Regression with Many Predictors

17.12.2014

Goals of Today's Lecture

- Get a (limited) overview of different approaches to handle data-sets with (many) more variables than observations.

Introduction

Example

- Can the concentration of a (specific) component be **predicted** from spectra?
- This sound like a regression problem. We have
  - a **response variable** Y (the concentration)
  - many **predictor variables** $x^{(1)}, \ldots, x^{(m)}$ (the spectrum)
- Hence, once we have a good model, we can hopefully **predict** the concentration based on the spectrum.
- This is especially useful if the spectrum is cheap and measuring the concentration is expensive.
- **But...**

- As is the case with spectra, we have **many predictor variables** (one for each wavelength).

- In such a situation, we typically have many more predictor variables than observations! ↯

- Hence, if we want to use all predictor variables, we **can't fit the model** because it would give a perfect fit.

- Therefore, we need methods that can deal with this new situation.

Stepwise Forward Selection of Variables

A simple approach is **stepwise forward regression**.

It works as follows:

- Start with empty model, only consisting of intercept.
- Add the predictor to the model that has the **smallest p-value**. For that reason fit all models with just one predictor and compare p-values.
- Add all possible predictors to the model of the last step, expand the model with the one with smallest p-value.
- Continue until some stopping criterion is met.

Pro's: Easy
Con's: Unstable: small perturbation of data can lead to (very) different results, may miss "best" model.

Principal Component Regression

Idea: Perform PCA on (centered) design matrix $\mathbf{X}$.

PCA will give us a "new" design matrix $\mathbf{Z}$. Use first $p < m$ columns. Perform an ordinary linear regression with the "new data".

Pro's

New design matrix $\mathbf{Z}$ is orthogonal (by construction).

Con's

We have **not** used Y when doing PCA. It could very well be that some of the "last" principal components are useful for predicting Y!

Extension

Select those principal components that have largest (simple) correlation with the response Y.

Ridge Regression

- Ridge regression "shrinks" the regression coefficients by adding a penalty to the least squares criterion.

$$\widehat{\underline{\beta}}_\lambda = \arg\min_{\underline{\beta}} \left\{ \|\underline{Y} - \mathbf{X}\underline{\beta}\|_2^2 + \lambda \sum_{j=1}^{m} \beta_j^2 \right\},$$

  where $\lambda \geq 0$ is a tuning parameter that controls the size of the penalty.

- The first term is the usual residual sum of squares.
- The second term penalizes the coefficients.
- **Intuition:** Trade-off between goodness of fit (first-term) and penalty (second term).

- There is a **closed form** solution

$$\underline{\widehat{\beta}}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\underline{Y},$$

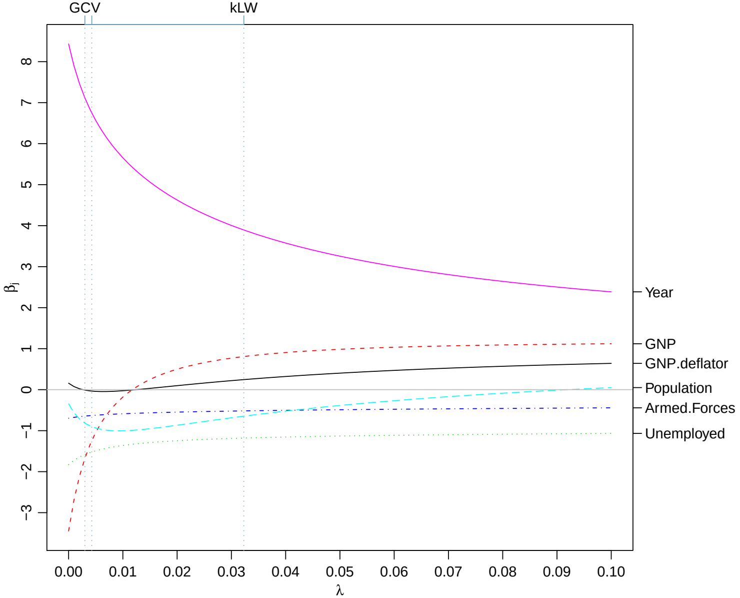
  where $\mathbf{I}$ is the identity matrix.

- Even if $\mathbf{X}^T\mathbf{X}$ is singular, we have a unique solution because we add the diagonal matrix $\lambda\mathbf{I}$.

- We can vary λ:
  - For $\lambda = 0$ we have the usual least squares fit (if it exists).
  - For $\lambda \to \infty$ we have $\underline{\widehat{\beta}}_\lambda \to \underline{0}$ (all coefficients shrunken to zero in the limit).

- This means, we can draw **paths** of coefficients (as a function of λ). At the end of the day we have to select a specific λ.

Illustration: Ridge Paths



Different curves are different coefficients.

Lasso

- Lasso = **L**east **A**bsolute **S**hrinkage and **S**election **O**perator.

- This is similar to Ridge regression, but "more modern".

$$\underline{\widehat{\beta}}_\lambda = \arg\min_{\underline{\beta}} \left\{ \|\underline{Y} - \mathbf{X}\underline{\beta}\|_2^2 + \lambda \sum_{j=1}^{m} |\beta_j| \right\},$$

- It has the property that it also **selects** variables, i.e. $\widehat{\beta}_{j,\lambda} = 0$ for large enough λ.

Statistical Consulting Service

Get help/support for

- **planning** your experiments.
- doing **proper analysis** of your data to answer your scientific questions.

Information available at

http://stat.ethz.ch/consulting