MIT OpenCourseWare
http://ocw.mit.edu

14.30 Ekonomide İstatistiksel Yöntemlere Giriş
Bahar 2009

Bu materyale atıfta bulunmak ve kullanım koşulları için http://ocw.mit.edu/terms sayfasını ziyaret ediniz.

14.30 Ekonomide İstatistiksel Yöntemlere Giriş
Ders Notları 13

Konrad Menzel

31 Mart 2009

1. Kovaryans

X ile Y'nin kovaryansı iki rasgele değişken arasındaki ilişkinin gücünün ölçüsüdür.

Tanım 1. *İki rasgele değişken X ve Y için, kovaryans şöyle tanımlanır:*

$$\mathrm{Cov}(X,Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$

İlk olarak, sadece tanımı uygulayarak aşağıdakileri elde ederiz.

Özelik 1.

$$\mathrm{Cov}(X, X) = \mathrm{Var}(X)$$

Özelik 2.

$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$$

Dahası, elimizde kovaryans hesaplamasında çok yararlı olan aşağıdaki sonuç var.

Özelik 3.

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Bu, varyansın benzer özeliğinin genelleştirilmişidir ve ispatı da aynı tür argümanları kullanır. Bir örnek ile bu sonucun nasıl yararlı olduğunu görelim:

Örnek 1. *Varsayalım ki X ile Y'nin bileşik p.d.f.'si aşağıdaki gibidir.*

$$f_{XY}(x,y) = \begin{cases} 8xy & \text{eğer } \; 0 \le x \le y \le 1 \; \text{ ise} \\ 0 & \text{diğer bütün durumlarda} \end{cases}$$

Kovaryans Cov(X, Y) nedir? – Özelik 7'deki denklemin sağ tarafına göre dahil olan bileşenleri hesaplayalım:

$$\begin{aligned}\mathbb{E}[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{XY}(x,y)dxdy = \int_0^1\int_x^1 8x^2y^2dydx \\ &= \int_0^1 8x^2\left(\int_x^1 y^2dy\right)dx = \int_0^1 \frac{8}{3}x^2(1-x^3)dx \\ &= \frac{8}{3}\int_0^1 (x^2-x^5)dx = \frac{8}{3}\left[\frac{x^3}{3}-\frac{x^6}{6}\right]_0^1 \\ &= \frac{8}{3}\left(\frac{1}{3}-\frac{1}{6}\right) = \frac{4}{9}\end{aligned}$$

Aynı şekilde, yukarıdaki adımları takip ederek şunları elde ederiz:

$$\begin{aligned}\mathbb{E}[Y] &= \int_0^1 8x\left(\int_x^1 y^2dy\right)dx = \int_0^1 \frac{8}{3}x(1-x^3)dx \\ &= \frac{8}{3}\int_0^1 (x-x^4)dx = \frac{8}{3}\left[\frac{x^2}{2}-\frac{x^5}{5}\right]_0^1 = \frac{8}{3}\cdot\frac{3}{10} = \frac{4}{5}\end{aligned}$$

ve

$$\begin{aligned}\mathbb{E}[X] &= \int_0^1 8x^2\left(\int_x^1 ydy\right)dx = \int_0^1 \frac{8}{2}x^2(1-x^2)dx \\ &= 4\int_0^1 (x^2-x^4)dx = 4\left(\frac{1}{3}-\frac{1}{5}\right) = \frac{8}{15}\end{aligned}$$

Bütün parçaları biraya getirerek ve özelik 7'yi uygulayarak şu sonuca varırız:

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \frac{4}{9} - \frac{8}{15}\cdot\frac{4}{5} = \frac{4\cdot 25 - 32\cdot 3}{225} = \frac{4}{225}$$

İki bağımsız rasgele değişken X ve Y için toplamların varyansının varyansların toplamına eşit olduğunu daha önce göstermiştik. Şimdi ise bağımsız olması gerekli olmayan rasgele değişkenlerin genelleştirilmesini görelim:

Özelik 4.

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$$

İspatın arkasındaki düşünce, özelik 3 ve 7'yi uygulayarak aşağıdakini elde etmektir.

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathbb{E}[(X+Y)^2] - \mathbb{E}[X+Y]^2 \\ &= \mathbb{E}[X^2+2XY+Y^2] - \mathbb{E}[X]^2 - 2\mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]^2 \\ &= \left(\mathbb{E}[X^2]-\mathbb{E}[X]^2\right) + \left(\mathbb{E}[Y^2]-\mathbb{E}[Y]^2\right) + 2\left(\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]\right) \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)\end{aligned}$$

Özelik 5. *X, Y, Z rasgele değişkenler için,*

$$\mathrm{Cov}(X, aY+bZ+c) = a\mathrm{Cov}(X,Y) + b\mathrm{Cov}(X,Z)$$

Özelik 6.

$$\mathrm{Cov}(aX+b, cY+d) = ac\mathrm{Cov}(X,Y)$$

Son özeliğe göre, kovaryans X ile Y'nin ölçeğine göre değiştiği için, X ile Y arasındaki ilişkinin gücünü veren, iki değişkenin diyelim ki ölçüm birimindeki değişiminden etkilenmeyen, standart bir ölçüye sahip olmak istiyoruz. Çok sıklıkla kullanılan o ölçü *korelasyon katsayısı*dır.

Tanım 2. *X ve Y'nin korelasyon katsayısı şöyle hesaplanır:*

$$\rho(XY) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

Korelasyon katsayısı bir şekilde normalleşir (bkz. Özelik 7).

Özelik 7.

$$-1 \leq \rho(X,Y) \leq 1$$

Genel konuşmak gerekirse, üç durumu ayırt ederiz.

- ϱ(X, Y) > 0: "X ile Y arasında pozitif korelasyon vardır"
- ϱ(X, Y) = 0: "X ile Y arasında korelasyon yoktur"
- ϱ(X, Y) < 0: "X ile Y arasında negatif korelasyon vardır"

Özelik 8.

Bazı a ≠ 0 ve b sabitleri için

$$|\rho(X,Y)| = 1 \Leftrightarrow Y = aX + b$$

Yani, eğer iki rasgele değişken arasında deterministtik *doğrusal* bir ilişki varsa, korelasyon katsayısının mutlak değeri 1'e eşittir. O durumda X ile Y arasında *mükemmel korelasyon* olduğunu söyleriz.

Açıklama 1. *Veri analizinin çok önemli bir ilkesi, iki rasgele değişken arasındaki istatiksel ilişkinin mekanik veya nedensel ifadelere dayanmamasıdır ki, biz bunun gerçekte veriye dayalı olmasını arzularız. Örneğin, insanların jimnastik salonlarında spor yaparak harcadıkları zaman ile sağlıkları arasında pozitif bir korelasyon olduğunu veri setleriyle gözlemleriz, ancak bu sporun sağlık durumunu iyileştirdiği anlamına gelmez. Diğer taraftan sağlık durumları çok kötü olan ve spor salonlarına gitmeyi akıllarına bile getirmeyen kimi insanlarda vardır.*

Neden X ile Y'nin korelasyonun X'in Y'nin nedeni olmasından tamamen farklı kavramlar olduğunu görmenin daha soyut bir yolu, X ile Y'nin kovaryansının simetrik olduğuna, böylece değişkenlerin rollerini değiştirebileceğimize dikkat etmektir. Ancak nedensellik için, ilişkinin spesifik bir yönünü düşünürüz, yani X → Y veya "X Y'nin nedenidir/etkileyendir", ancak aynı zamanda "Y X'in nedeni/etkileyeni değildir" deriz, böylece X ile Y'nin rollerini değiştiremeyiz. Bunda ötürü,

KORELASYON NEDENSELLİĞE EŞİT DEĞİLDİR

Ekonometri derslerinde bunun (çok) daha fazlası vardır.

1.1.Önizleme: Regresyon

Diyelim ki, bir işçinin geliri Y ile onun okullaşma yılı X ile ölçülen eğitimi arasındaki ilişkiyle ilgileniyoruz (kolaylık açısından her ikisinin de sürekli olduğunu varsayalım). Bu durumda her zaman X ile Y arasındaki ilişkiyi şöyle yazabiliriz:

$$Y = \alpha + \beta X + U$$

Burada U E|U| = 0 ve Cov(X, U) = 0 özeliklerine sahip rasgele bir değişkendir(regresyon kapsamında bu *hata terimi* olarak adlandırılacaktır).

(α, β) parametrelerini belirlemenin bir yolu aşağıdakini çözmektir.

$$(\alpha, \beta) = \arg\min_{\alpha,\beta} \mathbb{E}[(Y - X\beta - \alpha)^2]$$

β'ya göre birinci-derece koşulunu belirleyince (türevini almak) (dikkat: beklenen değerler doğrusaldır, bu nedenle türevi integral aracılığıyla alabiliriz) aşağıdakini elde ederiz:

$$
\begin{aligned}
0 &= \frac{d}{d\beta}\mathbb{E}[(Y - X\beta - \alpha)^2] \\
&= \frac{d}{d\beta}\int (y - x\beta - \alpha)^2 f_{XY}(x,y)dydx \\
&= \int \frac{d}{d\beta}\left[(y - x\beta - \alpha)^2\right] f_{XY}(x,y)dydx \\
&= \mathbb{E}\left[\frac{d}{d\beta}[(Y - X\beta - \alpha)^2]\right] \\
&= \mathbb{E}\left[2X(Y - X\beta - \alpha)\right]
\end{aligned}
$$

Aynı şekilde, α'ya göre birinci-derece koşulu şöyledir:

$$
0 = \frac{d}{d\alpha}\mathbb{E}[(Y - X\beta - \alpha)^2] = \mathbb{E}\left[\frac{d}{d\alpha}[(Y - X\beta - \alpha)^2]\right] = \mathbb{E}[2(Y - X\beta - \alpha)]
$$

Son ifadeyi α için çözünce aşağıdaki şu edilir:

$$
\alpha = \mathbb{E}[Y] - \mathbb{E}[X]\beta
$$

Bunu b için elde edilen birinci derece koşulda yerine koyunca,

$$
0 = \mathbb{E}\left[X(Y - X\beta - (\mathbb{E}[Y] - \mathbb{E}[X]\beta))\right] = \mathbb{E}[XY] - \mathbb{E}[X^2]\beta - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]^2\beta
$$

elde ederiz, böylece artık parametre için çözümü yapabiliriz:

$$
\beta = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[X^2] - \mathbb{E}[X]^2} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}
$$

Şimdi şunları doğrulayabiliriz:

$$
\mathbb{E}[U] = \mathbb{E}[Y - X\beta - \alpha] = 0
$$

ve

$$
\mathrm{Cov}(X,U) = \mathrm{Cov}(X, Y - X\beta - \alpha) = 0
$$

(doğurusu, ilki doğrudan α'nın birinci derece koşulundan, ikincisi ise β'nın birinci derece koşulundan elde edilir).

O zaman, "tahmin edilen regresyon" $\alpha + \beta X$ X ile ilişkili (X tarafından açıklanan) Y'nin parçasıdır ve U ise Y'nin X ile ilişkili olmayan parçasıdır. α ve β parametreleri genellikle *regresyon parametreleri* veya *en küçük kareler katsayıları* olarak adlandırılır. Doğrusal regresyon ekonometrinin “temel taşı”dır ve bunu 14.32 ve diğer ekonometri derslerinde çok değişik varyasyonlarda göreceksiniz.

2. Koşullu Beklenen Değerler

Örnek 2. *Her sene, bir firmanın AR-GE bölümü rasgele bir süreç sonucunda X kadar buluş üretmektedir, burada E|X| = 2 ve Var(X) = 2’dir. Her buluş p = 0.2 olasılıkla ticari bir başarı gösterecektir (bağımsızlık varsayalım). Bir yıl içerisindeki ticari başarı sayısını S ile gösterelim. Bir yıl içerisinde X = x buluş sayısına koşullanmış S’nin ortalamasının S ~ B(x, p) = xp olduğunu bildiğimiz için, buluşların ortalama olarak xp kadarının başarılı olması gerekir.*

Y veriyken X’in koşullu beklenen değeri koşullu p.d.f’den elde edilen X’in beklenen değeridir:

Tanım 3.

$$\mathbb{E}[Y|X] = \begin{cases} \sum_y y f_{Y|X}(y|X) & \text{eğer Y kesikli ise} \\ \int_{-\infty}^{\infty} y f_{Y|X}(y|X) dy & \text{eğer Y sürekli ise} \end{cases}$$

$f_{Y|X}(y|X)$ *rasgele değişken* X’i kendi argümanı olarak taşıdığı için, koşullu beklenen değerin de aynı zamanda rasgele bir değişken olduğunu not ediniz. Ancak, X’in belli bir değeri veriyken, Y’nin koşullu beklenen değerini de tanımayabiliriz

$$\mathbb{E}[Y|X = x] = \begin{cases} \sum_y y f_{Y|X}(y|\text{x}) & \text{eğer Y kesikli ise} \\ \int_{-\infty}^{\infty} y f_{Y|X}(y|\text{x}) dy & \text{eğer Y sürekli ise} \end{cases}$$

burada koşullu yoğunluk tanımlandığı sürece bu herhangi bir verili *x* değeri için sadece bir sayıdır.

Hesaplamalar tam olarak önceki gibi yapıldığı için (tek fark şimdi *koşullu* dağılım üzerinden integral alıyoruz), sayısal bir örnek yapmayacağız (problem seti için sadece tanımı uygulayınız). Bunun yerine, koşullu ve koşulsuz örnekler arasındaki farkı göstermek için daha kalitatif örnekleri tartışalım.

Örnek 3 *(“Limon” Piyasası)*. *Aşağıdaki ekonomist George Akerlof’un meşhur kullanılmış araba piyasası modelinin basitleştirilmiş bir versiyonudur. Varsayalım ki üç tür X kullanılmış araba vardır: mükemmel durumdaki arabalar (karpuzlar), orta kalite arabalar (istatistiksel manada kesinlik ifade eden “ortalama” değil), ve çok kötü durumdaki arabalar (limonlar). Her tür araba eşit frekansa sahiptir, yani*

P(“limon”) = P(“orta”) = P(“karpuz”) = 1/3

Satıcı ve bir alıcı her bir araba türü için, sırasıyla Y_S ve Y_B, kadar aşağıda verilen değerleri biçiyorlar:

Tür	Satıcı	Alıcı
Limon	5000	6000
Orta	6000	10000
Karpuz	10000	11000

İlk dikkat edilmesi gereken şey, her araba çeşidi için, alıcının biçtiği değerin satıcının biçtiğinden daha yüksek olduğudur, dolayısıyla her bir tür araba için, alış veriş alıcı ile satıcının biçtiği değerin arasındaki bir fiyattan gerçekleşmeli. Ancak, kullanılmış arabalar da, kalite ilk anda görülen şey değildir. Bu nedenle, eğer ne alıcı ne de satıcı söz konusu arabanın türü X'i bilmiyorsa, onların beklenen değeri yinelenen beklentiler kanununa göre aşağıdaki gibidir:

$$\begin{aligned}
\mathbb{E}[Y_S] &= \mathbb{E}[Y_S|\text{" limon "}]P(\text{" limon "}) + \mathbb{E}[Y_S|\text{" orta "}]P(\text{" orta "}) + \mathbb{E}[Y_S|\text{" karpuz "}]P(\text{" karpuz "}) \\
&= \frac{1}{3}(5,000 + 6,000 + 10,000) = 7,000 \\
\mathbb{E}[Y_S] &= \mathbb{E}[Y_B|\text{" limon "}]P(\text{" limon "}) + \mathbb{E}[Y_B|\text{" orta "}]P(\text{" orta "}) + \mathbb{E}[Y_B|\text{" karpuz "}]P(\text{" karpuz "}) \\
&= \frac{1}{3}(6,000 + 10,000 + 11,000) = 9,000
\end{aligned}$$

Dolayısıyla alış verişin gerçekleşmesi gerekir.

Daha gerçekçi bir düzenlemede, arabanın satıcısı arabanın kalitesini alıcıdan daha iyi bilir (tamirat geçmişini, kazalarını vs.) ve arabayı satmaya arzulayacağı fiyat belirler. Eğer satıcı üç araba türünü de mükemmel bir şekilde ayırt edebilirse, ki alıcı bunu yapamaz, alıcının satıcının belirtilen fiyattan arabayı satma arzusuna koşullanmış beklentiler oluşturması gerekir.

Eğer satıcı 6000 dolardan daha düşük bir fiyat belirtirse, alıcı kesin olarak arabanın "limon" olduğunu bilirdi, çünkü diğer durumlarda satıcı en az 6000 dolar talep ederdi, yani

$$\mathbb{E}[Y_B|Y_S < 6000] = \mathbb{E}[Y_B|\text{" limon "}] = 6000$$

ve alış veriş gerçekleşirdi. Ancak, eğer araba "karpuz" ise, satıcı en az 10000 dolar talep ederdi, hâlbuki alıcı en fazla

$$\mathbb{E}[Y_B|Y_S \leq 10,000] = \mathbb{E}[Y_B] = 9,000 < 10,000$$

kadar ödeyecekti, bu nedenle de satıcı yüksek kalite bir arabayı makul bir fiyata satamayacaktı.

Piyasanın “karpuz” için çalışmamasının(break down) nedeni bu modelde satıcının alıcıya arabanın kalitesinin düşük olmadığı konusunda kabul edilebilir bir garanti verememesidir, bu nedenle alıcı alış verişinde kötü bir araba alma ihtimalini hesaba katar.

Koşulu ve koşulsuz beklenen değerler arasındaki önemli bir ilişki Yinelenen Beklentiler Kanunudur (daha önce gördüğümüz Toplam Olasılık Kanunu’na çok benzer):

Önerme 1 (Yinelenen Beklentiler Kanunu).

$$\mathbb{E}\left[\mathbb{E}[Y|X]\right] = \mathbb{E}[Y]$$

İSPAT: g(x) = $\mathbb{E}[Y|X = x]$ olsun. g(x) x’in bir fonksiyonudur. Beklenen değeri şimdi hesaplayabiliriz:

$$\begin{aligned}
\mathbb{E}[g(X)] &= \int_{-\infty}^{\infty} g(x) f_X(x)dx = \int_{-\infty}^{\infty} \mathbb{E}[Y|X = x] f_X(x)dx \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} y \frac{f_{XY}(x,y)}{f_X(x)} dy \right) f_X(x)dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{XY}(x,y)dydx \\
&= \int_{-\infty}^{\infty} y \left(\int_{-\infty}^{\infty} f_{XY}(x,y)dx \right) dy \\
&= \int_{-\infty}^{\infty} y f_Y(y)dy = \mathbb{E}[Y]
\end{aligned}$$