HAFTA 12

REGRESYON ANALİZİNE GİRİŞ

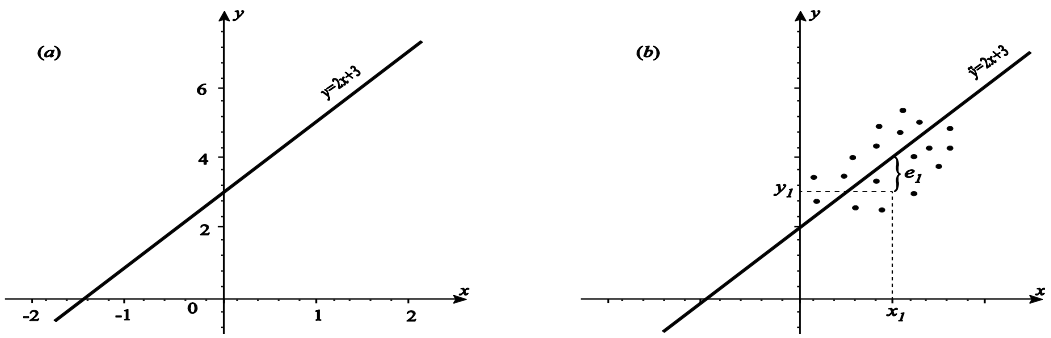
Kısım (7.9.4) de tahmin edicilerin bulunma yöntemleri incelenirken, en küçük kareler yönteminden kısaca bahsedilmişti. Herhangi iki değişken arasında, $y = f(x)$ gibi matematiksel bir ilişki varsa, x in her bir değerine karşılık y değeri bellidir. Oysa X ve Y rasgele değişkenler ise aynı $X = x$ için farklı Y değerleri gözlenebilir. Bu iki değişken arasında, e hata değişkenini göstermek üzere, $Y = f(x) + e$ şeklinde bir ilişki daha gerçekçidir. X in değerleri biliniyorsa, eşitlik $i = 1, 2, 3, ..., n$ için $Y_i = f(x_i) + e_i$ şeklinde yazılabilir. Bazı varsayımlar altında, bu eşitliğe *regresyon eşitliği* denir. Regresyonda önemli olan f fonksiyonunun belirlenmesidir. Bu bölümde, X ler bilinen (rasgele olmayan) değişkenler olarak $(X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)$ belirlendiğinde,

$$Y_i = \alpha_0 + \alpha_1 x_{1,i} + \alpha_2 x_{2,i} + ... + \alpha_p x_{p,i} + e_i \ , \ i = 1, 2, 3, ..., n$$

gibi bir model incelenmeye çalışılacaktır. Burada, Y_i ler bağımlı değişkeni (Y_i ler bağımsız rasgele değişkenler olup, bağımlılık Y_i lerin $x_{k,i}$ açıklayıcı değişkenlerine bağımlılığını ifade etmektedir), $x_{k,i}$, $k = 1, 2, 3, ..., p$ açıklayıcı değişkenleri (biliniyor, rasgele değil), e_i ler hata değişkenini ve $k = 0, 1, 2, ..., p$ için α_k ler de parametreleri göstermektedir.

10.1. Koşullu Beklenen Değer ve Regresyon

Regresyon, rasgele değişkenler arasındaki koşullu beklenen değerdir. Örneğin, $Y, X_1, X_2, \ldots, X_p$ rasgele değişkenlerinin ortak olasılık veya olasılık yoğunluk fonksiyonu bilindiğinde, $E(Y \mid X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)$ koşullu beklenen değeri, Y nin $X_1, X_2, \ldots, X_p$ değişkenleri üzerine regresyonudur. Bu koşullu beklenen değer, X lerin lineer bir fonksiyonu ise, lineer (veya doğrusal) regresyon, aksi halde lineer olmayan regresyondur. X ve Y gibi iki değişken arasında, $Y = f(X)$ şeklinde bir ilişki, f fonksiyonu biliniyorsa deterministik bir ilişkidir (Şekil (10.1.1a)). Örneğin, X ve Y gibi iki değişken arasında $Y = 2X + 3$ gibi fonksiyonel bir ilişkide $X = 1$ için $Y = 5$, $X = 0$ için $Y = 3$ ve $X = -1$ için $Y = 1$ değerleri elde edilir.



Şekil 10.1.1 $y = 2x + 3$ *doğrusunun grafiği*

Diğer taraftan, iki değişken arasındaki ilişki doğrusal olmak zorunda değildir. $Y = X^2 + 3$ veya $Y = 2X^2 + 3X + 7$ şeklinde parabolik bir ilişki de olabilir. Burada, X in seçilen bir değeri için Y nin değeri bellidir. Gerçek hayatta, değişkenlerden bir tanesi sabit tutularak (genellikle X) deney tekrar edilerek diğerinin değerleri gözlenir.

Aynı $X = x$ değeri için Y nin değerleri farklı gözlenebilir (Y ler bir hata ile gözlenir). Yani, Y ile x arasında, $Y = f(x) + e$ gibi (bilinmeyen f fonksiyonuna bağlı) bir ilişkiden söz edilir (Şekil 10.1.1b)). Böyle bir ilişkiye istatistiksel (stokastik) bir ilişki ya da Y nin x üzerine regresyonu denir. Regresyonda önemli olan bilinmeyen f fonksiyonunun bazı koşullar altında belirlenmesidir. Bu koşullar (veya varsayımlar) genellikle e hata terimi üzerindedir.

$X_1, X_2, ..., X_p$ değişkenlerinin değerleri $X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p$ şeklinde biliniyor olsun. x_i ler ile Y arasında $Y = f(x_1, x_2, ..., x_p) + e$ gibi istatistiksel bir ilişki, Y nin (rasgele), X_k açıklayıcı değişkenleri üzerine regresyonudur. Regresyonda, Y bağımlı değişken, $x_1, x_2, ..., x_p$ ler açıklayıcı değişkenlerin değerleri, e hata terimi olmak üzere, f bilinmeyen parametre içeren ve ilişkinin yapısını gösteren bir fonksiyondur.

Amaç, f fonksiyonunu hata kareler toplamı minimum olacak şekilde tahmin etmektir. $x_1, x_2, ..., x_p$ ler $X_1, X_2, ..., X_p$ değişkenlerinin değerleri olarak düşünüldüğünde, Y nin $X_1, X_2, ..., X_p$ ler üzerine regresyonu, $E(Y \mid X_1 = x_1, ..., X_p = x_p)$ şeklinde koşullu beklenen değerdir. Bu koşullu beklenen değer, bazen lineer bazen de lineer değildir. Buna göre regresyon denklemleri, lineer ve lineer olmayan regresyon olarak iki gruba ayrılır. Regresyon ile koşullu beklenen değer arasındaki ilişkiyi daha iyi görebilmek için aşağıdaki örneği inceleyelim.

Örnek 10.1.1 a) X ve Y rasgele değişkenlerinin ortak olasılık yoğunluk fonksiyonu

$$f(x,y) = \begin{cases} x e^{-x(y+1)} & , \quad x, y > 0 \\ 0 & , \quad d.y. \end{cases}$$

şeklinde verilsin. $X = x$ verildiğinde Y nin koşullu olasılık yoğunluk fonksiyonu,

$$f_{Y|X=x}(y \mid x) = \begin{cases} x e^{-xy} & , \quad x > 0, y > 0 \\ 0 & , \quad d.y. \end{cases}$$

olup Y nin X üzerine regresyonu,

$$E(Y \mid X = x) = \int_0^\infty y\, f(y \mid x)\, dy = \int_0^\infty y\, x\, e^{-xy}\, dy = \frac{1}{x}$$

şeklinde lineer olmayan bir denklemdir. Yani, $Y = f(x) + e$ şeklinde bir regresyon eşitliği için f fonksiyonu lineer değildir. Ancak böyle bir model de, $X^* = 1/X$ gibi bir dönüşüm ile $Y = a + b\,x^* + e$ şeklinde lineer regresyon modeli gibi yazılabilir.

b) X ve Y değişkenleri çok terimli binom dağılımına (multinomial) sahip olsunlar. Yani ortak olasılık fonksiyonu, $x = 0,1,2,...,n\,;\, y = 0,1,2,...,n\,;\, x + y \le n$ ve $0 < \theta_1, \theta_2 < 1$ olmak üzere

$$f(x,y) = \binom{n}{x, y, n-x-y} \theta_1^x \, \theta_2^y (1-\theta_1-\theta_2)^{n-x-y}$$

şeklinde olup,

$$\binom{n}{x, y, n-x-y} = \frac{n!}{x!\,y!(n-x-y)!}$$

dir (Örnek (2.5.6b). Ayrıca, $X \sim Binom(n, \theta_1)$ olmak üzere, $X = x$ verildiğinde Y nin koşullu olasılık fonksiyonu,

$$f(y \mid x) = \frac{f(x,y)}{f(x)} = \frac{(n-x)!}{y!(n-x-y)!} [\theta_2 / (1-\theta_1)]^y [1-\theta_2 / (1-\theta_1)]^{n-x-y}$$

dir. Yani koşullu dağılım, $\theta_1 + \theta_2 \le 1$ olmak üzere $Y \mid X = x \sim Binom(n-x,\, \theta_2 / (1-\theta_1))$ şeklinde Binomdur. Buradan Y nin X üzerine regresyonu, bu koşullu dağılımın beklenen değeri olup regresyon denklemi, $\alpha = n\theta_2 / (1-\theta_1)$ ve $\beta = -\theta_2 / (1-\theta_1)$ olmak üzere $E(Y \mid X = x) = \alpha + \beta x$ olarak bulunur. Burada da, Y nin X üzerine regresyonu lineerdir.

c) X ve Y rasgele değişkenleri iki boyutlu normal dağılıma sahip olsun. Yani ortak dağılım

$$\underset{\sim}{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \text{ ve } \Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$$

olmak üzere, $(X,Y) \sim N(\underset{\sim}{\mu},\, \Sigma)$ şeklinde olsun.

İki boyutlu normal dağılımın özelliklerinden, koşullu dağılımlar da normaldir. Yani, koşullu dağılım $Y \mid X = x \sim N(\mu_{y|x}, \sigma_{y|x})$ olup koşullu beklenen değer ve varyans

$$\mu_{y|x} = \mu_y + \rho(\sigma_y / \sigma_x)(x - \mu_x) \text{ ve } \sigma_{y|x} = \sigma_y^2 (1-\rho^2)$$

şeklindedir. Dolayısı ile Y nin X üzerine regresyonu,

$$E(Y \mid X = x) = \mu_y + \rho \frac{\sigma_y}{\sigma_x}(x - \mu_x) = \alpha + \beta x$$

şeklinde basit doğrusal regresyon denklemidir. Burada

$$\alpha = \mu_y - \rho \frac{\sigma_y}{\sigma_x} \mu_x \text{ ve } \beta = \rho \frac{\sigma_y}{\sigma_x}$$

için Y nin X üzerine regresyonu, $E(Y \mid X = x) = \alpha + \beta x$ şeklindedir. Yani, Y nin X üzerine regresyonu lineerdir.

d) X , Y ve Z rasgele değişkenlerinin ortak olasılık yoğunluk fonksiyonu, $0 < \alpha < 1$ için

$$f(x,y,z)=\begin{cases}1+\alpha(2x-1)(2y-1)(2z-1) & , \quad 0<x,y,z<1\\ 0 & , \quad d.y\end{cases}$$

olarak verilmiş olsun. Y nin X ve Z üzerine regresyon eşitliğini elde edelim. Önce,

$$f_{X,Z}(x,z)=\int\limits_{y\in D_Y} f(x,y,z)dy=\int_0^1(1+\alpha(2x-1)(2y-1)(2z-1))\,dy=1$$

olduğundan X ve Z nin marjinal ortak olasılık yoğunluk fonksiyonu,

$$f_{X,Z}(x,z)=\begin{cases}1 & , \quad 0<x,z<1\\ 0 & , \quad d.y.\end{cases}$$

olup, $X=x$ ve $Z=z$ verildiğinde Y nin koşullu olasılık yoğunluk fonsiyonu

$$f_{Y|X=x,Z=z}(y\,|\,x,z)=\frac{f(x,y,z)}{f_{X,Z}(x,z)}=\begin{cases}1+\alpha(2x-1)(2y-1)(2z-1) & , \quad 0<y<1\\ 0 & , \quad d.y\end{cases}$$

dir. Buradan, $X=x$ ve $Z=z$ verildiğinde Y nin koşullu beklenen değeri

$$E(Y\,|\,X=x,Z=z)=\int\limits_{y=0}^{1} y\,f_{Y|X=x,Z=z}(y\,|\,x,z)\,dy=\int_0^1 y\left[(1+\alpha(2x-1)(2y-1)(2z-1)\right]dy$$

$$=\frac{1}{2}+\frac{\alpha(2x-1)(2z-1)}{6}=\frac{1}{2}+\frac{\alpha}{6}+\frac{2\alpha xz}{3}-\frac{\alpha\,x}{3}-\frac{\alpha\,z}{3}$$

olarak hesaplanmıştır. Burada,

$$\alpha_0=\frac{1}{2}+\frac{\alpha}{6}\ ,\ \alpha_1=\alpha_2=-\frac{\alpha}{3}\ \text{ ve }\ \alpha_3=\frac{2\alpha}{3}$$

denirse, koşullu beklenen değer,

$$E(Y\,|\,X=x,Z=z)=\alpha_0+\alpha_1x+\alpha_2z+\alpha_3xz$$

şeklinde α_i parametrelerine göre lineer, x ve z ye göre lineer olmayan çoklu regresyon eşitliği elde edilir⊕

10.2. Basit Doğrusal Regresyon

Bu kısımda, X ve Y gibi iki değişken arasındaki koşullu beklenen değerin $E(Y\,|\,X=x)=a+b\,x$ şeklinde olduğu durum incelenecektir. $X_i=x_i$ verildiğinde x_i lere karşılık gözlenen Y_i değerleri kullanılarak X ile Y arasındaki doğrusal ilişki belirlenmeye çalışılacaktır. Kısaca, $Y_i=\alpha_0+\alpha_1\,x_i+e_i$, $i=1,2,3,...,n$ basit doğrusal regresyon modeli ele alınacaktır. Burada, Y_i ler bağımlı değişken, x_i ler ise açıklayıcı değişkendir. e_i ler de hata terimleri olup α_0 ile α_1 de parametrelerdir. Bu eşitliğe bir regresyon modeli denebilmesi için denklemin

i) x_i lerin rasgele olmayan bilinen değerler

ii) $E(e_i)=0$, $i \neq j$ için $Cov(e_i,e_j)=0$ ve $Var(e_i)=\sigma^2$

koşullarını sağlanması gerekir. İstatistiki sonuç çıkarımlar için e_i hata terimlerinin normal olduğu varsayılır. Bu regresyon modeli için $E(Y_i)=\alpha_0+\alpha_1 x_i$ ve $Var(Y_i)=\sigma^2$ dir.

Regresyonda en önemli aşama, α_0 ile α_1 parametrelerinin hata kareler toplamı minimum olacak şekilde tahmin edilmesidir. Hatalar pozitif ya da negatif olabilir. Bu sebeble hataların minimizasyonu yerine hata karelerinin minimizasyonu üzerinde durulur. Bu nedenle, uygulamada genellikle "*regresyon*" ifadesi yerine genellikle "*en küçük kareler yöntemi*" ifadesi de kullanılır. Bunun için,

$$Q(\alpha_0,\alpha_1)=\sum_{i=1}^{n} e_i^2=\sum_{i=1}^{n}(Y_i-\alpha_0-\alpha_1 x_i)^2$$

ifadesi minimum olacak şekilde α_0 ve α_1 parametreleri x ve Y lerin değerlerine göre tahmin edilir. Bu minimizasyon problemi için $Q(\alpha_0,\alpha_1)$ nin her iki değişkenine göre türevleri alınıp sıfıra eşitlenmesi ile,

$$\frac{\partial Q(\alpha_0,\alpha_1)}{\partial \alpha_0}=-2\sum_{i=1}^{n}(Y_i-\alpha_0-\alpha_1 x_i)=0 \quad , \quad \frac{\partial Q(\alpha_0,\alpha_1)}{\partial \alpha_1}=-2\sum_{i=1}^{n} x_i(Y_i-\alpha_0-\alpha_1 x_i)=0$$

eşitlikleri elde edilir. Buradan da,

$$\sum_{i=1}^{n}(Y_i-\alpha_0-\alpha_1 x_i)=0 \quad \Rightarrow \; n\alpha_0 \; +\alpha_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} Y_i$$

$$\sum_{i=1}^{n} x_i(Y_i-\alpha_0-\alpha_1 x_i)=0 \Rightarrow \alpha_0\sum_{i=1}^{n} x_i+\alpha_1\sum_{i=1}^{n} x_i^2=\sum_{i=1}^{n} x_i Y_i$$

denklemleri elde edilir. α_0 ve α_1 parametrelerinin en küçük kareler tahmin edicileri $\hat{\alpha}_0$ ve $\hat{\alpha}_1$ olmak üzere denklemler,

$$n\hat{\alpha}_0 \; +\hat{\alpha}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} Y_i \quad , \quad \hat{\alpha}_0\sum_{i=1}^{n} x_i+\hat{\alpha}_1\sum_{i=1}^{n} x_i^2=\sum_{i=1}^{n} x_i Y_i$$

şeklinde yazılır. Bu denklemlere "*normal denklemler"* denir. Normal denklemlerin çözümleri,

$$\hat{\alpha}_1=\left[\sum_{i=1}^{n}(x_i-\overline{x}_n)^2\right]^{-1}\left[\sum_{i=1}^{n}(x_i-\overline{x}_n)(Y_i-\overline{Y}_n)\right] \quad \text{ve} \quad \hat{\alpha}_0=\overline{Y}_n-\hat{\alpha}_1\overline{x}_n$$

şeklinde olup bu çözümler α_0 ve α_1 parametrelerinin *en küçük kareler tahmin edicileridir*. Bu tahmin edicilerin yansızlık etkinlik gibi istatistiksel özellikleri ileride incelenecektir. Parametrelerin en küçük kareler tahmin edicileri kullanılarak *kestirim denklemi,*

$$\hat{Y}_i=\hat{\alpha}_0+\hat{\alpha}_1 x_i \quad , \quad i=1,2,3,...,n$$

şeklinde yazılır. Kestirim denkleminden elde edilen kestirimler kullanılarak artıklar da $i=1,2,3,...,n$ için $\hat{e}_i=Y_i-\hat{Y}_i$ şeklinde hesaplanır. Bu kestirim ve artıklardan,

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y}_n) = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}_n^2\ ,\ SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y}_n) = \sum_{i=1}^{n} \hat{Y}_i^2 - n\bar{Y}_n^2\ ,\ \ SSE = SST - SSR$$

kareler toplamları hesaplanarak ANOVA (**AN**alysis **O**f **VA**riance) tablosu aşağıdaki gibi oluşturulur.

Kaynak	Serbestlik Derecesi	Kareler Toplamı	Ortalama Kareler Toplamı	F Değeri
Regresyon	1	SSR	$MSR = SSR/1$	$F = MSR/MSE$
Artıklar	$n-2$	SSE	$MSE = SSE/(n-2)$	
Toplam	$n-1$	SST		

Buradaki F değeri, parametrelerin veya öne sürülen modelin uygunluğunu sınamak için kullanılır. MSE istatistiğinin değeri de hata terimlerinin varyansı olan σ^2 nin (yansız) bir tahmin değeridir. Bununla birlikte, $R^2 = SSR/SST$ *oranı, "x lerin Y leri açıklama yüzdesidir* (coefficient of multiple determination)". Bu değer ne kadar büyük ise "model de o kadar iyidir" denebilir.

Tahmin edicilerinin istatistiki özelliklerine geçmeden, $Y_i = \alpha_0 + \alpha_1 x_i + e_i\ , i = 1, 2, ..., n$ şeklindeki basit doğrusal regresyon modelini

$$\underset{\sim}{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix},\ X = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ . & . \\ . & . \\ . & . \\ 1 & X_n \end{bmatrix},\ \underset{\sim}{\beta} = \begin{bmatrix} \alpha_0 \\ \\ \alpha_1 \end{bmatrix},\ \underset{\sim}{e} = \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix} \text{ için } \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ . & . \\ . & . \\ . & . \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \\ \alpha_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

biçiminde yazalım. Yani basit doğrusal regresyon modeli, $\underset{\sim}{y} = X\,\underset{\sim}{\beta} + \underset{\sim}{e}$ şeklinde yazılabilir. Bu yazılış ile normal denklemler, $X'X\,\underset{\sim}{\hat{\beta}} = X'\,\underset{\sim}{y}$ şeklinde olup çözümü $\underset{\sim}{\hat{\beta}} = (X'X)^{-1}X'\,\underset{\sim}{y}$ dir.

Basit doğrusal regresyon modeli $\underset{\sim}{y} = X\,\underset{\sim}{\beta} + \underset{\sim}{e}$ şeklinde verildiğinde varsayımlar; I_n, $n \times n$ boyutlu birim matrisi göstermek üzere, X (tasarım matrisi) biliniyor ve $E(\underset{\sim}{e}) = \underset{\sim}{0}, Var(\underset{\sim}{e}) = \sigma^2 I_n$ şekline dönüşür. Bu durumda, $E(\underset{\sim}{y}) = X\,\underset{\sim}{\beta}$ ve $Var(\underset{\sim}{y}) = \sigma^2 I_n$ dir. Böylece, parametrelerin en küçük kareler tahmin edicisinin beklenen değer ve varyansı sırası ile,

$$E(\underset{\sim}{\hat{\beta}}) = E[(X'X)^{-1}X'\,\underset{\sim}{y}] = (X'X)^{-1}X'E(\underset{\sim}{y}) = (X'X)^{-1}X'X\,\underset{\sim}{\beta} = \underset{\sim}{\beta}$$

ve

$$Var(\hat{\underset{\sim}{\beta}}) = Var[(X'X)^{-1}X'\underset{\sim}{y}] = (X'X)^{-1}X'Var(\underset{\sim}{y})\,X(X'X)^{-1}$$
$$= (X'X)^{-1}X'(\sigma^2 I_n)\,X(X'X)^{-1} = \sigma^2(X'X)^{-1}$$

şeklinde olur. Yani, $\hat{\underset{\sim}{\beta}} = (X'X)^{-1}X'\underset{\sim}{y}$ en küçük kareler tahmin edicisi $\underset{\sim}{\beta}$ için yansızdır. Ayrıca $\hat{\underset{\sim}{\beta}}$, $\underset{\sim}{y}$ ye göre lineer olup $\underset{\sim}{\beta}$ nın lineer yansız tahmin edicisidir. Diğer taraftan, verilen regresyon modeli Teorem (8.3.4) de verilen Gauss-Markov Teoreminin koşullarını sağlar. Dolayısı ile, bütün $\underset{\sim}{\lambda}$ lar için $\underset{\sim}{\lambda}'\underset{\sim}{\beta}$ tahminlenebilir olup, $\underset{\sim}{\lambda}'\hat{\underset{\sim}{\beta}}$ bütün lineer yansız tahmin ediciler arasında en küçük varyanslıdır. Yani, $\hat{\alpha}_0$ ve $\hat{\alpha}_1$ bütün lineer yansız tahmin ediciler arasında en küçük varyanslıdır. Bir başka ifade ile, $\hat{\alpha}_0$ ve $\hat{\alpha}_1$ "e*n iyi lineer yansız tahmin ediciler"* (**B**est **L**ineer **U**nbiased **E**stimator, ***BLUE***) dir. Kısaca, $\underset{\sim}{\beta}$ nın BLUE tahmin edicisi $\hat{\underset{\sim}{\beta}} = (X'X)^{-1}X'\underset{\sim}{y}$ dir.

Şimdi, hata terimlerinin normal dağılıma sahip olduğunu varsayalım. İlişkisiz ($i \neq j$ için $Cov(e_i, e_j) = 0$) normal dağılıma sahip rasgele değişkenler bağımsızdır. Böylece, Y_i ler de normaldir. Yani $\underset{\sim}{y} \sim MN(X\underset{\sim}{\beta}, \sigma^2 I_n)$ olup, en küçük kareler tahmin edicisi de $\hat{\underset{\sim}{\beta}} \sim MN(\underset{\sim}{\beta}, (X'X)^{-1}\sigma^2)$ şeklinde iki boyutlu normaldir. Çok değişkenli normal dağılımın özelliklerinden, bileşenlerinin de normal olduğunu biliyoruz. O halde, $i = 0,1$ için $(X'X)^{-1}_{i+1,i+1}$, $(X'X)^{-1}$ matrisinin $(i+1)$. satır ve $(i+1)$. sütun elemanını göstermek üzere, en küçük kareler tahmin edicileri de $i = 0,1$ için

$$\hat{\alpha}_i \sim N(\alpha_i, (X'X)^{-1}_{i+1,i+1}\sigma^2)$$

şeklinde normal dağılır. MSE nin σ^2 nin yansız bir tahmin edicisi olduğunu söylemiştik. Dolayısı ile, $i = 0,1$ için, $S_n^2(\hat{\alpha}_i) = MSE\,(X'X)^{-1}_{i+1,i+1}$ olmak üzere, $\hat{\alpha}_i$ ile MSE bağımsız olup $(\hat{\alpha}_i - \alpha_i)/S_n(\hat{\alpha}_i) \sim t_{n-2}$ dir. Buna göre, parametreler için hipotez testleri yapılabilir, güven aralıkları oluşturulabilir. $H_0 : \alpha_i = \alpha_{i,0}$ hipotezi $H_a : \alpha_i \neq \alpha_{i,0}$ alternatif hipotezine karşı test edilmek istendiğinde $t_h = (\hat{\alpha}_i - \alpha_{i,0})/s(\hat{\alpha}_i)$ olmak üzere, $|t_h| > t_{n-2}(\alpha/2)$ ise $H_0 : \alpha_i = \alpha_{i,0}$ hipotezi $\alpha-$anlam düzeyinde $H_a : \alpha_i \neq \alpha_{i,0}$ alternatif hipotezine karşı red edilir. Benzer şekilde tek yönlü hipotezler de test edilebilir. Altenatif hipotez $H_a : \alpha_i > \alpha_{i,0}$ şeklinde ise, $t_h > t_{n-2}(\alpha)$ için $H_0 : \alpha_i = \alpha_{i,0}$ hipotezi red edilir. $H_a : \alpha_i < \alpha_{i,0}$ alternatif hipotezi için " $t_h < -t_{n-2}(\alpha)$ ise $H_0 : \alpha_i = \alpha_{i,0}$ hipotezi red edilir". Parametreler için güven aralıkları da $\hat{\alpha}_i \pm s(\hat{\alpha}_i)t_{n-2}(\alpha/2)$ şeklindedir.

Parametreler tahmin edildikten sonra kestirimler yapılır. Kestirimlerin $\hat{Y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 x_i$ şeklinde hesaplandığını biliyoruz. Yani kestirim vektörü $\underset{\sim}{\hat{y}} = X\underset{\sim}{\hat{\beta}}$ olup hata terimlerinin normalliği varsayımı altında $\underset{\sim}{\hat{y}}$ nın dağılımı, $\underset{\sim}{\hat{y}} \sim MN(X\underset{\sim}{\beta}, X(X'X)^{-1}X'\sigma^2)$

dir. Buradan kestirimler için de istatistiki sonuç çıkarımlar yapılabilir. Örneğin, $X = x_h$ için bir kestirim $\underset{\sim}{x}_h = (1, x_h)'$ olmak üzere $\hat{y}_h = \underset{\sim}{x}_h' \underset{\sim}{\hat{\beta}}$ dir. Ayrıca,

$$E(\hat{y}_h) = E(\underset{\sim}{x}_h' \underset{\sim}{\hat{\beta}}) = \underset{\sim}{x}_h' \underset{\sim}{\beta} \quad \text{ve} \quad Var(\hat{y}_h) = Var(\underset{\sim}{x}_h' \underset{\sim}{\hat{\beta}}) = \underset{\sim}{x}_h' (X'X)^{-1} \underset{\sim}{x}_h \sigma^2$$

$s^2(\hat{y}_h) = \underset{\sim}{x}_h' (X'X)^{-1} \underset{\sim}{x}_h MSE$ den y_h için bir güven aralığı $\hat{y}_h \pm s(\hat{y}_h) t_{n-2}(\alpha/2)$ şeklinde oluşturulur.

Kestirimler hakkında istatistiki sonuç çıkarımlar yerine, örneklemin dışında belirlenen yeni bir $x_{h,yeni}$ değerine karşılık gelen Y rasgele değişkeninin değeri de kestirilebilir. Böyle bir problem için, Y_h rasgele değişkeni, $Y_1, Y_2, \ldots, Y_n$ lerden bağımsızdır. Parametre tahminleri $Y_1, Y_2, \ldots, Y_n$ örneklemine bağlı olarak yapıldığı için hesaplanan $\hat{Y}_{h,yeni}$ değeri örneklemin bir fonksiyonudur. Yani, $Y_{h,yeni}$ örneklemden bağımsız yeni bir rasgele değişkendir. Buna göre,

$$\hat{Y}_h - \hat{Y}_{h,yeni} \sim N(0, Var(\hat{Y}_h) + Var(\hat{Y}_{h,yeni}))$$

veya

$$\hat{Y}_h - \hat{Y}_{h,yeni} \sim N(0, \sigma^2[\underset{\sim}{x}_h'(X'X)^{-1}\underset{\sim}{x}_h + 1])$$

yazılır. Ayrıca,

$$(\hat{Y}_h - \hat{Y}_{h,yeni}) / \sqrt{MSE[\underset{\sim}{x}_h'(X'X)^{-1}\underset{\sim}{x}_h + 1]} \sim t_{n-2}$$

olup, $Y_{h,yeni}$ için bir güven aralığı da

$$\hat{Y}_h \pm t_{n-2}(\alpha/2)\sqrt{MSE[\underset{\sim}{x}_h'(X'X)^{-1}\underset{\sim}{x}_h + 1]}$$

şeklinde oluşturulur.

Şimdi, buraya kadar yapılanları özetlemek için aşağıdaki örneği inceleyelim. Buradaki veriler tamamen uydurma verilerdir.

Örnek 10.2.1 Bir fabrikada çalışan işçi sayısı (X) ile fabrikanın üretim miktarı (ton olarak) arasında basit doğrusal bir ilişkinin var olduğunu kabul edelim. Bu modele uygun olduğu varsayılan aşağıdaki verileri kullanarak buraya kadar yapılanları özetleyelim.

X	40	80	120	160	200
Y	440	450	690	820	930

Bu verilere $Y_i = \alpha_0 + \alpha_1 x_i + e_i$, $i = 1,2,3,4,5$ şeklinde bir model uygun olsun.

a) Parametrelerin en küçük kareler tahmin değerleri için bazı özet bilgiler:

$$\sum_{i=1}^{n} x_i = 600 \ , \ \bar{x}_n = 120, \ \sum_{i=1}^{n} x_i^2 = 88000, \ \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 = 16000 \ , \ \sum_{i=1}^{n} x_i y_i = 453600$$

$$\sum_{i=1}^{n} y_i = 3330 \ , \ \bar{y}_n = 666, \ \sum_{i=1}^{n} y_i^2 = 2409500, \ \sum_{i=1}^{n} (y_i - \bar{y}_n)^2 = 191720$$

olarak hesaplanmıştır. Buna göre modelin eğiminin (α_1) en küçük kareler tahmin değeri,

$$\hat{\alpha}_1 = \left[\sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \right]^{-1} \left[\sum_{i=1}^{n} (x_i - \bar{x}_n)(y_i - \bar{y}_n) \right] = \left[\sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \right]^{-1} \left[\sum_{i=1}^{n} x_i y_i - n\bar{x}_n \bar{y}_n \right]$$

$$= [16000]^{-1} [453600 - 5(120)(666)] = 54000 / 16000 = 3.375$$

ve modelin kesim noktasının (intercept, α_0) en küçük kareler tahmin değeri de,

$$\hat{\alpha}_0 = \bar{y}_n - \hat{\alpha}_1 \bar{x}_n = 666 - (3.375)(120) = 261$$

olarak hesaplanmıştır. Yani, kestirim denklemi $\hat{Y}_i = 261 + 3.375\, x_i$ dir.

b) $\hat{Y}_i = 261 + 3.375\, x_i$ kestirim denklemi kullanılarak kestirim ve artık değerler,

$$\hat{Y}_1 = 261 + 3.375\, x_1 = 261 + 3.375(40) = 396 \quad , \quad \hat{e}_1 = Y_1 - \hat{Y}_1 = 440 - 396 = 44$$

$$\hat{Y}_2 = 261 + 3.375\, x_2 = 261 + 3.375(80) = 531 \quad , \quad \hat{e}_2 = Y_2 - \hat{Y}_2 = 450 - 531 = -81$$

$$\hat{Y}_3 = 261 + 3.375\, x_3 = 261 + 3.375(120) = 666 \quad , \quad \hat{e}_3 = Y_3 - \hat{Y}_3 = 690 - 666 = 24$$

$$\hat{Y}_4 = 261 + 3.375\, x_4 = 261 + 3.375(160) = 801 \quad , \quad \hat{e}_4 = Y_4 - \hat{Y}_4 = 820 - 801 = 19$$

$$\hat{Y}_5 = 261 + 3.375\, x_5 = 261 + 3.375(200) = 936 \quad , \quad \hat{e}_5 = Y_5 - \hat{Y}_5 = 930 - 936 = -6$$

şeklindedir. Hesaplanan bu değerlerden,

$$\sum_{i=1}^{n} \hat{e}_i = 0 \ , \ \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i = 3330 \ , \ \sum_{i=1}^{n} x_i \hat{e}_i = 0 \ \text{ ve } \ \sum_{i=1}^{n} \hat{y}_i \hat{e}_i = 0$$

olduğu görülür. Yani $\bar{y}_n = \bar{\hat{y}}_n$ dir. Model kesim noktası (α_0) içerdiği sürece bu eşitlikler her zaman doğrudur. Bu sonuçlar, ileride göreceğimiz çoklu regresyon modelleri için de geçerli olup normal denklemlerin bir sonucudur (Probloem (10.5.1)).

c) ANOVA tablosunu hazırlayalım. Bunun için gerekli toplamlar,

$$SST = \sum_{i=1}^{n} (y_i - \bar{y}_n)^2 = 191720 \ , \ SSR = \sum_{i=1}^{n} \hat{y}_i^2 - n\bar{y}_n^2 = 2400030 - 5(666)^2 = 182250 ,$$

$$SSE = SST - SSR = 191720 - 182250 = 9470$$

şeklinde hesaplanmıştır. ANOVA tablosu aşağıdaki gibi olup $R^2 = SSR / SST \cong 0.95$ dir. Fabrikanın üretim miktarı yaklaşık %95 oranında çalışan işçi sayısı ile açıklanmaktadır.

Kaynak	Serbestlik Derecesi	Kareler Toplamı	Ortalama Kareler Toplamı	F Değeri
Regresyon	1	$SSR = 182250$	$MSR = 182250$	$F = 57.74$
Artıklar	3	$SSE = 9470$	$MSE = 3156.67$	
Toplam	4	$SST = 191720$		

d) Parametreler için %95 lik güven aralığı için standart hataların

$s_n^2(\hat{\alpha}_i) = MSE\,(X'X)^{-1}_{i+1,i+1}$, $i = 0,1$

ile hesaplandığını biliyoruz. Buradan $X'X$ matrisi ile bu matrisin tersi

$$XX = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} = \begin{bmatrix} 5 & 600 \\ 600 & 88000 \end{bmatrix},\ (X'X)^{-1} = \begin{bmatrix} 1.1 & -0.0075 \\ -0.0075 & 0.0000625 \end{bmatrix}$$

şeklinde hesaplanmıştır. $\alpha = 0.05$ için $t_3(0.025) = 3.182$ olup standart hatalar

$s_n^2(\hat{\alpha}_0) = MSE\,(X'X)^{-1}_{11} = (3156.67)(1.1) = 3472.23$, $s_n(\hat{\alpha}_0) = \sqrt{3472.23} = 58.92$

$s_n^2(\hat{\alpha}_1) = MSE(X'X)^{-1}_{22} = (3156.67)(0.0000625) \cong 0.1973$, $s_n(\hat{\alpha}_1) = \sqrt{0.1973} \cong 0.444$

olarak bulunmuştur. Buna göre parametreler için güven aralıkları,

α_0 için %95 lik güven aralığı : $\hat{\alpha}_0 \pm s_n(\hat{\alpha}_0)\,t_{n-2}(\alpha/2)$ den $(73.52,\ 448.48)$

α_1 için %95 lik güven aralığı : $\hat{\alpha}_1 \pm s_n(\hat{\alpha}_1)\,t_{n-2}(\alpha/2)$ den $(1.962, 4.788)$

şeklinde elde edilmiştir.

e) Fabrika sahibi 250 işçi çalıştırmış olsaydı, yapılacak üretim miktarı için bir kestirimde bulunup, bu kestirim için %95 lik güven aralığı oluşturalım. Önce, $x = 250$ için Y nin bir kestirimi $\hat{y}_{x=250} = 261 + 3.375(250) = 1104.75$ dir. Bu kestirimin standart hatası da $\underset{\sim}{x}_h = (1, 250)'$ olmak üzere, $s_n^2(\hat{Y}_{x=250}) = MSE(1 + \underset{\sim}{x}_h'(X'X)^{-1}\underset{\sim}{x}_h) = 7122.23$ den $s_n(\hat{y}_{x=250}) = \sqrt{7122.23} \cong 84.4$ dir. %95 lik güven aralığı ise, $t_3(0.025) = 3.182$ olmak üzere, $\hat{y}_{x=250} \pm s_n(\hat{y}_{x=250})\,t_{n-2}(\alpha/2)$ de değerler yerine yazılırsa (836.19,1373.31) olarak elde edilmiştir. Yani, yukarıdaki veriler için regresyon varsayımları geçerli ise, fabrika sahibi 250 işçi çalıştırdığında yapılacak üretim miktarının %95 ihtimalle 836 ile 1374 ton arasında olması beklenir ⊕