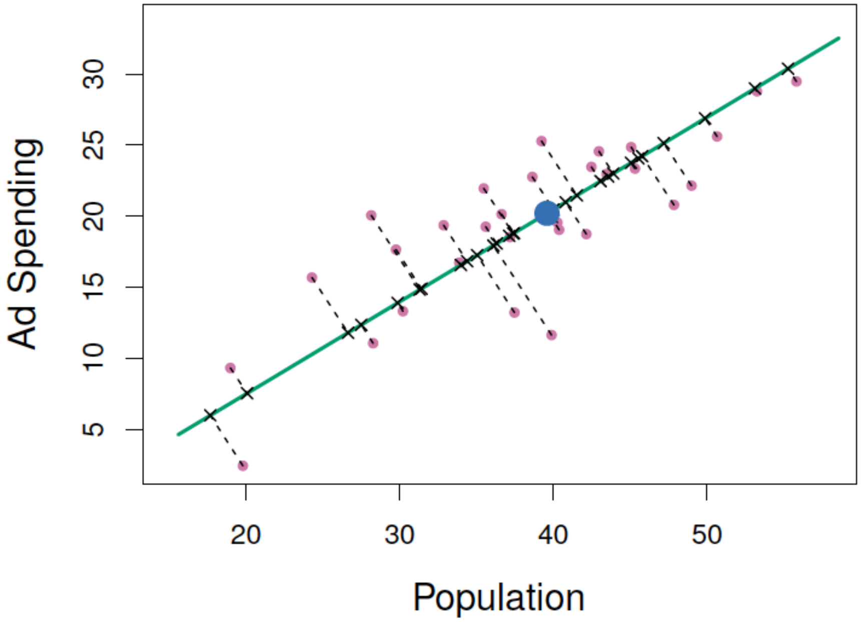


Fig: 6.15: Projections into PCs and distances for subset of samples.

Principal Component

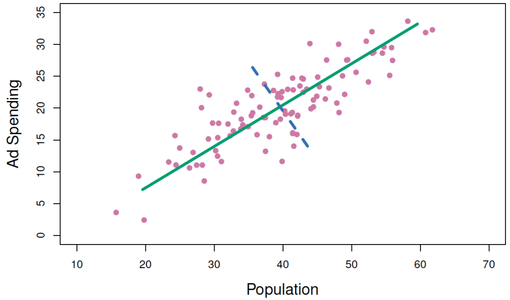
$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

Recall: We are looking for coeffs. ϕ *and* z

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, \quad i = 1, \ldots, n,$$

In this example:

$$\phi_{11} = 0.839, \;\; \phi_{21} = 0.544$$



Hence, for ith training data point x_i

$$z_{i1} = 0.839 \times (\mathtt{pop}_i - \overline{\mathtt{pop}}) + 0.544 \times (\mathtt{ad}_i - \overline{\mathtt{ad}})$$

- New 1D variable expressed as linear combination of two input variables: pop and adv.
- Our original Linear Regression with two input variables changes to LR with 1 input.
- Some constraints on $\phi_{j,m}$

$$\phi_{1,m}^2 + \phi_{2,m}^2 + \cdots \phi_{p,m}^2 = 1. \;\; (0.839^2 + 0.544^2 = 1)$$

$$\phi_{j,m1}^T \; \phi_{j,m2} = 1 \quad \text{(Orthonormality)}$$

Dimension Reduction:

Why Dimension Reduction is important in some Regression Problems?

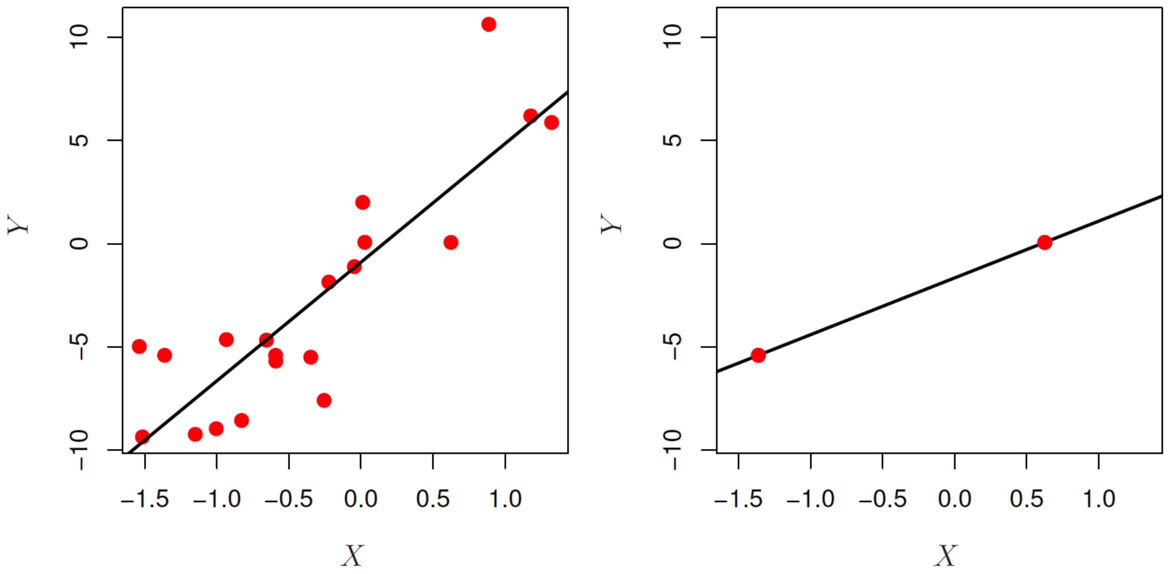


Fig: 6.22, Example: Linear Regression when n<<p

For Linear Regression (and other data-based learning models) to be reliable, n>>p

Model Selection and Regularization

Take away points:

- Subset selection, shrinkage and dimension reduction are three key approaches to model selection and regularization.
- L1 constrained Linear Regression gives closest approximation of subset selection.
- When n<<p, it is recommended we do dimension reduction, shrinkage to make data-based learning models reliable.