고급회귀분석론

Ch2. Simple Linear Regression

양성준

단순선형회귀모형

- 예측변수 하나와 반응변수 하나의 관계를 직선관계(linear relationship)로 모형화
- 먼저 얻게 된 관측치 쌍이 $(x_i, y_i),\ i = 1, 2, \ldots, n$이라 하자.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- 회귀계수 : β_0 - 절편(intercept), β_1 - 기울기(slope)
- ϵ_i들은 서로 uncorrelated 되어 있다고 보통 가정한다. 이는 y_i들도 서로 uncorrelared임을 함의한다.
- $Var(\epsilon_i) = \sigma^2$ 가정이 추가되면 반응변수의 분산은 예측변수의 값에 상관없이 동일하다는 의미이다.
- 추정대상은 β_0, β_1 혹은 오차항의 분산 σ^2이지만 보통 β_1에 대한 추정 및 추론이 주 관심사이다.

최소제곱추정(least-squares estimation)

- 수많은 직선들 중 어떤 직선이 best인가?
- 최소제곱추정법은 모형에 의한 반응변수의 추정치와 실제 반응변수의 관측치 사이의 거리의 제곱합을 최소화하는 직선을 추정모형으로 선택하는 것이다.

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- 위 식은 β_0, β_1의 값에 의해 결정된 하나의 직선에 대한 오차제곱합을 나타낸다. 즉, 이 식이 어떤 β_0, β_1에서 최소가 되는지를 푸는 문제로 귀결된다.

최소제곱추정량

- β_0, β_1에 대해서 각각 편미분한 뒤 0으로 놓는다.

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

- 위 두식을 정규방정식(normal equation)이라 한다. 연립하여 풀면

$$\hat{\beta}_1 = S_{xy}/S_{xx}, \ \hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}$$

여기서 $S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}), S_{xx} = \sum_i (x_i - \bar{x})^2$

- Q : 추정된 직선이 항상 지나게 되는 지점이 있는가?

적합치 및 잔차

- 주어진 x_i에서 최소제곱직선에 의해 결정되는 y_i의 값을 적합치(fitted value)라 한다.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- 반응변수에서 적합치와 관측치 사이의 차이를 잔차(residual)이라 한다. 잔차는 오차의 실현값(realized value)로 간주할 수 있다.

$$e_i = y_i - \hat{y}_i$$

- 잔차는 후에 모형의 가정을 체크하는 데 있어서 매우 중요한 역할을 하게 된다.

Example : Rocket Propellant Data

- x : 추진제 연식, y : 결합전단강도

$$\hat{y} = 2627.82 - 37.15x$$

- 연식에 따라 전단강도는 하강한다. 1년마다 평균적으로 37.15 정도 줄어든다.
- 10년 된 로켓의 전단강도는 평균적으로 2256.32정도로 예측된다.

최소제곱추정량의 성질

- $\sum_i (x_i - \bar{x}) = 0$임
- 최소제곱추정량은 linear estimator임

$$\hat{\beta}_1 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} y_i$$

- $E(\hat{\beta}_1) = \sum_i (x_i - \bar{x}) E(y_i)/S_{xx} = \sum_i (x_i - \bar{x})(\beta_0 + \beta_1 x_i)/S_{xx}$
  즉, $E(\hat{\beta}_1) = \beta_1$
- 마찬가지로 $E(\hat{\beta}_0) = \beta_0$
- 최소제곱추정량은 불편추정량임

최소제곱추정량의 성질

- y_i가 서로 uncorrelated이므로

$$Var(\hat{\beta}_1) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{S_{xx}^2} Var(y_i) = \frac{\sigma^2}{S_{xx}}$$

- 또한

$$Var(\hat{\beta}_0) = Var(\bar{y}) + \bar{x}^2 Var(\hat{\beta}_1) - 2\bar{x} Cov(\bar{y}, \hat{\beta}_1) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}}{S_{xx}} \right)$$

여기서 $Cov(\bar{y}, \hat{\beta}_1) = 0$임을 이용하였다.

몇가지 이론적 성질

- 정규방정식으로부터 $\sum_i e_i = \sum_i e_i x_i = 0$
- $\sum_i y_i = \sum_i \hat{y}_i$
- $\sum_i \hat{y}_i e_i = 0$

오차항의 분산 추정

- σ^2은 회귀계수의 추정에서는 중요하지 않으나 추정량의 분산과 연관된다. 즉, 계수의 신뢰구간을 구성하거나 검정등을 실시할 때 필요하다.
- 만약 오차항을 관측할 수 있다면 관측된 오차들의 표본분산으로 추정이 가능할 것이다.
- 실제로는 오차항이 직접 관측되지 않으므로 잔차를 통해 추정해야 한다.

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SS_{Res}}{d.f.} = MS_{Reg}$$

- 위 추정량은 Mean squared error 혹은 Residual mean square등으로 지칭한다.
- 모형의 잔차로부터 추정되므로 모형에 깊게 의존한다. 즉, 모형이 잘못 설정된 경우 유용성이 심각하게 저하된다.

회귀모형의 유의성

- 회귀모형은 본질적으로 변수들간의 유의미한 관계를 전제로 하는 것이다. 단순선형회귀모형에서 이 유의성은 $\beta_1 = 0$ 여부에 따라 결정된다.
- 유의성검정

$$H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$$

- 분포에 대한 가정 등은 우선 생략하자. 위와 같은 가설은 β_1에 대한 추정량과 그 표준오차로부터 간단히 검정할 수 있다.

$$t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$$

위 통계량의 절대값이 적절한 자유도의 t분포 분위수보다 크면, 혹은 표본이 충분히 큰 경우 정규분포의 분위수보다 크면 H_0를 기각할 수 있다.

$$|t| > t_{\alpha/2,n-2} \quad or \quad |t| > z_{\alpha/2}$$

cf) 표준오차는 추정량의 표준편차

- 단, $\beta_1 = 0$ 여부는 모형가정에 의존할 수 있음을 기억하자.

Example : Rocket Propellant Data

- $\hat{\beta} = -37.15,\ se(\hat{\beta}_1) = 2.89$로부터

$$t = -37.25/2.89 = -12.85$$

유의수준 $\alpha = 0.05$에서 $|t| = 12.85 > 2.101 = t_{0.025,18}$이므로
귀무가설을 기각하여 변수들간의 직선관계가 유의미하다고 볼 수 있다.