Nov 12, 2018

1 Linear Model

1.1 Ordinary Least Square

$$\hat{\beta}^{ols} = \min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \tag{1}$$

Ordinary Least Square의 경우 오차의 평균이 0, 오차의 분산이 등분산, 오차가 서로 uncorrelated 일 경우 선형 모형 중에서 최적의 모델이다 (BLUE) 이때 오차의 분포가 정규분포라면 MLE와 동일한 결과를 보여준다

OLS가 아닌 다른 방법을 사용하는 이유는 변수의 개수 p보다 관측치의 개수 n이 충분히 많지 않은 경우 정확한 예측을 하지 못하는 문제가 발생하기 때문이다. 또한 실제 β값중에서 0이 되어야 하는 β 대해서도 0에 가까운 값으로 예측 할 뿐 실제 0으로 예측하지 못하는 문제가 발생한다.

1.2 Ridge regression

$$\hat{\beta}^{ridge} = \min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2}$$

L2 regularization term $\lambda \sum_{j=1}^{p} \beta_j^2$ 이 추가되어 있다 기존 OLS와 동일한 부분에서 데이터에 가장 잘 적합하는 동시에 L2 penalty를 통해서 계수들이 0으로 가게 하는 shrink 효과를 부여한다. 이때 L2 regularization term 은 β_j 가 X_j의 sclae에 따라 영향을 받는다 따라서 우선적으로 각 변수들을 표준화 해주어야 한다

least square방법은, unbiased하나 높은 variance를 가진 계수를 추정하게 된다. 이는 데이터가 조금만 바뀌어도 계수들이 크게 변동할 수 있음을 의미한다. 특히, $p > n$, 즉 설명변수가 많아질때 least square는 심지어 유일한 해가 없게 된다. 이러한 상황에서 ridge regression은 약간의 bias에서의 손해로 variance를 크게 줄여 least square보다 좋은 결과를 가져올 수 있다. 쉽게 말해 설명변수p 보다 데이터의 수n이 적을때, 더욱 덜 flexible한 적합을 하여 소수의 데이터의 특성에 국한되지 않는 모델을 만드는 것이다.

1.3 Lasso regression

$$\hat{\beta}^{lasso} = \min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3}$$

Ridge regression의 단점으로는 중요하지 않은 변수의 β_j를 정확히 0으로 보내지 못한다. Lasso는 적당한 λ만으로 몇몇 계수를 정확하게 0으로 가게 만들 수 있다. 따라서 몇몇 중요하지 않은 변수가 사라진 효과이므로 해석력에서 ridge보다 강력한 강점을 가지고 있다.

2 Nonlinear Model

2.1 Polynomial Regression

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 .. \beta_d x_i^d \tag{4}$$

$$\hat{\beta}^{ols} = \min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \tag{5}$$

각 계수들은 OLS를 사용하여서 적합한다. y의 전체 구간에 대해서 non linear형태를 부여한다

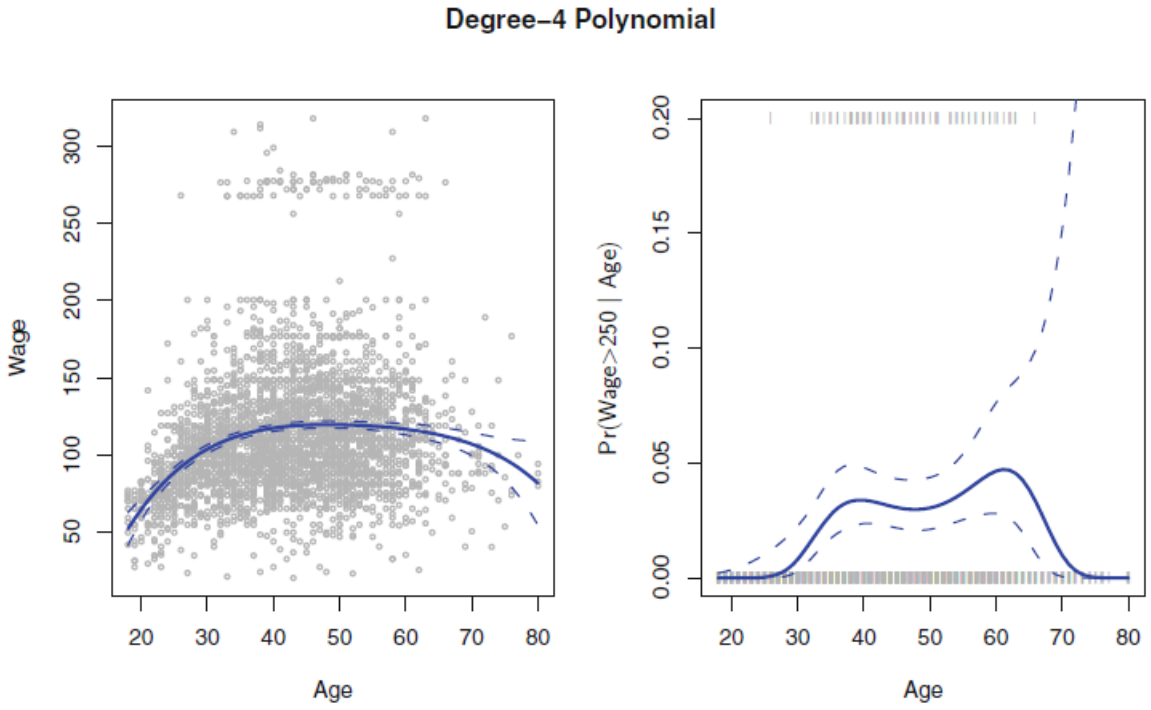

Figure 1: Polynomial Regression

2.2 Step Functions

$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) .. \beta_K C_K(x_i) \tag{6}$$

$$\hat{\beta}^{ols} = \min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \tag{7}$$

$$C_0(X) = I(X < c_1) \tag{8}$$

$$C_1(X) = I(c_1 \leq X < c_1) \tag{9}$$

$$C_{K-1}(X) = I(c_{K-1} \leq X < c_K) \tag{10}$$

$$C_K(X) = I(c_K \leq X) \tag{11}$$

전체 X를 K개의 범주로 나누고 각 범주에 대해서 상수를 부여한다. 이 식의 계수들 역시 least square 로 적합하여서 구한다

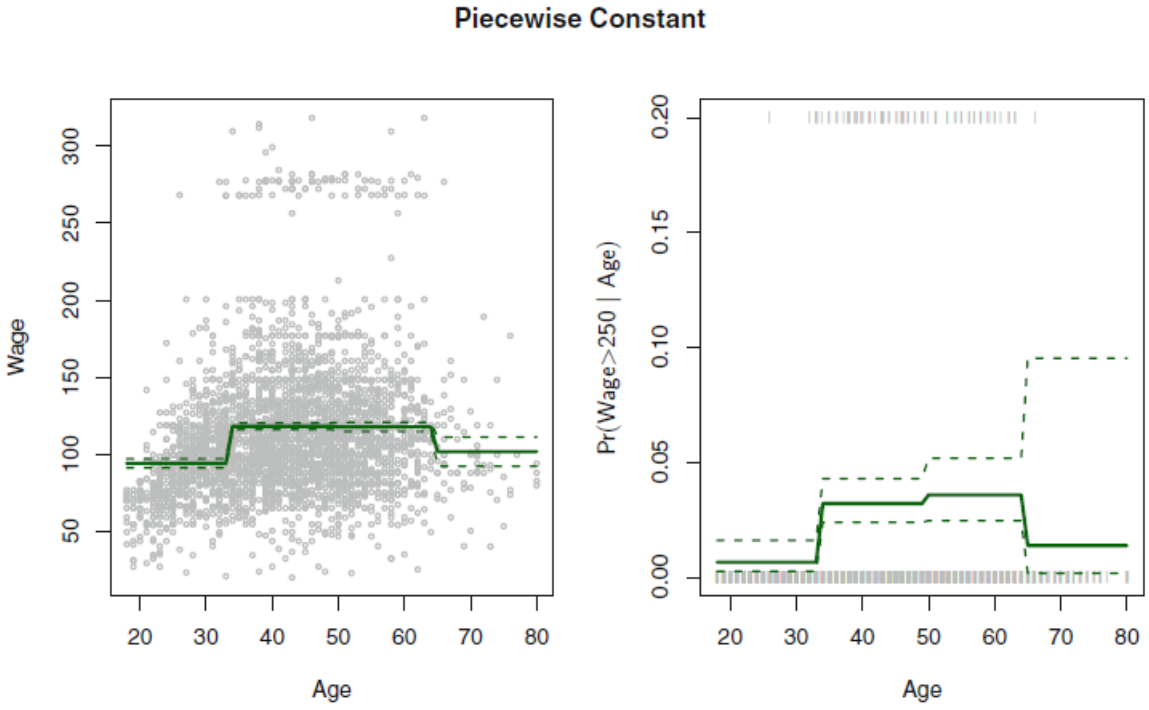

Figure 2: Step Functions

2.3 Regression Splines

전체 X의 범위에 대해서 Polynomial Regression를 적합하는 것이 아니라 구간별로 낮은 차수의 적합을 따로하는 것을 piece wise polynomial regression 이라고 한다 이때 각 범주가 바뀌는 지점을 knots이라고 한다. 이때 knots τ에서 회귀선이 연결되지 않는 문제가 발생한다. 또한 단순히 연결되는 제약조건을 부여할 경우 knots에서 선이 급격하게 꺽여있는 모습을 보인다. 따라서 합리적인 모형을 만들기 위해서 다항식의 차수가 d라고 할때 d-1차 까지의 미분이 가능하게 하여야 한다

$$h(x,\xi) = (x-\xi)_+^3 \tag{12}$$

$$y_i = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 h(X,\xi_1) + .. + \beta_{K+3} h(X,\xi_K) \tag{13}$$

OLS를 통하여 β 값들을 추정하게 된다.

3 Group Lasso

y 가 N개의 관측치를 갖고 있고 X는 $N \times P$ 매트릭스일때 , X 매트릭스를 L개의 그룹으로 나눌 수 있다고 가정하자 p_l은 l번째 집합의 변수 갯수라 고 할 때 우리는 각각의 X_l에대해서 계수 β_l을 추정할 수 있다.

$$\hat{\beta}^{glasso} = \min_{\beta}(\frac{1}{2}||y - \sum_{l=1}^{L} X_l\beta_l||_2^2 + \lambda \sum_{l=1}^{L} \sqrt{p_l}||\beta_l||_2) \tag{14}$$

Group Lasso를 통해서 각 그룹간 설명력을 비교하여 설명력이 낮은 그룹을 제거하는 효과를 얻을 수 있다. 하지만 그룹내에서도 유의한 변수와 유의하지 않은 변수가 있을 경우 이를 반영하지 못한다 따라서 추가적인 L2 panelty를 통해 수정할 수 있다.

$$\hat{\beta}^{sparse\_glasso} = \min_{\beta}(\frac{1}{2n}||y - \sum_{l=1}^{L} X_l\beta_l||_2^2 + \lambda_1 \sum_{l=1}^{L} \sqrt{p_l}||\beta_l||_2 + \lambda_2||\beta||_1) \tag{15}$$