201111774 응용통계학과 박성진

(Homework) ¶

Search the origin of the word "regression" by googling and then find the actual meaning of the word at that time

- The term **regression** was coined by Francis Galton in the nineteenth century to describe a biological phenomenon.

  회귀라는 용어는 19 세기에 Francis Galton에 의해 생물학적 현상을 설명하기 위해 만들어졌습니다.

- The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean).

  이 현상은 키가 큰 조상의 자손의 높이가 평균(회귀라는 의미의 현상)으로 퇴보하는 경향이 있다는 것을 나타냅니다.

- For Galton, regression had only this biological meaning, but his work was later extended by Udny Yule and Karl Pearson to a more general statistical context.

  Galton의 경우 회귀는 생물학적 의미만 가지고 있었지만 그의 연구는 나중에 Udny Yule과 Karl Pearson에 의해 보다 일반적인 통계적 맥락으로 확장되었습니다.

(Homework)

1. 3 Interpretation of the model

$$y = f(x) + \epsilon = B_0 + B_1 x + \epsilon$$

- model $f(x)$: conditional expectation of $y \; given \; x$
- intercept B_0 : conditional mean of y given $x = 0$

$$B_0 = E(y|x = 0)$$

- slope B_1: difference between two conditional expectations

$$B_1 = E(y|x = k + 1) - E(y|x = k), \quad \forall k \in \mathbb{R}$$

  In this sense, we often say B_1 is the effect of x on y

---

1. 3 .1 Centering/ scaling / standardization

Centering x often helps the interpretation, here centering implies subsituting

$$z = x - \mu_x$$

- for x, where $\mu_x = E(x)$ so that the model becomes

$$y = f(x) + \epsilon = B_0 + B_1 x + \epsilon \iff y = g(z) = \alpha_0 + \alpha_1 z + \epsilon$$

- model $g(z)$: conditional expectation of $y \; given \; z$

$$g(z) = \alpha_0 + \alpha_1 z$$

- intercept α_0 : conditional mean of y given $z = 0 \iff x = \mu_x$

$$\begin{aligned} \alpha_0 &= E(y|z=0) \\ &= E(y|x=\mu_x) \\ &= B_0 + B_1\mu_x \end{aligned}$$

- slope α_1: difference between two conditional expectations

$$\begin{aligned} \alpha_1 &= E(y|z=k+1) - E(y|z=k) \\ &= E(y|x=\mu_x+k+1) - E(y|x=\mu_x+k) \\ &= B_1 \end{aligned}$$

Scaling x implies subsituting

$$z = x/\sigma_x$$

- for $\sigma_x^2 = Var(x)$

$$y = f(x) + \epsilon = B_0 + B_1 x + \epsilon \iff y = g(z) = \alpha_0 + \alpha_1 z + \epsilon$$

- model $g(z)$: conditional expectation of y *given* z

$$g(z) = \alpha_0 + \alpha_1 z$$

- intercept α_0 : conditional mean of y given $z = 0 \iff x = 0$

$$\begin{aligned} \alpha_0 &= E(y|z=0) \\ &= E(y|x=0) = B_0 \end{aligned}$$

- slope α_1: difference between two conditional expectations

$$\begin{aligned} \alpha_1 &= E(y|z=k+1) - E(y|z=k) \\ &= E(y|x=\sigma_x(k+1)) - E(y|x=\sigma_x(k)) \\ &= (B_0 + \sigma_x B_1 k + B_1\sigma_x) - (B_0 + \sigma_x B_1 k) \\ &= \sigma_x B_1 \end{aligned}$$

Standardization x implies subsituting

$$z = (x - \mu_x)/\sigma_x$$

- for x, where $\mu_x = E(x)$, $\sigma_x^2 = Var(x)$

$$y = f(x) + \epsilon = B_0 + B_1 x + \epsilon \iff y = g(z) = \alpha_0 + \alpha_1 z + \epsilon$$

- model $g(z)$: conditional expectation of y *given* z

$$g(z) = \alpha_0 + \alpha_1 z$$

- intercept α_0 : conditional mean of y given $z = 0 \iff x = \mu_x$

$$\begin{aligned} \alpha_0 &= E(y|z=0) \\ &= E(y|x=\mu_x) = B_0 \end{aligned}$$

- slope α_1: difference between two conditional expectations

$$
\begin{aligned}
\alpha_1 &= E(y|z = k+1) - E(y|z = k) \\
&= E(y|x = \sigma_x(k+1) + \mu_x) - E(y|x = \sigma_x(k) + \mu_x) \\
&= (B_0 + \sigma_x B_1 k + B_1 \sigma_x + \mu_x) - (B_0 + \sigma_x B_1 k + \mu_x)
\end{aligned}
$$

201111774 응용통계학과 박성진

1.5 Estimation based on a loss function

- ### Steps for estimating two parameters(coefficients) B_0 and B_1

Determine an appropriate loss function $L(B_0, B_1)$

Find the population mimimizer $\hat{B}_0^{pop}$ and $\hat{B}_1^{pop}$ by minimizing the risk function

$$R(B_0, B_1) = E\{L(B_0, B_1)\}$$

with respect to B_0 and B_1

here, note that we often cannot determine the minimizer for some possible reason

---

- ### For example, we can see that

$$\begin{aligned}(\hat{B}_0^{pop}, \hat{B}_1^{pop}) &= \text{argmin}_{B_0, B_1} E(y - B_0 - B_1 x)^2 \\ &= (\mu_y - \hat{B}_1 \mu_x, \sigma_{xy}^2 / \sigma_x^2) \qquad \cdots (6)\end{aligned}$$

where $\mu_y = E(y)$, $\sigma_{xy}^2 = Cov(x, y)$and $\sigma_x^2 = Var(x)$

- It is impossible to specify the minimizer unless the joint distribution of x and y is known

  we do not know μ_x, μ_y, σ_x and σ_{xy}

(Homework)

Prove(6)

- 회귀분석의 1차적인 목적은 표본으로부터 모회귀계수 $\hat{B}_0^{pop}, \hat{B}_1^{pop}$ 를 추정하여 추정된 회귀식을 만드는 것이다.

- 최소 제곱법은 $\sum_{i=1}^{n}(e_i)^2 = \sum_{i=1}^{n}(y_i - (B_0 + B_1 X_i))^2$ 식을 각각 B_1과 B_0로 각각 편미분하여 0과 같다고 놓는다. 그러면

$$\begin{aligned}\sum y_i &= n\hat{B}_0 + \hat{B}_1 \sum x_i \\ \sum x_i y_i &= \hat{B}_0 \sum x_i + \hat{B}_1 \sum x_i^2\end{aligned}$$

- 의 식이 나타난다. 이를 정리하면

$$\hat{B}_0 = E(y) - \hat{B}_1 E(x)$$

$$\hat{B}_1 = \frac{\sum_{i=1}^{n}(x_i - E(x))(y_i - (E(y))}{\sum_{i=1}^{n}(x_i - E(x))^2}$$