*J Vet Clin* 27(4) : 427-434 (2010)

Journal of Veterinary Clinics
www.ksvc.or.kr

# 상관성과 단순선형회귀분석

**박선일[1] · 오태호***

강원대학교 수의과대학 및 동물의학종합연구소, *경북대학교 수의과대학

(게재승인: 2010년 6월 14일)

# Correlation and Simple Linear Regression

**Son-Il Pak[1] and Tae-Ho Oh***

*College of Veterinary Medicine and Institute of Veterinary Science, Kangwon National University, Chuncheon 200-701, Korea*
**College of Veterinary Medicine, Kyungpook National University, Daegu, 702-701, Korea*

**Abstract :** Correlation is a technique used to measure the strength or the degree of closeness of the linear association between two quantitative variables. Common misuses of this technique are highlighted. Linear regression is a technique used to identify a relationship between two continuous variables in mathematical equations, which could be used for comparison or estimation purposes. Specifically, regression analysis can provide answers for questions such as how much does one variable change for a given change in the other, how accurately can the value of one variable be predicted from the knowledge of the other. Regression does not give any indication of how good the association is while correlation provides a measure of how well a least-squares regression line fits the given set of data. The better the correlation, the closer the data points are to the regression line. In this tutorial article, the process of obtaining a linear regression relationship for a given set of bivariate data was described. The least square method to obtain the line which minimizes the total error between the data points and the regression line was employed and illustrated. The coefficient of determination, the ratio of the explained variation of the values of the independent variable to total variation, was described. Finally, the process of calculating confidence and prediction interval was reviewed and demonstrated.

***Key words*** **:** correlation coefficient, linear regression, least square.

## 서 론

두 연속변수(continuous variable) 간의 관계를 평가하는 목적으로 상관분석(correlation)과 회귀분석(regression)을 사용한다(1,5). 일반적으로 두 변수의 관계는 방향(direction)과 강도(strength)로 설명되는데 방향은 양(positive) 혹은 음(negative)의 관계이고, 크기는 연관성(association)의 크기를 나타낸다. 상관분석은 두 변수 간의 선형 연관성(linear association)의 강도를 측정하는 반면 회귀분석은 관찰된 자료를 대표할 수 있는 최적의 선형 (비선형 관계도 분석이 가능하지만 본 논문에서는 선형관계만을 다룸) 연관성을 수리모형으로 표현하는 기법이다. 이를테면 특정 질병으로 진단받은 개에서 혈청 blood urea nitrogen (BUN)과 크레아티닌(creatinine) 농도 간의 관련성을 평가하는 연구에서 연구자가 혈청 BUN 농도 수준을 크레아티닌 농도로 예측할 수 있는지에 관심을 두는 경우다. 또한 크레아티닌 농도와 사구체 여과율(glomerular filtration rate) 간의 관계, 개의 연령과 홍역 백신 역가간의 관계, 약물 투여 후 시간경과에 따른 대사율의 관계 등은 이러한 분석을 필요로 한다.

[1]Corresponding author.
E-mail : paksi@kangwon.ac.kr

## 결 론

### 산점도

두 변수 간의 연관성을 분석하는 전통적인 방법은 한 변수에 대한 측정값을 x축으로 하고 이에 대응하는 다른 변수의 측정값을 y축으로 설정하여 모든 자료를 산점도(scatter diagram, scatter plot)로 표현하는 것이다. 여기에서 x축을 독립변수(independent variable), 설명변수(explanatory), 예측변수(predictor)라고 하며, y축은 종속변수(dependent), 반응변수(response), 결과변수(outcome)라고 한다. 예를 들어 질병 'a'로 진단받은 개의 혈청 BUN과 크레아티닌 농도를 측정한 자료(Table 1)에 대하여 산점도를 그려보면 두 변수는 대략적으로 양의 관계가 있는 것으로 판단된다(Fig 1).

**Table 1.** Blood urea nitrogen (BUN, mg/dl) and creatinine by age (month) for 20 dogs with disease 'a'

| ID | Age (month) $x_i$ | BUN $y_i$ | Creatinine $y_i$ | ID | Age (month) $x_i$ | BUN $y_i$ | Creatinine $y_i$ |
|---|---|---|---|---|---|---|---|
| 1 | 32 | 24 | 1.4 | 11 | 18 | 9 | 0.7 |
| 2 | 48 | 28 | 1.2 | 12 | 27 | 11 | 1.2 |
| 3 | 34 | 27 | 1.7 | 13 | 31 | 24 | 1.4 |
| 4 | 52 | 29 | 0.9 | 14 | 39 | 27 | 1.1 |
| 5 | 57 | 25 | 0.8 | 15 | 58 | 31 | 1.0 |
| 6 | 23 | 12 | 1.2 | 16 | 47 | 23 | 0.9 |
| 7 | 42 | 22 | 1.3 | 17 | 36 | 13 | 1.2 |
| 8 | 26 | 20 | 1.5 | 18 | 48 | 26 | 1.1 |
| 9 | 23 | 19 | 1.1 | 19 | 0 | 22 | 1.3 |
| 10 | 37 | 22 | 0.8 | 20 | 21 | 23 | 0.8 |

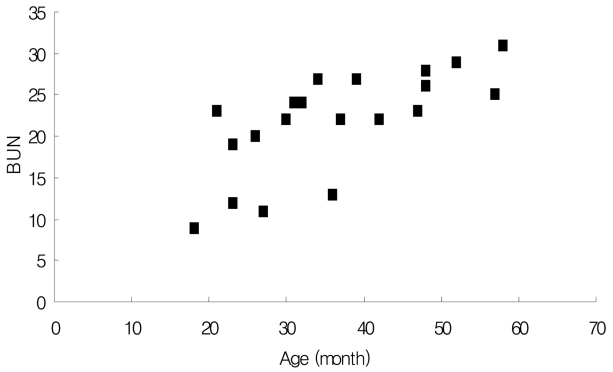


**Fig 1.** Scatter diagram for blood urea nitrogen (BUN, mg/dl) and age (month).

### 상관계수

#### Pearson 상관계수

등간격(interval scale) 이상의 척도로 측정된 두 변수를 x, y라 하면 $(x_1,y_1)$, $(x_2,y_2)$, ...$(x_n,y_n)$등 n개의 관찰치 쌍(pair)이 만들어지고, x 값과 y 값의 평균을 각각 $\bar{x}, \bar{y}$라고 하면 상관계수(r)는 다음의 공식으로 계산되며 이를 Pearson 상관계수라고 한다.

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

상관계수 r은 선형 회귀선(linear regression line)에 관찰치가 근사하는 정도를 평가하는 통계량으로 두 변수 간 선형 연관성의 강도를 측정한다. 즉 산점도에서 관찰치들이 직선상에 위치하면 두 변수 간의 선형관계가 강하다는 것을 의미한다. 상관계수는 –1과 1 사이의 값을 갖고 1에 근사할수록 양(positive)의 선형관계가 매우 강하고, -1에 근사할수록 음(negative)의 선형관계가 매우 강하며, 0에 근사하면 선형관계가 없다고 해석한다(Fig. 2). Pearson 상관계수의 제곱을 결정계수라 하며 자세한 내용은 회귀분석에서 설명한다. Table 1의 자료에 대한 상관계수는 0.656으로 계산되며 혈청 BUN 농도와 연령은 양의 선형관계가 있음을 알 수 있다. Fig 3은

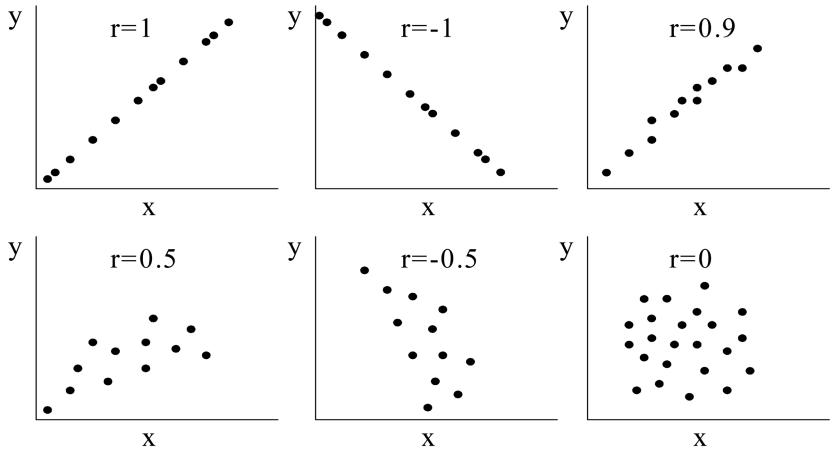


**Fig 2.** Examples of linear relationship (r = correlation coefficient).

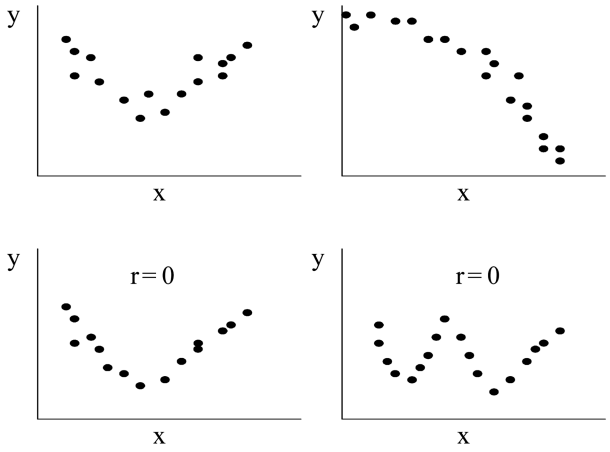


**Fig 3.** Examples of nonlinear relationship (r = correlation coefficient).

두 변수 간 비선형관계(nonlinear relationship)를 예시한 것이다. Pearson 상관계수는 모수적 기법이고 이에 상응하는 비모수기법으로는 Spearman's 순위(rank) 상관계수(rho, $\rho$)를 사용한다(3). Pearson 상관계수는 관찰치가 전체 자료범위에 대하여 균질하게 분포하지 않거나 표본크기가 작을 때, 두 변수의 관계가 선형이 아닐 때, 이상점(outlier)이 존재하거나 자료의 분포가 정규분포에 근사하지 않을 때 불확실성이 증가하므로 이러한 자료에 대해서는 비모수적 접근이 바람직하다.

### 상관계수에 대한 가설검정

두 변수 간 선형 상관관계 여부는 상관계수 r = 0이라는 귀무가설을 검정한다. Table 1의 자료를 분석하면 모집단 상관계수가 0이라는 귀무가설을 기각하고 BUN 농도와 연령은 선형관계가 있다는 결론을 얻는다($p < 0.01$).

### 상관계수의 신뢰구간

상관계수에 대한 가설검정은 선형관계의 여부를 판단하지만 이러한 관계의 강도 즉 모집단의 상관계수가 포함될 구간을 정량화하기 위해서는 신뢰구간을 계산하는 것이 바람직하다. 신뢰구간을 계산하기 위해서는 Fisher의 z 변환을 통하여 상관계수를 정규분포로 변환하며, 변환된 상관계수($z_r$)의 표준오차 근사추정치와 $z_r$의 95% 신뢰구간 $[z_{r(L)}, z_{r(U)}]$은 다음과 같다(6).

$$z_r = \frac{1}{2} log_e \left[\frac{1+r}{1-r}\right], \quad SE(z_r) = \frac{1}{\sqrt{n-3}}$$

$$z_{r(L)},\ z_{r(U)}: [r - z_{1-\alpha/2}/(\sqrt{n-3}),\ r + z_{1-\alpha/2}/(\sqrt{n-3})]$$

따라서 $r$의 95% 신뢰구간은 다음과 같이 계산된다.

$$95\%\ 신뢰구간: \left[\frac{e^{2z_{r(L)}}-1}{e^{2z_{r(L)}}+1}, \frac{e^{2z_{r(U)}}-1}{e^{2z_{r(U)}}+1}\right]$$

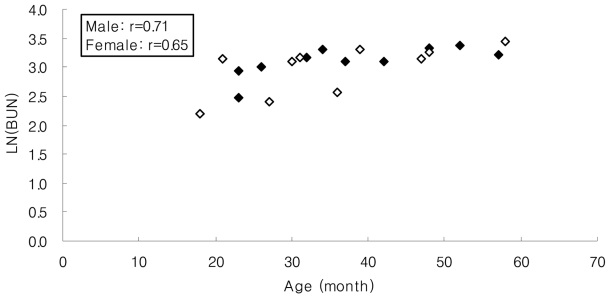


**Fig 4.** Correlation coefficient by sexes. Regression line is ln $y = 0.019 \times Age + 2.35$. Note that the BUN level was logarithmic transformed prior to fitting. Female (open diamond); Male (filled diamond).

Table 1의 자료에 대하여 $z_r$에 대한 신뢰구간을 계산하면 [0.311, 1.260]을 얻는다. 따라서 상관계수 r에 대한 신뢰구간은 Fisher 변환된 상, 하한값을 역변환해야 하며 신뢰구간은 [0.301, 0.851]로 계산된다.

$$z_r = \frac{1}{2} log_e \left[\frac{1+0.656}{1-0.656}\right] = 0.786,$$

$$SE(z_r) = \frac{1}{\sqrt{20-3}} = 0.242$$

대수변환 95% 신뢰구간: $0.786 \pm 1.96 \times 0.242 \leftrightarrow [0.311, 1.260]$

95% 신뢰구간:

$$\left[\frac{e^{2\times 0.311}-1}{e^{2\times 0.311+1}} = 0.301, \quad \frac{e^{2\times 1.26}-1}{e^{2\times 1.26}+1} = 0.851\right]$$

### 상관계수의 오용

상관계수에 대한 해석과 관련하여 오류를 범하는 경우가 많은데 가장 흔한 것은 두 변수 간의 상관관계를 분석할 때 상관관계에 잠재적으로 영향을 미칠 수 있는 제3의 변수를 고려하지 않고 해석하거나, 특히 상관계수가 유의할 때 이러한 결과에 대하여 두 변수 간 인과관계(causality)가 있다는 것으로 해석하는 것이다(7). 예를 들어 동물병원에 내원한 개의 폐사율이 병원의 규모와 양의 상관성이 있는 것으로 분석되었다고 하자. 규모가 큰 병원일수록 폐사율이 높다는 연구결과에 대하여 고려해야할 사항은 규모가 큰 병원의 진료의 질이 낮기 때문에 폐사율이 높은 것이 아니라 대형 병원에 내원할 정도로 환자의 질병경과가 보다 중증인 경우가 많기 때문에 이러한 결과를 얻을 수 있다는 점을 고려해야 한다. 이 경우 환자의 질병경과는 두 변수의 상관관계에 영향을 미치기 때문에 분석시 질병경과를 보정하여 상관계수를 계산할 필요가 있다. 둘째, r = 0은 두 변수 간 상관관계가 없다는 것을 의미하는 것이 아니다. 두 변수 간 선형관계는 없지만 비선형관계가 존재할 수 있다(Fig 3). 셋째, 두 변수 간의 관계를 하부집단(subgroup)으로 구분할 수 있는 경우 하부집단별 상관계수에 비하여 전체 자료의 상관계수가 과장되어 나타날

수 있다(Fig 4). Table 1의 자료에서 BUN 농도와 연령간의 관계에서 전체 자료의 상관계수는 0.656이지만, 예를 들어 성별로 구분할 때 수컷 0.71, 암컷 0.65로 분석되었다고 하자. 이 경우 수컷의 상관관계는 과소평가된 반면 암컷의 상관관계는 미미하지만 과대평가되었다고 볼 수 있다. 극단적인 예로 개별 집단에서는 음의 상관관계가 전체 자료에서는 양의 상관관계를 보이는 경우도 있다(Fig 5). 넷째, 상관계수가 높다는 결과를 두 측정치 간의 일치도(agreement)가 높다고 해석하는 것은 잘못된 것이다. 일치도 분석이 연구의 목적이라면 다른 분석기법을 사용하는 것이 바람직하다(4).

### 회귀분석

#### 최소제곱 회귀선

상관분석에서 예시한 BUN 농도와 연령 자료에서 연구자는 질병 'a로 진단받은 개체의 연령이 BUN 농도에 미치는 영향에 관심을 가질 수 있다. 즉 BUN 농도를 연령으로 예측할 수 있는지를 평가하는 것으로 이러한 연구에서는 두 변수 간의 관계를 방정식으로 표현해야 하며 이를 회귀선(regression line)이라고 한다.

Table 1의 자료에 대하여 회귀식을 유도하여 보자. 예를 들어 5번째 관찰치는 질병 'a'로 진단받은 환자의 연령(x)이 57개월이고 이 환자의 BUN (y)은 25 mg/dl 이다. 만일 57개월령의 다른 환자에 대하여 BUN을 측정한다면 아마도 25 mg/dl과 다른 (같거나 작거나 큰) 결과를 얻을 수 있다. 이러한 차이가 나타나는 이유는 BUN 농도에 영향을 미칠 수 있는 요인이 연령 이외의 많은 다른 요인이 관여할 수 있기 때문이다. 따라서 x = 57인 많은 환자를 대상으로 측정하면 x = 57의 평균값($\mu_5$)을 얻을 수 있을 것이다. 이를 회귀식으로 표현하면 $y_5 = \mu_5 + \varepsilon_5$가 된다. 동일한 방법으로 20두($i$)에 대한 회귀식을 얻을 수 있으며 이를 일반화하면 다음과 같다.

$$y_i = \mu_i + \varepsilon_i$$

여기에서 $\mu_i$는 환자의 연령이 $x_i$일 때 관찰할 수 있는 모집단 BUN 농도의 평균값이고, 오차항 $\varepsilon_i$는 환자의 연령을 제외한 모든 다른 요인이 $y_i$에 미치는 효과의 크기를 나타낸다. 20두 각각의 평균값 $\mu_1$, $\mu_2$, $\cdots\mu_{20}$이 $x_1$, $x_2$, $\cdots x_{20}$과 직선의 관계가 있다고 가정하면 이 직선은 다음과 같이 정의할 수 있다.

$$\mu_i = \beta_0 + \beta_1 x_i$$

이 식에서 $\beta_0$는 직선의 절편, $\beta_1$은 직선의 기울기(slope)가 된다. 예를 들어 $x_i = 0$을 가정하면 위의 식은 $\mu_i = \beta_0$이 된다. 따라서 절편에 대한 해석은 환자의 연령이 0인 모든 가능한 환자들의 평균 BUN 농도가 된다 (그러나 실제로 환자의 연령이 0인 경우는 없기 때문에 큰 의미를 갖는 것은 아님). 기울기를 해석하기 위하여 2두의 환자 즉 $x_1 = c$와 $x_2 = c + 2$을 가정하면 두 환자의 회귀식은 다음과 같이 표현할 수 있으며, 이들 두 환자의 평균 BUN 농도 차이는 $\beta_1$이 된다.

$$\mu_1 = \beta_0 + \beta_1(c),\ \mu_2 = \beta_0 + \beta_1(c+1)$$

$$\mu_2 - \mu_1 = [\beta_0 + \beta_1(c+1)] - [\beta_0 + \beta_1(c)] = \beta_1$$

따라서 기울기 $\beta_1$은 환자의 연령 1 단위 증가와 관련된 평균 BUN 농도의 변화량이 된다는 것을 알 수 있다. 요약하면 $i$번째 환자의 평균 BUN 농도 $\mu_i$가 $\beta_0 + \beta_1 x_1$이라고 하면 $i$번째 환자에서 관찰된 BUN 농도는 다음과 같이 정리할 수 있다.

$$y_i = \mu_i + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

이 방정식을 선형 회귀모형(linear regression model)이라고 한다. $\beta_0$와 $\beta_1$을 모형의 모수(parameter)라고 하며, 두 모수의 참값은 알 수 없다. 그러나 관찰 자료 $x_1$, $x_2$, $\cdots$, $x_n$와 $y_1$, $y_2$, $\cdots$, $y_n$을 사용하여 $\beta_0$와 $\beta_1$의 점추정치로 $b_0$와 $b_1$를 최소제곱법(least square)으로 계산할 수 있다. 즉 모든 편차의 제곱합을 최소화하는 $b_0$와 $b_1$의 값을 찾는 것으로 다음의 공식으로 계산된다.

$$b_0 = \bar{y} - b\bar{x}; \quad b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

주어진 값 x에 대하여 관찰된 값($y_i$)과 회귀식으로 예측된 값($\hat{y}_i$) 간의 수직 차이를 편차(deviation) 혹은 잔차(residual)라고 하며 ε로 표기한다(Fig 6).

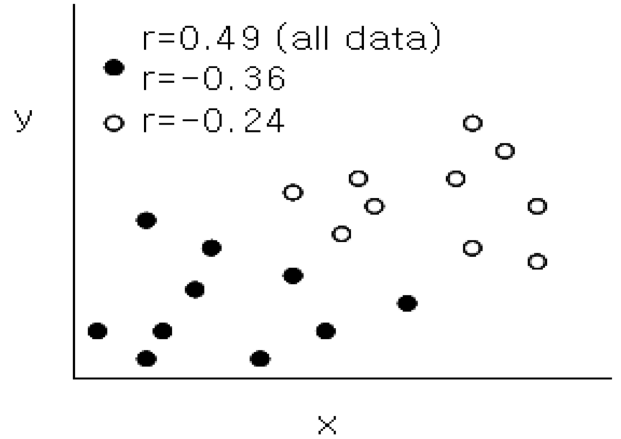


**Fig 5.** Comparison of correlation coefficients by subgroup and combined data.

#### 선형회귀모형의 가정

독립변수 x, 종속변수 y로 구성된 단순 선형회귀식(regression of y on x)은 세가지 가정을 전제로 한다(5). 첫째, 분산의 동질성(constant variance)은 독립변수 x의 특정한 값 $x_i$에 상응하는 종속변수의 값은 $x_i$에 무관하게 동일한 분산 $\sigma^2$을 갖는다는 것을 의미한다. 둘째, 독립성(independence)은 종속변수 y의 값은 모든 다른 y값과 통계적으로 상호 독립이다.

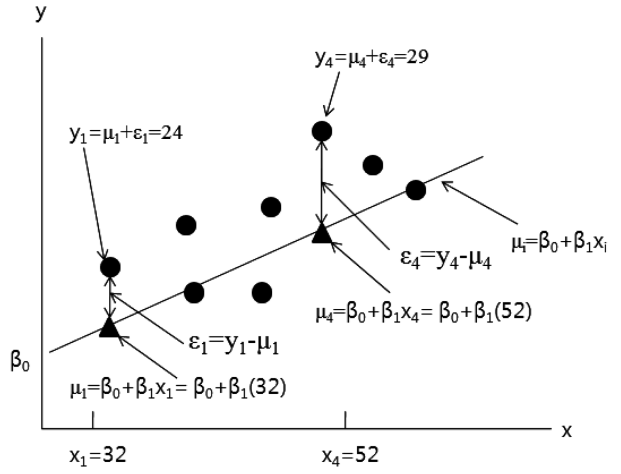


**Fig 6.** Simple linear regression line relating y to x, minimizing the sum of squares of all of the deviations. ε = deviation (residual). $\mu_i$ represents the effect on $y_i$ of the $x_i$, and $y_i$ denotes the value observed in $i$ of the dependent variable.

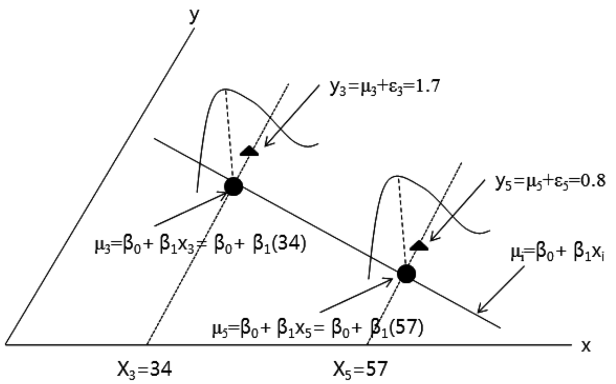


**Fig 7.** An illustration of the constant variance and normality assumptions of regression equation for the data in Table 1. x = age (month), y = creatinine level (mg/dl).

이는 특정한 x값에 상응하는 y값들은 다른 x값에 상응하는 y값과 관련이 없다는 것으로 이를테면 시간(time)대별로 반복하여 측정한 자료는 독립성 가정에 위배된다. 마지막으로 정규성(normality)은 독립변수 x의 특정한 값 $x_i$에 상응하는 종속변수의 값은 정규분포를 따른다는 것을 의미한다. Fig 7은 Table 1의 자료에서 연령과 크레아티닌 농도에 대한 분산의 동질성과 정규성 가정을 예시한 것이다. 즉 환자의 연령이 34개월과 57개월일 때 관찰할 수 있는 모집단 크레아티닌 농도를 나타낸 것으로, 이들 두 모집단 (34개월과 57개월)은 분산이 동일하고 정규분포를 따른다.

**회귀모수에 대한 가설검정과 신뢰구간**

모집단의 절편과 기울기가 각각 0이라는 귀무가설은 회귀계수의 추정치를 각각의 표준오차로 나누어 검정한다.

절편의 표준오차 $[SE(b_0)]$: $s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$

기울기의 표준오차 $[SE(b_1)]$: $\frac{s}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$

표준편차 $(s)$: $\sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2-b\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-2}}$

계산된 검정통계량을 자유도가 n – 2(여기에서 2는 추정하려는 회귀계수의 개수)인 t 분포의 값과 비교하여 유의성을 결정하며, 회귀계수의 95% 신뢰구간은 다음과 같이 계산한다.

절편: $b_0 \pm t_{\alpha/2,df=n-2} \times SE(b_0)$

기울기: $b_1 \pm t_{\alpha/2,df=n-2} \times SE(b_1)$

Table 1의 자료에서 원시자료의 정규성(normality)을 충족하기 위하여 대수변환을 가정하자. 원시자료는 Shapiro-Wilk 검정에서 P = 0.0722로 유의수준 1%에서 정규성을 만족하지 못함). 이 자료를 분석하면 $s$ = 0.269로 계산되며 (SAS 패키지를 사용하는 경우 출력결과에서 root MSE 통계량) 검정통계량과 유의확률을 정리하면 Table 2와 같다. 따라서 절편과 기울기 추정치에 대한 신뢰구간은 다음과 같이 계산된다.

(i) 기울기
$SE(b_0) = 0.005$
95% CI: $= 0.019 \pm 2.101 \times 0.005 \Leftrightarrow [0.0085,\ 0.0295]$

(ii) 절편
$SE(b_1) = 0.195$
95% CI: $= 2.35 \pm 2.101 \times 0.195 \Leftrightarrow [1.94,\ 2.76]$

본 자료에서 최소제곱법에 의한 회귀선은 $\ln y = 0.019 \times Age + 2.35$로 추정된다. 여기에서 기울기 0.019는 연령이 1단위 (개월) 증가할 때 BUN 농도가 ln 0.019 단위 증가함을 의미한다. 또한 신뢰구간이 0을 포함하지 않기 때문에 연령이 증가함에 따라 BUN 농도가 증가한다는 분명한 증거가 있다(P = 0.0017). 이 자료에서 BUN 농도는 대수변환된 값이므로 ln 0.019 단위를 원래의 단위로 환산하면 $e^{0.017}$ =

**Table 2.** Estimation of regression parameters, standard error (SE) and 95% confidence interval (CI) for the data in Table 1

| | Coefficient | SE of coefficient | t | p | CI |
|---|---|---|---|---|---|
| Intercept | 2.3479 | 0.1954 | 12.01 | < 0.0001 | 1.94 - 2.76 |
| Age | 0.0188 | 0.0051 | 3.69 | 0.0017 | 0.00 - 0.03 |

1.019 mg/dl가 된다. 다른 예로 50개월령 개의 대수변환 BUN 농도는 $2.35 + 0.019 \times 50 = 3.3$ 단위이고 이 값은 $e^{3.3} = 27.1$ mg/dl에 해당한다. 전술하였듯이 절편 2.35는 연령이 0일 때 BUN 농도이지만 관찰된 자료에서 연령 =0은 없기 때문에 실질적인 의미는 없다. 연령의 회귀계수에 대한 유의확률 $p = 0.0017$은 회귀계수 =0이라는 귀무가설을 기각하는 강한 증거이므로 BUN 농도와 연령 간에 선형관계가 있으며 이러한 결과는 신뢰구간이 0을 포함하지 않은 것으로도 알 수 있다. 절편 $\beta_0 = 0$에 대한 귀무가설을 기각하는 증거가 있으므로 ($p < 0.0001$) 회귀식에 포함되어야 한다.

### 결정계수

회귀분석을 통하여 얻을 수 있는 유용한 정보의 하나는 결정계수(coefficient of determination)로 y의 총 변동 중 회귀선으로 설명될 수 있는 변동의 비율이다. 이 값이 1에 가까울수록 y 변동의 대부분을 회귀모형으로 설명이 가능하다 것을 의미하며 아래의 공식이나 피어슨 상관계수의 제곱($r^2$)으로 계산한다. Spearman 상관계수를 제곱하여 결정계수를 계산하는 것은 잘못된 분석이다(7).

$$r^2 = \frac{regression\ \ sum\ \ of\ \ square}{total\ \ sum\ \ of\ \ square}$$

Table 1의 자료에 대한 결정계수는 $0.986/2.289 = 0.43$ ($0.656^2$)로 혈청 BUN 농도의 변동 중 43%만이 연령으로 설명되며 나머지 57% ($1 - r^2$)는 다른 요인에 기인한 변동이 있음을 시사한다.

### 예측

회귀식의 용도는 측정하지 않은 미래의 관찰치에 대한 예측(prediction)에 있으며 예측은 종속변수의 평균값에 대한 추정과 종속변수의 개별 관찰치에 대한 추정 등 두가지 목적으로 수행된다. 전자는 이를테면 Table 1의 자료에서 독립변수 즉 환자의 연령 x가 $x_0$일 때 종속변수인 평균 크레아티닌 농도($\mu_0$)에 대한 점추정치(point estimate, $\hat{y_0}$)를 알고자 하는 것이다.

$$\mu_0 = \beta_0 + \beta_1 x_0$$

$$\hat{y_0} = b_0 + b_1 x_0$$

반면에 후자는 환자의 연령 x가 $x_0$일 때 크레아티닌 농도의 (미래에 관찰하게 될) 개별 관찰치($y_0$)에 대한 점예측치(point prediction, $\hat{y_0}$)를 추정하는 것이다. 여기에서 $y_0 - \hat{y_0}$을 예측오차(prediction error)라고 한다.

$$y_0 = \mu_0 + \varepsilon_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$$

$$\hat{y_0} = b_0 + b_1 x_0$$

이러한 점추정치에는 언제나 불확실성이 존재한다. 즉 독립변수 x가 $x_0$일 때 종속변수의 평균 $\mu_0$에 대한 점추정치 $\hat{y_0}$는 $\mu_0$와 동일하지 않다는 것이다. 따라서 이러한 불확실성을 고려하기 위하여 흔히 구간추정치로 표현하며 종속변수의 평균값에 구간을 신뢰구간(confidence interval), 종속변수의 개별 관찰치에 대한 구간을 예측구간(prediction interval)이라고 한다.

### 첫째, 모집단의 평균 예측

예를 들어 Table 1의 자료에서 $x_0 = 44$일 때 평균 BUN 농도($\mu_0$)에 대한 신뢰구간을 계산하는 상황으로, 특정한 1두가 아니라 모집단에서 BUN농도의 평균값에 관심을 두는 경우다. 특정한 값 x에 대하여 예측된 y의 평균값은 모집단 평균에 대한 추정치가 되므로 모집단 평균의 신뢰구간을 계산할 수 있다. 독립변수 x의 특정한 값 $x_0$에 대한 표준오차와 95% 신뢰구간은 다음과 같다.

$$\hat{y_0} \pm t_{df=n-2} \times SE(\hat{y_0})$$

$$SE(\hat{y_0}) = s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

회귀식 $\ln y = 2.35 + 0.019 \times Age$에 환자의 연령 44개월을 대입하면 대수변환 BUN 값은 $y = 3.186$(mg/dl)이므로 원래 단위로 환산하면 24.2 mg/dl이 된다.

$$\mu_0 = \beta_0 + \beta_1 x_0 = \beta_0 + \beta_1(44)$$

$$\hat{y_0} = b_0 + b_1 x_0 \Leftrightarrow \ln\hat{y_0} = 2.35 + 0.019(44) = 3.186$$

$$e^{3.186} = 24.2$$

표준오차가 0.07145이므로 대수변환 점추정치의 95% 신뢰구간은 [2.60, 3.77]이고, 이 값을 역변환하면 BUN 농도는 [20.8, 28.1]로 계산된다.

$$SE(\hat{y_0}) = 0.07145$$

$$3.186 \pm 2.101 \times 0.07145 \leftrightarrow 3.186 \pm 0.150 \leftrightarrow [3.036,\ 3.336]$$

$$[e^{3.036},\ e^{3.336}] = [20.8,\ 28.1]$$

요약하면 $x_0 = 44$일 때 평균 BUN 농도($\mu_0$)의 점추정치는 24.2 mg/dl이고, 평균값이 20.8에서 28.1 mg/dl이라는 것을 95% 신뢰한다고 해석한다.

### 둘째, 종속변수의 개별 관찰치에 대한 예측

예측구간은 신뢰구간과 동일한 방법으로 계산되지만 표준오차가 다르다.

$$\hat{y_0} \pm t_{df=n-2} \times SE(\hat{y_0})$$

$$SE(\hat{y_0}) = s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Table 1의 자료에서 예를 들어 44개월령 개에서 측정된 BUN 수치가 23 mg/dl이라고 할 때 이 값의 예측구간을 계산하여 보자. 회귀식 $\ln y = 2.35 + 0.019 \times Age$에 환자의 연령 44개월을

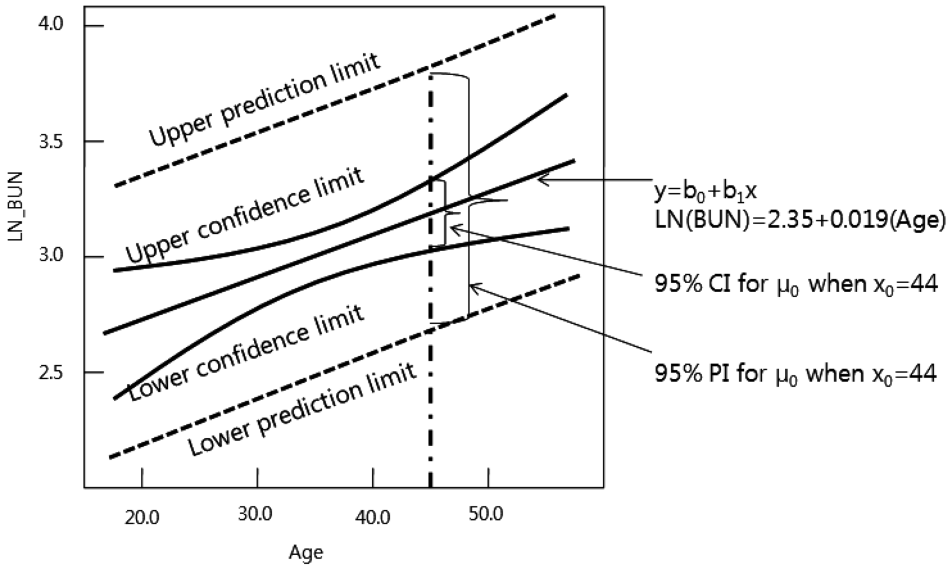


**Fig 8.** Regression line and 95% confidence and prediction interval for the data in Table 1. Note that the BUN level was logarithmic transformed prior to fitting. CI = confidence interval, PI = prediction interval.

대입하면 24.2 mg/dl이다. $\bar{x} = 36.45$; $x_p = 44$; $\Sigma_{i=1}^{n}(x_i - \bar{x})^2 = 2780.95$; $s = 0.2691$ 임계값 $t_{0.05/2,df=18} = 2.101$이므로 점추정치의 신뢰구간은 [2.60, 3.77]로 계산된다.

$$SE(\hat{y_0}) = 0.2784$$

$$3.186 \pm 2.101 \times 0.2784 \leftrightarrow 3.186 \times 0.5849 \leftrightarrow [2.60,\ 3.77]$$

$$[e^{2.60},\ e^{3.77}] = [13.5\ \ 434.4]$$

요약하면 $x_0 = 44$일 때 이 환자에서 기대할 수 있는 (미래의) BUN 농도($\mu_0$)의 점추정치는 24.2 mg/dl이고, 이러한 점추정치가 13.5 – 43.4 mg/dl 범위에 해당하는 것을 95% 신뢰한다고 해석한다.

이러한 계산에서 볼 때 평균 BUN 농도($\mu_0$)에 대한 신뢰구간 [20.8, 28.1]은 종속변수의 개별 관찰치($y_0 = \mu_0 + \varepsilon_0$)에 대한 예측구간 [13.5, 43.4]에 비하여 상대적으로 좁게 계산된다는 것을 알 수 있다. 이러한 차이는 특정 개체 $x$에 대하여 $y$ (BUN)를 추정하는 것과 비교할 때 다수의 개체에서 $y$의 평균을 추정할 때 정확도가 더 높다는 것을 의미한다. 특정한 값 x에 대하여 $y$의 평균에 대한 표준오차는 모든 x에 대하여 동일하지 않고, 회귀식을 작성하는데 사용한 자료의 평균($\bar{x}$)과 $x$간의 차이가 클수록 표준오차는 증가한다. 따라서 $\bar{x}$와 큰 차이를 보이는 특정한 값 x 즉 실험범위(experimental region)를 벗어난 x에 대하여 회귀식을 이용하여 예측하는 경우에는 정확성이 저하된다는 것을 주의해야 한다(7). 종속변수의 평균에 대한 신뢰구간과 개별 관찰치에 대한 예측구간은 Fig 8과 같다.

### 분석의 가정

상관분석과 회귀분석을 사용하기 위해서는 몇가지 가정을 전제로 한다. 첫째, 관찰 자료가 독립성을 만족해야 한다. 둘째, 상관분석의 경우 두 변수가 정규분포를 따르는 반면, 회귀분석에서는 종속변수가 정규분포이고 종속변수의 변동이 독립변수의 모든 값에 대하여 동일해야 한다. 셋째, 두 분석 모두 두 변수 간 선형관계를 가정한다. 흔히 이러한 가정의 충족여부는 일차적으로 산점도를 이용하여 평가하며, 스크리닝 과정을 거친 후 보다 정밀한 분석은 적합된 값 (y축)과 잔차 (x축)에 대한 잔차분석(residual analysis)으로 수행된다. 만일 선형관계와 자료변동이 일정하다는 가정이 충족되면 잔차는 모든 적합된 값에 대하여 0을 중심으로 균일하게 분포하며, 정규확률도표(normal plot)를 사용하여 평가할 수 있다. 즉 잔차가 표준정규분포 (평균 0, 표준편차 1)에 근사할 때 기대되는 값을 나타낸 것으로 잔차가 정규분포를 따르면 직선의 형태를 보인다. 회귀식을 이용하여 예측할 때 예측오차는 무작위 변동에 기인하지만 모형의 적합성이 결여되어 나타난 결과일 수 있기 때문에 특히 회귀식이 유도된 원시자료(raw data)의 범위를 벗어난 x값에 대하여 예측하는 것은 매우 신중해야 한다(7).

본 예에서는 연령과 BUN 농도의 관계에 대한 회귀식을 설명하였으며, 관심이 있는 독자는 BUN과 크레아티닌 농도의 관계에 대한 회귀식을 적합해보기 바란다.

## 감사의 글

본 연구는 강원대학교 동물의학종합연구소의 지원에 의해 이루어졌으며 이에 감사드립니다.

## 참 고 문 헌

1. Bewick V, Cheek L, Ball J. Statistical review 7: correlation and regression. Crit Care 2003; 7: 451-459.

2. Bland M, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; I: 307-310.
3. Daniel WW. Applied nonparametric statistics. 2nd ed. Thompson Information Publishing Group, Boston, 1990.
4. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. pp. 212-236, Wiley & Sons, New York, NY, 1981.
5. Glantz SA, Slinker BK. Primer of applied regression and analysis of variance. McGraw-Hill, New York, 1990.
6. Zar JH. Biostatistical analysis. 4th ed. pp. 381-386, Prentice-Hall International, UK, 1999.
7. Zou KH, Tuncali K, Silverman SG. Correlation and simple linear regression. Radiology 2003; 227: 617-622.