응용통계학

Simple Linear Regression

양성준

회귀모형 (Regression model)

- 예측변수 x와 반응변수 y 사이의 함수적 연관성 모형화
- 확률변수인 (unobservable) 오차항을 고려하며 보통 가법적으로 모형에 포함

$$y = f(x) + \epsilon$$

- $f \in \mathcal{F}$에서 $\mathcal{F}$를 어떻게 설정하느냐가 1차적인 고려 요소
- (x, y)에 대해 다음의 순서쌍의 관측을 전제

$$(x_i, y_i),\ i = 1, \ldots, n.$$

- 같은 모형 하에서도 여러 접근(추정 방식)이 존재

모수적 방법

- 여러 가능한 접근법들

- $f(x) = \beta_0 + \beta_1 x$

- $f(x) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p$

- $f(x) = \beta_2 \dfrac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)}$

- $\ldots$

모수적 방법

- 모수적 접근들은 공통적으로 $dim(\mathcal{F}) < \infty$
- f의 형태를 사전에 결정 ($\mathcal{F}$)
- $\mathcal{F}$에 따라 유한개의 모수를 추정하는 문제로 바뀌게 되어 추정과 해석이 상대적으로 단순함
- $\mathcal{F}$를 벗어나는 형태의 연관성을 고려되지 않으므로 데이터의 특성을 충분히 반영 못할 수 있고, $\mathcal{F}$의 형태 결정이 주관적일 수 있음
- $\mathcal{F}$ 안에 실제 연관성을 나타내는 함수가 있다고 가정할 수 있을 때 가장 효율적

비모수적 방법

- 비모수적 접근들은 보통 $dim(\mathcal{F}) = \infty$
- f의 형태에 대한 가정을 보통 하지 않음. 최소한의 가정만 수립
- 추정과 해석이 상대적으로 복잡
- $\mathcal{F}$의 범위가 매우 커서 데이터의 특성을 그대로 잘 반영함
- 모수적인 방법으로 충분하지 않을 때, 혹은 데이터 탐색과정에서 사용하기도 함
- “Let the data speak for themseleves”

어떤 모형을 써야 하는가?

- 획일적인 정답은 없다
- 고려해야 할 여러 요소들

  - 목적이 데이터생성프로세스의 이해 및 해석에 있는가? 반응변수에 대한 예측에 있는가?

  - 모형이 데이터를 잘 적합하는가?

  - 모형의 가정에는 큰 무리가 없는가?

  - 해당 분야에서 받아들여질 수 있는 방법인가?

  - 분석결과를 받아볼 사람이 좋아할 것인가?

- “All models are wrong, but some of them are useful”
- 모형은 어디까지나 모형. 실제에 대한 approximation임을 기억하라

회귀모형에서 우리는 무엇을 추정하는가?

- 오차항에 대해서 다음의 가정을 한다면

$$E(\epsilon|x) = 0$$

다음과 같은 등식을 얻을 수 있다.

$$f(x) = E(y|x)$$

- 즉 추정대상이 되는 f는 주어진 예측변수에서의 반응변수의 평균을 나타내는 함수이다 (conditional mean function). 이런 의미에서 회귀모형을 mean regression model이라고 부르기도 한다.

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_i는 정규분포를 따른다고 보통 가정한다. (정규성)
- ϵ_i들은 서로 uncorrelated 되어 있다고 보통 가정한다. 이는 y_i들도 서로 uncorrelared임을 함의한다. (독립성)
- $Var(\epsilon_i) = \sigma^2$ 가정이 추가되면 반응변수의 분산은 예측변수의 값에 상관없이 동일하다는 의미이다. (등분산성)
- 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.

최소제곱추정량의 성질

- x_i는 확률변수가 아니라 상수인 것으로 가정 (fixed design).
  확률변수라 가정해도 결과는 크게 바뀌지 않는다.
- $\sum_i (x_i - \bar{x}) = 0$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.
- Q. 오차항에 대한 어떤 가정이 사용되었는가?

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

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

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

- Q. 오차항에 대한 어떤 가정이 사용되었는가?

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.

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SSE}{d.f.} = MSE$$

- 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

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

- 위와 같은 가설은 H_0가 참일 때 다음 사실로부터 간단히 검정할 수 있다.

$$t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim t(n-2)$$

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

- Q. 위 분포식은 모형의 어떤 가정에 의존하는가?

분산분석 (Analysis of Variance)

- 회귀모형의 유의성 검정을 위해 분산분석의 관점에서 접근할 수도 있다.
- 반응변수에 존재하는 총변동(분산)을 모형에 의한 변동과 나머지(오차에 의한) 변동으로 분해하는 것이다. 이러한 접근은 단순선형회귀모형 뿐 아니라 중선형회귀모형, 비선형모형, 더 일반화된 모형에도 사용되는 매우 범용적인 방법이다.
- 하나의 관측치에 대해서는 다음과 같은 분해가 가능하다.

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

- 모든 관측치에 의한 변동은 다음과 같이 분해가 가능하다.

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+2\sum_{i=1}^{n}(\hat{y}_i-\bar{y})(y_i-\hat{y}_i)$$

분산분석 (Analysis of Variance)

- $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 0$ 으로부터 다음이 성립한다.

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- 위 각항을 SST, SSR, SSE라 보통 지칭한다.
- $SSR = \hat{\beta}_1 S_{xy} = S_{xy}^2 / S_{xx}$임을 보일 수 있다.
- 제곱합의 자유도는 각각 $n-1,\ 1,\ n-2$가 된다. 제곱합을 자유도로 나눈 것을 평균제곱합이라 하고 각각 MST(잘 쓰지 않음), MSR, MSE라 지칭한다. 자유도는 n개의 제곱합에서 제약조건의 개수를 뺀 것으로 이해할 수 있다. 예를 들어 SST는 $\sum_{i=1}^{n}(y_i - \bar{y}) = 0$이라는 제약조건이 하나 존재하므로 자유도가 $n-1$이 된 것으로 볼 수 있다. SSE는 $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0,\ \sum_{i=1}^{n} x_i(y_i - \hat{y}_i) = 0$으로 두 개의 제약조건이 존재하는 것으로 볼 수 있다.

분산분석 (Analysis of Variance)

- 분산분석에서는 F 검정을 이용한다. $H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$ 검정을 위하여 다음과 같이 통계량을 정의한다.

$$F_0 = \frac{SSR/1}{SSE/(n-2)} = \frac{MSR}{MSE}$$

- 위 사실로부터 F_0의 관측값이 크면 $\beta_1 \neq 0$ 일 가능성이 크다 할 수 있다.

- 귀무가설 H_0 하에서 $F_0 \sim F_{1,n-2}$임이 알려져 있다.

t 검정과 F 검정

- 앞서 t 검정 통계량은 다음과 같이 정의되었다.

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{MSE/S_{xx}}}$$

따라서

$$t_0^2 = \frac{\hat{\beta}_1^2}{MSE/S_{xx}} = \frac{S_{xy}^2/S_{xx}^2}{MSE/S_{xx}} = \frac{MSR}{MSE} = F_0$$

- 즉, 단순선형회귀모형에서 양측 t검정과 F검정은 같은 결과를 준다.
- 중선형회귀모형에서는?

구간추정

- 추정량의 표준오차와 몇가지 분포성질로부터 모수에 대한 구간추정(신뢰구간)이 가능하다.
- 구간추정은 전반적인 추정의 질을 평가할 수 있게 해 준다.
- 오차항의 정규성, 독립성, 등분산성 가정 하에서 $\hat{\beta}_0, \hat{\beta}_1$이 자유도가 $n-2$인 t 분포를 따른다는 사실로부터 유도될 수 있다.
- $\beta_j,\ j=0,1$의 $100(1-\alpha)\%$ 신뢰구간은 다음과 같이 주어진다.

$$(\hat{\beta}_j - t_{\alpha/2,n-2} se(\hat{\beta}_j), \hat{\beta}_j + t_{\alpha/2,n-2} se(\hat{\beta}_j))$$

- σ^2에 대한 신뢰구간 또한 구성이 가능하다.

평균반응에 대한 구간추정

- 주어진 x_0에서의 평균반응에 대한 추정량은

$$\hat{E}(y|x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

- $Cov(\bar{y}, \hat{\beta}_1) = 0$임을 이용하여 분산을 계산해 보면

$$\begin{aligned} Var(\hat{\beta}_0 + \hat{\beta}_1 x_0) &= Var(\bar{y} + \hat{\beta}_1(x_0 - \bar{x})) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \end{aligned}$$

- 다음과 같이 평균반응에 대한 구간추정이 가능

$$(\hat{E}(y|x_0) - t_{\alpha/2,n-2} se(\hat{E}(y|x_0)), \hat{E}(y|x_0) + t_{\alpha/2,n-2} se(\hat{E}(y|x_0)))$$

- x_0가 $\bar{x}$에서 멀어질수록 구간의 폭이 커지게 된다.

결정계수 (Coefficient of determination)

- 반응변수의 전체 변동성 중 모형(독립변수)에 의해 설명되는 비율을 결정계수라 하고 다음과 같이 정의한다.

$$R^2 = \frac{SSR}{SST}$$

- SST>SSR 이므로 $0 \leq R^2 \leq 1$이 성립한다.
- 결정계수가 큰 것은 모형의 설명력이 높은 것을 의미한다. 하지만 큰 결정계수 값이 항상 바람직한 것은 아니며 과적합 여부를 살펴보아야 한다.
- 얼마 이상이어야 한다는 기준치는 없다. 데이터의 특성 (오차항의 분산 등)에 따라 다양한 값을 나타낼 수 있다. 결정계수만을 가지고 모형의 유용성을 평가하는 것은 위험할 수 있다.
- 결정계수가 크다고 해서 항상 현재 적합된 모형이 적절함을 의미하지는 않는다.

Some other issues

- 예측변수의 범위를 벗어나서 예측하는 것은 위험할 수 있다.
- 절편이 없는 모형의 적합이 필요한 경우도 있다. (원점을 지나는 직선)
- 최소제곱추정량은 분포의 형태에 대한 자세한 가정 없이 유도될 수 있다. 만약, 추정단계에서 오차항에 대한 정규성, 독립성, 등분산성 등을 가정한다면 다른 추정방법을 사용하는 것도 가능하다. 예를 들어 최대가능도추정량(Maximum likelihood estimator)을 들 수 있는데 이는 절편과 기울기에 대해서는 최소제곱법과 같은 추정량을 주며 오차항의 분산에 대한 추정량만 약간 다르게 나타난다. 일반적으로 최대가능도추정량 및 그 성질을 다루기 위해서는 더 강력한 통계이론이 필요하지만 추정량의 성질면에서는 최소제곱추정량보다 여러가지로 더 나은 성질을 가진다.

예측변수가 확률변수인 경우?

- 앞서서는 예측변수가 확률변수가 아니라고 가정하였다. 이 경우 주어진 예측변수의 수준에서 반복하여 반응변수를 관측하는 것이 가능하다. 하지만, 많은 경우 예측변수 또한 확률변수이고 예측변수와 반응변수가 적절한 결합분포를 형성하고 연구자는 그 관측치 쌍을 반복 관측하는 것으로 보는 것이 적절하다.

- 앞서 기술한 통계적인 절차들이 적당한 조건 하에서 그대로 사용가능하다.

- 통계적 절차들은 동일하더라도 해석 측면에서는 조금씩 다를 수 있음을 주의.

- 예측변수가 확률변수인 경우 반응변수와의 상관분석이 가능함

상관분석

- 예측변수와 반응변수와의 상관계수가 0인가?

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

- 위 가설은 두 변수의 결합분포가 정규분포이고 귀무가설이 참이라는 가정 하에서 다음과 같은 표본상관계수의 분포성질을 통해 검정이 가능하다.

$$t_0 = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}} \sim t_{n-2}$$

  위 검정은 $H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$ 과 동일한 검정임 (Hint : $SSR = \hat{\beta}_1^2 S_{xx}$)
- $H_0 : \rho = \rho_0 \quad vs \quad H_1 : \rho \neq \rho_0$ for $\rho_0 \neq 0$는 절차가 더 복잡하고 보통 관심대상이 아닌 경우가 많음

오차항의 가정에 대한 중요성

- 정규성 : 정규확률그림이나 정규성 검정 등으로 확인이 가능하마. 중심극한 정리에 의해 표본의 크기가 일정 수준 이상이 되면 위배되어도 추론결과에 큰 영향 없음
- 독립성 : 종속성은 매우 다양한 형태로 존재할 수 있으므로 모든 경우를 고려하여 검토하는 것은 사실상 불가능함. 예측변수의 수준 순서대로 독립성을 살펴보거나, 시간에 흐름에 따라 관측된 성질이 있는지를 살펴보는 것이 현실적임. 만약 자료수집 설계 자체가 독립성을 담보할 수 있다면 어느 정도 충족되는 것으로 보아도 좋음
- 등분산성 : 가장 위배되기 쉬우면서도 추론 결과에 큰 영향을 주는 가정.
- 가정에 대한 위배가 나타나는 경우 추론 결과의 신뢰성이 저해될 수 있음. 이 경우 Bootstrap등의 대안을 생각할 수 있음.
- 단, 추론 자체보다는 예측에 초점이 있거나, 데이터의 연관성에 대한 근사식 발견에 목적이 있는 경우에는 오차항에 대한 가정 검토에 지나치게 집중할 필요는 없음

비모수 상관분석

- 단순선형회귀모형과 상관분석은 두 변수 사이의 선형적 관계를 전제로 한 것임.

- 비선형적인 관계에 관심이 있는 경우 Kendall 의 τ, Spearman이 순위상관계수등을 활용할 수 있음