Review

December 6, 2018

Linear model

- We assume that we are given data Y, X where for unit i, Y_i is the response variable. The i-th row of X is the contains the explanatory variables (covariates) of unit i and it is denoted by $x_i = (1, x_{i,1}, \ldots, x_{i,p})$. The observation $x_{i,j}$ is the measurement of unit i associated with j-th explanatory variable.

Linear model

- We assume that we are given data Y, X where for unit i, Y_i is the response variable. The i-th row of X is the contains the explanatory variables (covariates) of unit i and it is denoted by $x_i = (1, x_{i,1}, \ldots, x_{i,p})$. The observation $x_{i,j}$ is the measurement of unit i associated with j-th explanatory variable.
- We write the model in matrix notation as

$$Y|X \;=\; X\beta \;+\; e$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ is an unknown vector. We make the assumption that $\mathbb{E}(Y|X) = X\beta$, or $\mathbb{E}(e) = 0$.

Linear model

- We assume that we are given data Y, X where for unit i, Y_i is the response variable. The i-th row of X is the contains the explanatory variables (covariates) of unit i and it is denoted by $x_i = (1, x_{i,1}, \ldots, x_{i,p})$. The observation $x_{i,j}$ is the measurement of unit i associated with j-th explanatory variable.
- We write the model in matrix notation as

$$Y|X \;=\; X\beta \;+\; e$$

  where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ is an unknown vector. We make the assumption that $\mathbb{E}(Y|X) = X\beta$, or $\mathbb{E}(e) = 0$.
- We proposed to estimate it by minimizing the sum of squares of the residuals

$$\hat{\beta} \;=\; \underset{\beta \in \mathbb{R}^n}{\arg\min} ||Y - X\beta||^2$$

- We showed that
  1. If solution is unique then $\hat{\beta} = (X^T X)^{-1} X^T Y = HY$, so $\hat{\beta}$ is a linear estimator.
  2. Under the model above the ols estimators satisfies $\mathbb{E}(\hat{\beta}) = \beta$.
  3. If in addition, $cov(e) = \sigma^2 I$, then $cov(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$.
  4. Gauss-Markov theorem.
  5. Estimable.

- **How to test** $H_0 : \beta_j = 0$ p-value for testing $H_0 : \beta_j = 0$ can be got by

$$\mathbb{P}\left(|t_{n-p-1}| > \left| \frac{\hat{\beta}_j}{s.e(\hat{\beta}_j)} \right| \right).$$

- **How to test** $H_0 : \beta_j = 0$ p-value for testing $H_0 : \beta_j = 0$ can be got by

$$\mathbb{P}\left(|t_{n-p-1}| > \left| \frac{\hat{\beta}_j}{s.e(\hat{\beta}_j)} \right| \right).$$

- Also, we can use the fact that under the null hypothesis that a smaller model (m) holds, for testing an alternative larger model

$$\frac{RSS(m) - RSS(M)}{RSS(M)/(n-p-1)} \sim F_{1,n-p-1}.$$

One way ANOVA

- Consider the model

$$y_{ij} = \mu_i + e_{ij}, \qquad \text{for } i = 1, \ldots, t, \quad \text{and} \quad j = 1, \ldots, n_i$$

where e_{ij} are i.i.d normal random variables with mean zero and variance σ^2. Let $\sum_{i=1}^{t} n_i = n$.

One way ANOVA

- Consider the model

$$y_{ij} = \mu_i + e_{ij}, \qquad \text{for } i = 1, \ldots, t, \quad \text{and} \quad j = 1, \ldots, n_i$$

  where e_{ij} are i.i.d normal random variables with mean zero and variance σ^2. Let $\sum_{i=1}^{t} n_i = n$.
- The F-statistic for testing $H_0 : \mu_1 = \cdots = \mu_t$ is

$$T = \frac{\sum_{i=1}^{t} n_i(\bar{y}_i - \bar{y})^2/(t-1)}{\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2/(n-t)}$$

  which has the F-distribution with $t-1$ and $n-t$ degrees of freedom under H_0.

One way ANOVA

- Consider the model

$$y_{ij} = \mu_i + e_{ij}, \qquad \text{for } i = 1, \ldots, t, \quad \text{and} \quad j = 1, \ldots, n_i$$

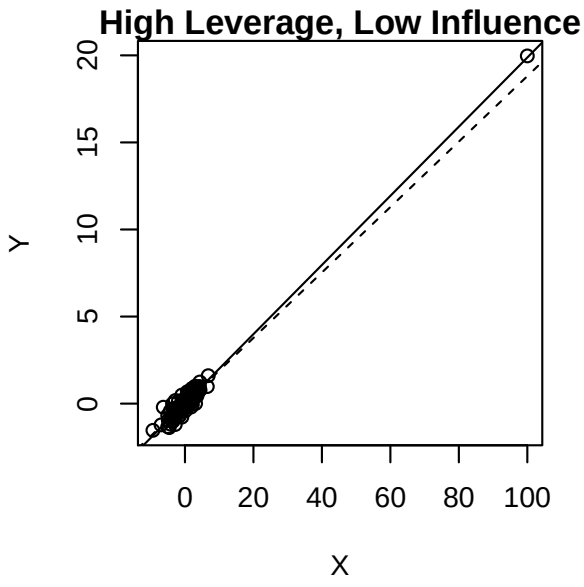
  where e_{ij} are i.i.d normal random variables with mean zero and variance σ^2. Let $\sum_{i=1}^{t} n_i = n$.
- The F-statistic for testing $H_0 : \mu_1 = \cdots = \mu_t$ is

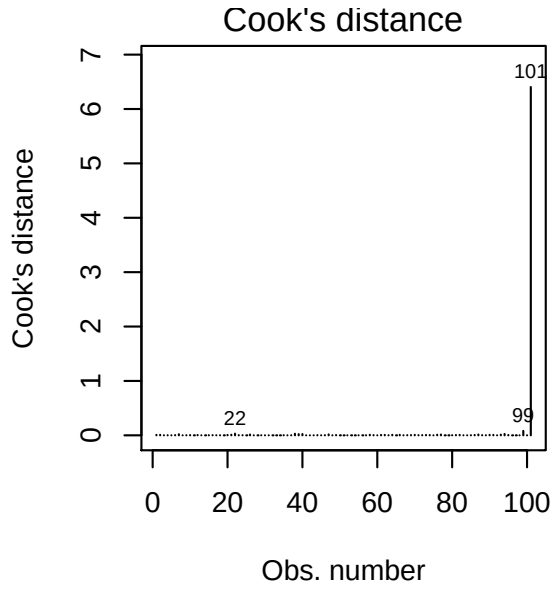
$$T = \frac{\sum_{i=1}^{t} n_i(\bar{y}_i - \bar{y})^2/(t-1)}{\sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2/(n-t)}$$

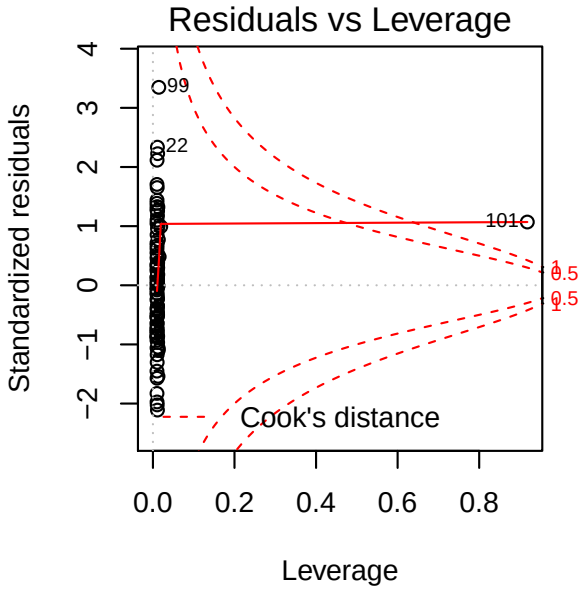
  which has the F-distribution with $t-1$ and $n-t$ degrees of freedom under H_0.
- Hat matrix is diagonal.

Regression diagnostics

- Residuals, standardized residuals, predicted residuals, standardized predicted residuals, levarage, Cook's distance.


High Leverage, Low Influence
Y
X
0
5
10
15
20
0
20
40
60
80
100


Cook's distance
Cook's distance
Obs. number
0
1
2
3
4
5
6
7
0
20
40
60
80
100
101
22
99


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
−2
−1
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
99
22
101
0.5
0.5
1
1

Variable selection

- Backward and forward selection.
- Adjusted R^2.
- BIC
- AIC
- Mallow's cp
- Cross-validation
- Leave one out cross-validation, generalized cross-validation.

GLM

- In GLMs, the response variables $y_1, \ldots, y_n$ can be either discrete (have pmfs) or continuous (have pdfs). It is assumed that $y_1, \ldots, y_n$ are independent.

GLM

- In GLMs, the response variables $y_1, \ldots, y_n$ can be either discrete (have pmfs) or continuous (have pdfs). It is assumed that $y_1, \ldots, y_n$ are independent.
- We also assume that the pmf or pdf of y_i can be modelled by two parameters θ_i and ϕ_i and can be written as

$$f(x; \theta_i, \phi_i) := h(x, \phi_i) \exp\left(\frac{x\theta_i - b(\theta_i)}{a(\phi_i)}\right). \tag{1}$$

θ_i is the main parameter (also called the canonical parameter). ϕ_i is called the dispersion parameter and one often assumes that ϕ_i is the same for all i. The function $b(\theta_i)$ is called the cumulant function.

- In GLM, we assume that $y_1, \ldots, y_n$ are independent with pmf or pdf of the form (1). We then write

$$g(\mu_i) := \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

for an increasing function g.

- In GLM, we assume that $y_1, \ldots, y_n$ are independent with pmf or pdf of the form (1). We then write

$$g(\mu_i) := \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

  for an increasing function g.
- This g is called the *link function*. In classical linear models, $g(\mu_i) = \mu_i$ which means that we have the identity link.

- In GLM, we assume that $y_1, \ldots, y_n$ are independent with pmf or pdf of the form (1). We then write

$$g(\mu_i) := \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

  for an increasing function g.
- This g is called the *link function*. In classical linear models, $g(\mu_i) = \mu_i$ which means that we have the identity link.
- The link function $g = (b')^{-1}$ is called the *canonical link function*. Recall that $(b')^{-1}(\mu_i) = \theta_i$. Thus GLM with the canonical link function models the canonical parameter θ_i as a linear function of the explanatory variables.

Regression and classification trees

- Given a variable X_j and a cut-off c, we can divide the subjects into two groups: G_1 given by $X_j \leq c$ and G_2 given by $X_j > c$. The *deviance* of this split is defined as:

$$RSS(j, c) := \sum_{i \in G_1} (y_i - \bar{y}_1)^2 + \sum_{i \in G_2} (y_i - \bar{y}_2)^2$$

where $\bar{y}_1$ and $\bar{y}_2$ denote the mean values of the response in the groups G_1 and G_2 respectively.

Regression and classification trees

- Given a variable X_j and a cut-off c, we can divide the subjects into two groups: G_1 given by $X_j \leq c$ and G_2 given by $X_j > c$. The *deviance* of this split is defined as:

$$RSS(j, c) := \sum_{i \in G_1} (y_i - \bar{y}_1)^2 + \sum_{i \in G_2} (y_i - \bar{y}_2)^2$$

  where $\bar{y}_1$ and $\bar{y}_2$ denote the mean values of the response in the groups G_1 and G_2 respectively.
- The values of j and c for which $RSS(j, c)$ is the smallest give the best split. The quantity $\min_{j,c} RSS(j, c)$ should be compared with $TSS = \sum_i (y_i - \bar{y})^2$. The ratio $\min_{j,c} RSS(j, c)/TSS$ is always smaller than 1 and the smaller it is, the greater we are gaining by the split.

- The recursive partitioning algorithm for constructing the regression tree proceeds as follows:

- The recursive partitioning algorithm for constructing the regression tree proceeds as follows:
  1. Find j and c such that $RSS(j, c)$ is the smallest.

- The recursive partitioning algorithm for constructing the regression tree proceeds as follows:
  1. Find j and c such that $RSS(j, c)$ is the smallest.
  2. Use the jth variable and the cut-off c to divide the data into two groups: G_1 given by $X_j \leq c$ and G_2 given by $X_j > c$.

- The recursive partitioning algorithm for constructing the regression tree proceeds as follows:
  1. Find j and c such that $RSS(j, c)$ is the smallest.
  2. Use the jth variable and the cut-off c to divide the data into two groups: G_1 given by $X_j \leq c$ and G_2 given by $X_j > c$.
  3. Repeat this process within each group separately.

Shrinkage

- We have the least squares criteria

$$\min_{\beta} ||y - \mathbf{X}\beta||^2$$

If we want to limit the contributions of variable $\boldsymbol{X}_j$ to the model, we can think of wanting $|\hat{\beta}_j|$ to be small. We don't know which variables we want to limit, though, so we want to write down some condition that is global on the vector β and then algorithmically let the data tell me which variables should get to contribute the most. One can think about these methods as smoother versions of variable selection that don't require 0/1 choices or as much user choices.

- Mathematically, we can write this as

$$\min_{\mathcal{S}(\beta)\leq c} ||y - \mathbf{X}\beta||^2$$

We can make different choices as to what is the allowable 'size' of β, e.g.

$$\mathcal{S}(\beta) = \sum_j \beta_j^2, \text{ or }, \mathcal{S}(\beta) = \sum_j |\beta_j|$$

As $c \to \infty$ we are putting on less constraint, so we get closer to the standard least squares model. Another way we can formulate this problem

$$\min_{\beta} ||y - \mathbf{X}\beta||^2 + \lambda\mathcal{S}(\beta)$$

In this way, we are adding a penalty for the size of β. λ controls how much weight we assign to minimizing the coefficient magnitude versus minimizing the error. There is a one-to-one relationship between c and λ so in theory these are equivalent.

Ridge Regression

Choosing $\mathcal{S}(\beta) = \sum_j \beta_j^2$ is called **Ridge Regression**,

$$\min_{\beta} \ ||y - \mathbf{X}\beta||^2 + \lambda||\beta||_2^2.$$

For this penalty, we can find a closed-form solution for $\hat{\beta}_{RR}$,

$$\hat{\beta}_{RR} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T y$$

Notice that $\hat{\beta}_{RR}$ is biased,

$$E(\hat{\beta}_{RR}) = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X}\beta$$

Lasso

Choosing $\mathcal{S}(\beta) = \sum_j |\beta_j|$ is called **Lasso** (Least Absolute Shrinkage and Selection Operator). Our measure of size ($\mathcal{S}$) is also a norm on $\mathbb{R}^p$ and is called the L_1 norm,

$$\min_{\beta} ||\boldsymbol{y} - \mathbf{X}\beta||^2 + \lambda||\beta||_1.$$

Lasso is particularly popular for model selection because it tends to zero-out values of β_j rather than just make them small. This has the effect of defining a subset model.