Regression Summary

X. Zeng

October 2015

1 Linear Regression

1.1 Model Introduction

Linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X.
Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares loss function as in ridge regression (L2-norm penalty) and lasso (L1-norm penalty).
Given a data set $\{y_i,\, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$ of n statistical units, a linear regression model assumes that the relationship between the dependent variable y_i and the p-vector of regressors x_i is linear. This relationship is modeled through a disturbance term or error variable ε_i — an unobserved random variable that adds noise to the linear relationship between the dependent variable and regressors. Thus the model takes the form

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where T denotes the transpose, so that $x_i^T \beta$ is the inner product between vectors x_i and β.

Often these n equations are stacked together and written in vector form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \mathbf{x}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \tag{3}$$

1.2 Major Assumptions

The following are the major assumptions made by standard linear regression models with standard estimation techniques (e.g. ordinary least squares):

- **Weak exogeneity**. This essentially means that the predictor variables x can be treated as fixed values, rather than random variables.
- **Linearity**. This means that the mean of the response variable is a linear combination of the parameters (regression coefficients) and the predictor variables.
- **Constant variance** (a.k.a. homoscedasticity). This means that different response variables have the same variance in their errors, regardless of the values of the predictor variables.
- **Independence of errors**. This assumes that the errors of the response variables are uncorrelated with each other.
- **Lack of multicollinearity in the predictors**. For standard least squares estimation methods, the design matrix X must have full column rank p,; otherwise, we have a condition known as multicollinearity in the predictor variables.

1.3 Estimation Methods

Some of the more common estimation techniques for linear regression are summarized below.

1.3.1 Least-squares estimation and related techniques

- **Ordinary least squares (OLS)** is the simplest and thus most common estimator.The general form is

$$Y = X\beta + \varepsilon, \qquad \mathrm{E}[\varepsilon|X] = 0, \qquad \mathrm{Var}[\varepsilon|X] = \sigma^2 I_n. \tag{4}$$

  The OLS method estimates the unknown parameters by minimizing the error sum of squares:

$$\hat{\beta} = \arg\min_{b\in\mathbb{R}^p} \sum_{i=1}^{n} (y_i - x_i^T b)^2 = \arg\min_{b\in\mathbb{R}^p} (y - Xb)^T (y - Xb), \tag{5}$$

  or equivalently in matrix form,

$$\hat{\beta} = (X^T X)^{-1} X^T y \ . \tag{6}$$

- **Generalized least squares (GLS)** is an extension of the OLS method, that allows efficient estimation of β when either heteroscedasticity, or correlations, or both are present among the error terms of the model, as long as the form of heteroscedasticity and correlation is known independently of the data. To handle heteroscedasticity when the error terms are uncorrelated with each other, GLS minimizes a weighted analogue to the sum of squared residuals from OLS regression, where the weight for the ith case is inversely proportional to $Var(\varepsilon_i)$. This special case of GLS is called "weighted least squares". The general form is

$$Y = X\beta + \varepsilon, \qquad \mathrm{E}[\varepsilon|X] = 0, \qquad \mathrm{Var}[\varepsilon|X] = \Omega. \tag{7}$$

$$\hat{\beta} = \arg\min_{b} (Y - Xb)' \, \Omega^{-1} (Y - Xb), \tag{8}$$

$$\hat{\beta} = (X'\Omega^{-1}X)^{-1} X'\Omega^{-1} Y. \tag{9}$$

  GLS can be viewed as applying a linear transformation to the data so that the assumptions of OLS are met for the transformed data.

1.3.2 Maximum-likelihood estimation and related techniques

- **Maximum likelihood estimation** can be performed when the distribution of the error terms is known to belong to a certain parametric

family f_θ of probability distributions. When f_θ is a normal distribution with zero mean and variance θ, the resulting estimate is identical to the OLS estimate.
Suppose there is a sample $x1, x2, \cdots, x_n$ of n independent and identically distributed observations, coming from a distribution with an unknown probability density function $f_0(\cdot)$. It is however surmised that the function f_0 belongs to a certain family of distributions $\{f(\cdot|\theta), \theta \in \Theta\}$ (where θ is a vector of parameters for this family), called the parametric model, so that $f_0 = f(\cdot|\theta_0)$. The value θ_0 is unknown and is referred to as the true value of the parameter vector. It is desirable to find an estimator $\hat{\theta}$ which would be as close to the true value θ_0 as possible. Either or both the observed variables x_i and the parameter θ can be vectors.
To use the method of maximum likelihood, one first specifies the joint density function for all observations. For an independent and identically distributed sample, this joint density function is

$$f(x_1, x_2, \ldots, x_n \mid \theta) = f(x_1 \mid \theta) \times f(x_2|\theta) \times \cdots \times f(x_n \mid \theta). \qquad (10)$$

Now we look at this function from a different perspective by considering the observed values $x1, x2, \cdots, x_n$ to be fixed "parameters" of this function, whereas θ will be the function's variable and allowed to vary freely; this function will be called the likelihood:

$$\mathcal{L}(\theta\,;\, x_1, \ldots, x_n) = f(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta). \qquad (11)$$

In practice it is often more convenient to work with the logarithm of the likelihood function, called the log-likelihood:

$$\ln \mathcal{L}(\theta\,;\, x_1, \ldots, x_n) = \sum_{i=1}^{n} \ln f(x_i \mid \theta), \qquad (12)$$

The method of maximum likelihood estimates θ_0 by finding a value of θ that maximizes the likelihood function.

- **Ridge regression** and other forms of penalized estimation such as **Lasso regression**, deliberately introduce bias into the estimation of β in order to reduce the variability of the estimate. The resulting estimators generally have lower mean squared error than the OLS estimates,

particularly when multicollinearity is present or when overfitting is a problem. They are generally used when the goal is to predict the value of the response variable y for values of the predictors x that have not yet been observed. These methods are not as commonly used when the goal is inference, since it is difficult to account for the bias.

Ridge regression places a particular form of constraint on the parameters (β's), $\hat{\beta}$ is chosen to minimize the penalized sum of squares:

$$\sum_{i=1}^{n}(y_i - \sum_{i=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{i=1}^{p} \beta_j^2 \tag{13}$$

The result is

$$\hat{\beta} = (X^T X - \lambda I_p)^{-1} X^T y \ . \tag{14}$$

Lasso regression is similar. That is

$$\sum_{i=1}^{n}(y_i - \sum_{i=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{i=1}^{p} |\beta_j| \tag{15}$$

The difference between Lasso and Ridge is that Lasso is L1-norm minimization, and Ridge is L2-norm minimization. In general, one might expect the lasso to perform better in a setting where a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or that equal zero. Ridge regression will perform better when the response is a function of many predictors, all with coefficients of roughly equal size. However, the number of predictors that is related to the response is never known a priori for real data sets. A technique such as cross-validation can be used in order to determine which approach is better on a particular data set.

References

- `https://en.wikipedia.org/wiki/Linear_regression`
- James, Witten, Hastie and Tibshirani. An Introduction to Statistical Learning.