Predictive Analytics

Week 8: Linear Methods for Regression II

Semester 2, 2018

Discipline of Business Analytics, The University of Sydney Business School

Week 8: Linear Methods for Regression II

1. Introduction

2. Principal components regression

3. Partial least squares (optional)

4. Illustration and discussion

5. Considerations in high dimensions

6. Robust regression

Reading: Chapters 6.3 and 6.4 of ISL.

Exercise questions: Chapter 6.8 of ISL, Q5 and Q6.

Introduction

Dimension reduction methods (key concept)

Dimension reduction methods consist of building $M < p$ transformed variables which are linear combinations (projections) of the predictors. We then fit a linear regression of the response on the new variables.

Dimension reduction methods

Given the original predictors $x_1, x_2, \dots, x_p$, we let $z_1, z_2, \dots, z_M$ represent $M < p$ linear combinations of the original predictors, that is

$$\boldsymbol{z}_m = \sum_{j=1}^{p} \phi_{jm} \boldsymbol{x}_j,$$

for some constants $\phi_{1m}, \phi_{2m}, \dots, \phi_{pm}$, $m = 1, \dots, M$ to be determined.

Dimension reduction methods

We consider a linear regression model for the transformed predictors

$$Y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i,$$

which we fit by OLS.

We therefore estimate only $M + 1 < p + 1$ regression parameters, reducing variance compared to OLS.

Dimension reduction methods

The model for the transformed predictors implies a model for the original predictors:

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right) = \sum_{j=1}^{p} \beta_j x_{ij},$$

where

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}.$$

Dimension reduction is therefore a constraint on the original linear regression model. The cost of imposing this restriction is bias.

Dimension reduction methods

- The reduction in variance compared to OLS can be substantial when $M << p$.

- If $M = p$ and Z_m are linearly independent, no dimension reduction occurs and dimension reduction methods are equivalent to OLS on original p predictors.

- Dimension reduction methods can be useful when $p > N$.

Principal components regression

Principal components analysis (key concept)

Principal Component Analysis (PCA) is a popular way of deriving a set of low dimensional set of features from a large dimensional set of variables. In our setting, we want to use PCA to reduce the dimension of the $N \times p$ design matrix $\boldsymbol{X}$.

In our discussion below, we assume that we first center and standardise all the predictors.

Principal components (key concept)

We define the first **principal component** of $\boldsymbol{X}$ as the linear combination

$$\boldsymbol{z}_1 = \phi_{11}\boldsymbol{x}_1 + \phi_{21}\boldsymbol{x}_2 + \ldots + \phi_{p1}\boldsymbol{x}_p,$$

such that $\boldsymbol{z}_1$ has largest sample variance among all linear combinations whose coefficients satisfy $||\phi_1||_2^2 = \sum_{j=1}^{p} \phi_{j1}^2 = 1$.

Principal components (key concept)

The m-th principal component of $\boldsymbol{X}$ is the linear combination

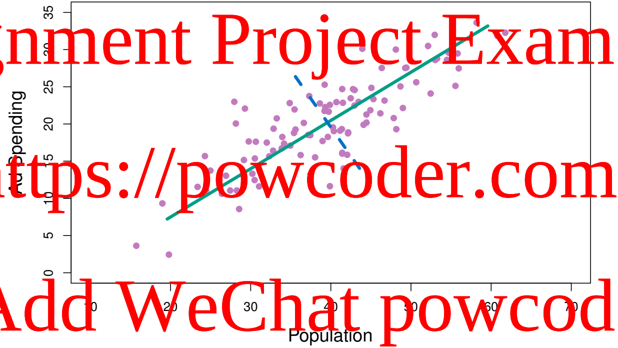
$$\boldsymbol{z}_m = \sum_{j=1}^{p} \phi_{jm} \boldsymbol{x}_j,$$

such $\boldsymbol{z}_m$ has largest sample variance among all linear combinations that are orthogonal to $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{m-1}$ and satisfy $||\phi_m||_2^2 = 1$.

Principal components (key concept)

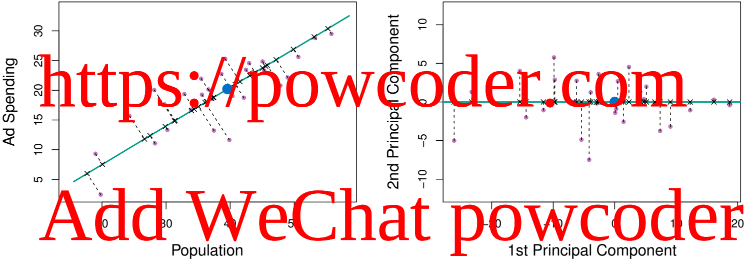
The first m principal components of the design matrix $\mathbf{X}$ provide the best m-dimensional linear approximation to it, in the sense of capturing variation in the predictor data.

Principal components analysis

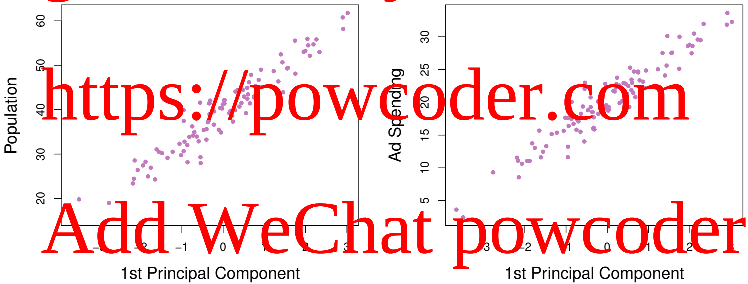


The two axes represent predictors. The green line indicates the first principal component and the blue dashed line shows the second principal component.

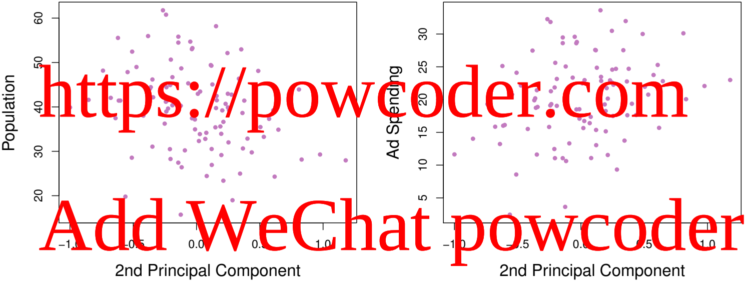
Principal components analysis



Principal components analysis



Principal components analysis



Principal Components Regression

The **principal components regression** (PCR) method consists of running a regression of Y on the first m principal components of $\boldsymbol{X}$.The PCR method implicitly assumes that directions of highest variance in $\boldsymbol{X}$ are the ones most associated with the response.

Principal components regression

Algorithm 1 Principal components regression

```
Center and standardise the predictors.
Use PCA to obtain $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$, the $p$ principal components of the design matrix $\boldsymbol{X}$.
for $m = 1, \ldots, p$ do
    Regress the response $\boldsymbol{y}$ on $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ (the first $m$ principal components) by OLS and call it $\mathcal{M}_m$.
end for
Select the best model out of $\mathcal{M}_1, \ldots, \mathcal{M}_p$ by cross-validation.
```

Principal Component Regression

- PCR can lead to substantial variance reduction compared to OLS when a small number of components account for a large part of the variation in the predictor data.

- Additional principal components leads to smaller bias, but larger variance.

- In PCR, the number of principal components is typically chosen by cross-validation.

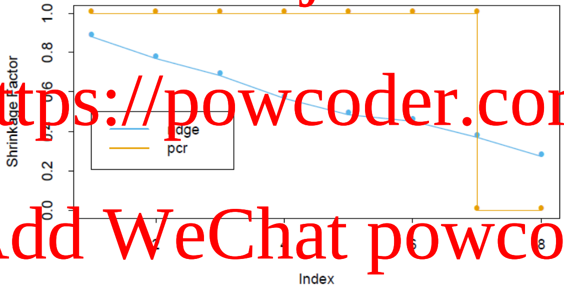
- Try to sketch learning cure for PCR (Train & Validation (Test) MSE vs Number of Components).

- PCR does not perform variable selection.

Comparison with ridge regression

- There is a close connection between the ridge regression and PCR methods.
- Ridge regression shrinks the coefficients of all principal components, with least shrinkage for the first component progressively smaller shrinkage factors for subsequent components.
- PCR leaves the components with largest variance alone and discards the ones with smallest variance.
- We can therefore think of ridge regression as a continuous version of PCR. Ridge may be preferred in most cases as it shrinks smoothly.

Comparison with ridge regression



Partial least squares (optional)

Partial Least Squares

The **partial least squares** method (PLS) tries to identify the best linear combinations of predictors in a *supervised* way, in a sense that it takes into account the information in $\boldsymbol{y}$ to construct the new features. When constructing z_m, PLS weights the predictors by the strength of their univariate effect on $\boldsymbol{y}$.

That contrasts with PCA, which identifies promising directions in an *unsupervised* way.

Partial Least Squares

Algorithm 2 Partial Least Squares (Initialisation)

```
Center and standardise the predictors.
Run $p$ simple linear regressions of $\boldsymbol{y}$ on each predictor $\boldsymbol{x}_j$ and denote the associated coefficients as $\phi_{1j}$.
Compute the first direction $\boldsymbol{z}_1 = \sum_{j=1}^{p} \phi_{1j} x_j$.
Run a SLR regression of $\boldsymbol{y}$ on $\boldsymbol{z}_1$ and let the coefficient be $\widehat{\theta}_1$. Call this model $\mathcal{M}_1$.
Orthogonalise each predictor with respect to $\boldsymbol{z}_1$: $\boldsymbol{x}_j^{(1)} = x_j - \boldsymbol{x}_j \left[ (\boldsymbol{x}_j^T \boldsymbol{z}_1)/(\boldsymbol{x}_j^T \boldsymbol{x}_j) \right]$. These are the residuals of a SLR of $\boldsymbol{x}_j$ on $\boldsymbol{z}_1$. (continues on the next slide)
```

Partial Least Squares

Algorithm Partial Least Squares (continued)

```
for $m = 2, \dots, p$ do
    Run $p$ simple linear regressions of $\boldsymbol{y}$ on each $x_j^{(m-1)}$ and denote the associated coefficients as $\phi_{mj}$.
    Compute the new direction $\boldsymbol{z}_m = \sum_{j=1}^{p} \phi_{mj} x_j$.
    Run a SLR regression of $\boldsymbol{y}$ on $\boldsymbol{z}_m$ and let the coefficient be $\theta_m$. Call the linear regression model with response $\boldsymbol{y}$, inputs $\boldsymbol{z}_1, \dots, \boldsymbol{z}_m$, and estimated coefficients $\widehat{\theta}_1, \dots, \widehat{\theta}_m$ model $\mathcal{M}_m$.
    Orthogonalise each $\boldsymbol{x}_j^{(m-1)}$ with respect to $\boldsymbol{z}_m$: $\boldsymbol{x}_j^{(m)} = \boldsymbol{x}_j^{(m-1)} - \boldsymbol{x}_j \left[(\boldsymbol{x}_j^T \boldsymbol{z}_1)/(\boldsymbol{x}_j^T \boldsymbol{x}_j)\right]$.
end for
Select the best model out of $\mathcal{M}_1, \dots, \mathcal{M}_p$ by cross-validation.
```

PCR and PLS: discussion

- While PCR seeks directions with high variance, PLS seeks directions with high variance *and* correlation with response.

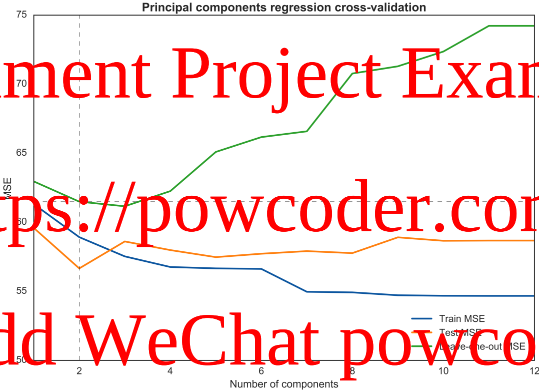
- The variance aspect tends to dominate, such that PLS behaves similarly to PCR and ridge regression.

- Using y reduces bias but potentially increases variance. PLS shrinks low variance directions, but can actually inflate high variance ones.

- In practice, PLS often does no better or slightly worse than PCR and ridge regression.

Illustration and discussion

Illustration: predicting the equity premium



Since the cross-validation performance is nearly identical with 2 and 3 components, we select $M = 2$ for the results below.

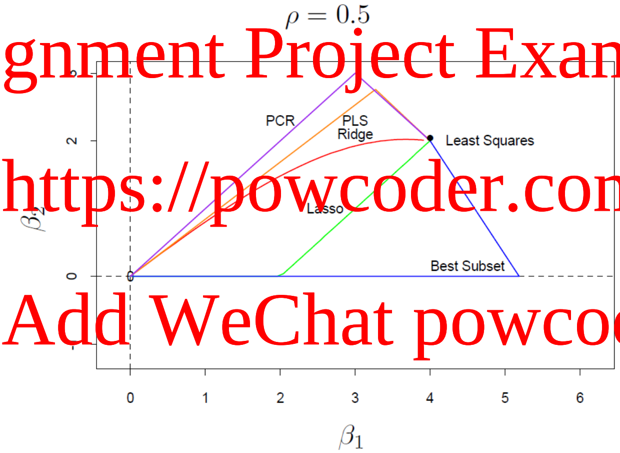
Dimension reduction methods

Equity premium prediction results

	Train R^2	Test R^2
OLS	0.108	0.014
PCR	0.039	0.048
PLS	0.085	0.036

For this example, PCR has the best test performance among all linear methods that we have discussed.

Comparison of shrinkage and selection methods



Considerations in high dimensions

Considerations in high dimensions

A **high-dimensional regime** occurs when the number of predictors is larger than the number of observations ($p > N$). Similar issues occur when $p \approx N$.

We cannot perform least squares in this setting, recall Lecture 2. If $p = N$, the training R^2 is always one. OLS is too flexible when $p > N$ and will overfit the data when $p \approx N$.

Example

Text analytics. In type of analysis, the predictors are often a large number of binary variables indicating the presence of words in a document, search history, etc. This is called a **bag of words** model. We thousands of possible words, the the number of predictors is very large in this type of analysis.

We can further extend the feature space to include **n-grams**, recording the appearance of words together in a sequence.

Considerations in high dimensions

- We can apply variable selection, shrinkage, and dimension reduction methods with carefully tuned hyper-parameters to high-dimensional settings.

- However, even these methods are subject to marked deterioration in performance as the number of irrelevant or very weak predictors increases relative to N.

- Therefore, we cannot blindly rely on standard methods in high dimensional regimes. We need to carefully consider appropriate dimension reduction and penalisation schemes, preferably based on understanding of the substantive problem.

- The next slide shows an example.

Supervised Principal Components

Algorithm 3 Supervised Principal Components

```
Center and standardise the predictors.
Run $p$ separate simple linear regressions of $\boldsymbol{y}$ on each individual predictor a record the estimated coefficients.
for $\theta$ in $0 \leq \theta_1 < \ldots < \theta_K$ do
    Form a reduced design matrix $\boldsymbol{X}_\theta$ consisting only of predictors whose SLR coefficient is higher than $\theta$ in absolute value.
    Use PCA to obtain $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$, the first $m$ principal components of $\boldsymbol{X}_\theta$.
    Use these principal components to predict the response.
end for
Select $\theta$ and $m$ by cross-validation.
```

Robust regression

Robust regression

All the linear regression methods that we have seen so far were based on the squared error loss function, which is equivalent to assuming a Gaussian likelihood for the data.

However, estimation based on the squared error loss can result in poor fit when there are **outliers**. This is because the squared error penalises deviations quadratically, so that points with larger residuals have more effect on the estimation than points with low residuals (near the regression line).

Robust regression

One way to achieve **robustness** to outliers is to replace the squared error losses with other losses that are less influenced by unusual observations.

Alternatively (and equivalently in some cases), we replace the Gaussian likelihood with that of a distribution with heavy tails. Such a distribution will assign higher likelihood to outliers, without having to adjust the regression fit to account for them.

Least absolute deviation

The **least absolute deviation** (LAD) estimator is

$$\widehat{\beta}_{\mathsf{lad}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left| y_i - \beta_0 - \sum_{j=1}^{p} \beta_i x_{ij} \right|$$

LAD estimation is equivalent to ML based on the **Laplace** distribution.

In the special case when we formulate the minimisation problem

$$\widehat{m} = \operatorname*{argmin}_{m} \sum_{i=1}^{n} |Y_i - m|$$

the LAD estimator $\widehat{m}$ is the sample median of the response.

Huber loss

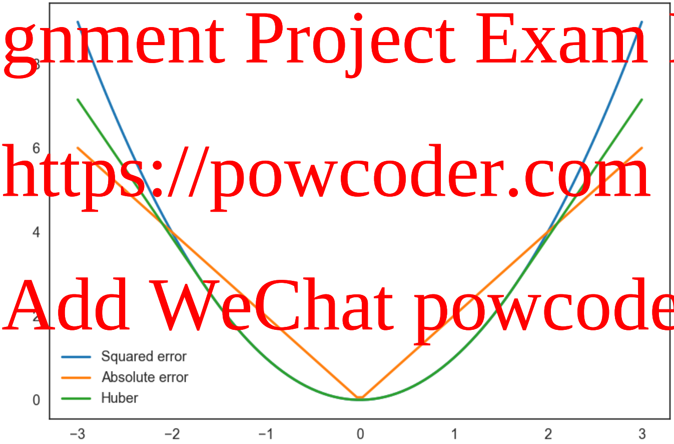
A popular method for robust regression is the **Huber loss**:

$$\widehat{\beta}_{\text{huber}} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{N} L_\delta \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_i x_{ij} \right) \right\}$$

$$L_\delta(e) = \begin{cases} e^2 & \text{if } |e| \leq \delta \\ 2\delta|e| - \delta^2 & \text{if } |e| \geq \delta \end{cases}$$

The Huber loss combines the good properties of squared and absolute errors.

Loss functions



Review questions

- What are dimension reduction methods for regression?
- What is principal components analysis (PCR)?

- What is the relationship between PCR and ridge regression?
- What is the high-dimensional regime?

- Explain the purpose of robust regression.

Technical appendix (Optional)

A column vector $\boldsymbol{v}$ is an **eigenvector** of a square matrix $\boldsymbol{A}$ if it satisfies the equation

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v},$$

where λ is a scalar known as the **eigenvalue** associated with $\boldsymbol{v}$. The eigenvectors of $\boldsymbol{A}$ do not change direction when multiplied by $\boldsymbol{A}$.

A scalar λ is an eigenvalue of $\boldsymbol{A}$ if $(\boldsymbol{A} - \lambda \boldsymbol{I})$ is singular:

$$\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0.$$

Principal components analysis

The **eigendecomposition** of a diagonalisable $p \times p$ symmetric real square matrix $\boldsymbol{A}$ has the form

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T,$$

where $\boldsymbol{\Lambda}$ is a $p \times p$ diagonal matrix whose diagonal elements $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ are the eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{V}$ is a $p \times p$ orthogonal matrix whose columns $\boldsymbol{v}_j$ are the eigenvectors of $\boldsymbol{A}$.

If one or more eigenvalues λ_j are zero then $\boldsymbol{A}$ is singular (non-invertible).

Principal components analysis

A orthogonal matrix $\boldsymbol{V}$ is a square matrix whose columns and rows are orthonormal, i.e.,

$$\boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I},$$

such that $\boldsymbol{V}^{-1} = \boldsymbol{V}^T$.

Principal components analysis

Now, let X be the $N \times p$ matrix of centered predictors. The sample variance-covariance matrix of $\boldsymbol{X}$ is

$$S = (\boldsymbol{X}^T\boldsymbol{X})/N,$$

where $\boldsymbol{X}^T\boldsymbol{X}$ has an eigendecomposition denoted as $\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T$. The eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}$ are all positive provided that there is no perfect multicollinearity. Eigenvalues near zero indicate the presence of multicollinearity.

Principal components analysis

The first principal component of $\boldsymbol{X}$ is

$$\boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{v}_1.$$

The sample variance of the first principal component is

$$s^2_{\boldsymbol{z}_1} = \frac{\boldsymbol{v}_1^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{v}_1}{N} = \frac{\boldsymbol{v}_1 \boldsymbol{V} \boldsymbol{\Lambda} \boldsymbol{V}^T \boldsymbol{v}_1}{N} = \frac{\lambda_1}{N},$$

where $\boldsymbol{v}_1^T \boldsymbol{v}_1 = 1$ and $\boldsymbol{v}_1^T \boldsymbol{v}_j = 0$ $(j \neq 1)$ since $\boldsymbol{V}$ is an orthogonal matrix. The first principal component is therefore the linear combination of the columns of $\boldsymbol{X}$ that has the largest variance among all possible normalised linear combinations.

Principal components analysis

The principal components of $\boldsymbol{X}$ are

$$\boldsymbol{z}_m = \boldsymbol{X}\boldsymbol{v}_m,$$

for $m = 1, \ldots, p$, with decreasing sample variance

$$s^2_{\boldsymbol{z}_m} = \frac{\lambda_m}{N},$$

since $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$. Since the eigenvectors are orthogonal, the principal components have sample correlation zero.

The principal component $\boldsymbol{z}_m$ is the direction of largest variance that is orthogonal to $z_1, \ldots, \boldsymbol{z}_{m-1}$.