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.

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.

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

분산분석 (Analysis of Variance)

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

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

- 위 각항을 SST, SSR, SSE라 보통 지칭한다.
- $SSR = \hat{\beta}_1 S_{xy} = S_{xy}^2 / S_{xx}$임을 보일 수 있다. (Try it!)
- 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. $H_0 = \beta = 0 \quad vs \quad H_1 \beta_1 \neq 0$ 검정을 위하여 다음과 같이 통계량을 정의한다.

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

- $E(MSE) = \sigma^2$, $E(MSR) = \sigma^2 + \beta_1^2 S_{xx}$ 임을 보일 수 있다. (Appendix C.3 참고)
- 위 사실로부터 F_0의 관측값이 크면 $\beta_1 \neq 0$ 일 가능성이 크다 할 수 있다.
- 귀무가설 하에서 $F_0 \sim F_{\alpha,1.n-2}$ 임이 알려져 있다.

Example : Rocket Propellant Data

- $SST = 1693737.60,\ S_{xy} = -41112.65$로부터

$$SSR = \hat{\beta}_1 S_{xy} = 1527334.95$$

$F_0 = 165.21 > F_{0.05,1,18} = 8.29$이므로 유의수준 0.05에서 회귀모형이 유의하다고 결론내릴 수 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.

$$(\hat{\beta}_j - t_{\alpha/2,n-2}se(\hat{\beta}_1), \hat{\beta}_j + t_{\alpha/2,n-2}se(\hat{\beta}_1))$$

- σ^2에 대한 신뢰구간 또한 구성이 가능하다.

Example : Rocket Propellant Data

- $se(\hat{\beta}_1) = 2.89,\ t_{0.025,18} = 2.101$로부터 기울기에 대한 95% 신뢰구간은

$$-43.22 \leq \beta_1 \leq -31.08$$

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

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

- 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 $0 \leq 1$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.
- 결정계수 값은 예측변수의 변동성에도 의존한다.
- 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.
  - 주어진 x에서 y의 조건부 분포가 $N(\beta_0 + \beta_1 x, \sigma^2)$
  - x는 모형의 모수들$(\beta_0, \beta_1, \sigma^2)$과는 상관없는 확률분포를 가짐
- 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 (why?)

- $H_0 : \rho = \rho_0 \quad vs \quad H_1 : \rho \neq \rho_0$ for $\rho \neq 0$는 절차가 더 복잡함.