# BÖLÜM 1

## BASİT REGRESYONUN İNCELENMESİ

# GİRİŞ

- Modelleme matematiksel ifadelerin gelişimini, bir anlamda ilgilenilen rassal bir değişkenin davranışlarının tanımlanmasını kapsar.

- Bu değişken dünya piyasasındaki buğday fiyatı, belirli bir tümör tipinin büyüme oranı veya metal telin çekme kuvveti olabilir.

- Her durumda, bu değişkene bağımlı değişken adı verilir ve Y ile gösterilir.

- Y'nin alt simgesi gözlemin yapılmış olduğu belirli birimi tanımlar; fiyatın kaydedildiği zaman, ölümün gerçekleştiği ülke, tümör büyümesi kaydedilen denek ve benzerleri gibi.

# DOĞRUSAL MODEL VE VARSAYIMLARI

En basit doğrusal model sadece bir bağımlı değişken içerir ve bağımsı değişkenin değeri artıyor ya da azalıyorken, bağımlı değişkenin gerçek ortalaması sabit bir oranda değişir. Bu yüzden, $\mathcal{E}(Y_i)$ ile gösterilen $Y_i$'nin gerçek ortalaması ile $X_i$ arasındaki fonksiyonel ilişki bir doğrunun eşitliğidir:

$$\mathcal{E}(Y_i) = \beta_0 + \beta_1 X_i. \tag{1.1}$$

$X = 0$ olduğu durumda $\mathcal{E}(Y_i)$ değeri olan $\beta_0$, sabit terim ve X'deki birim başına değişimin $\mathcal{E}(Y_i)$ deki değişim oranı olan $\beta_1$ ise doğrunun eğimidir.

Bağımlı değişken $Y_i$'nin gözlemlerinin, $\mathcal{E}(Y_i)$ kitle ortalamalı rassal kitlelerinden yığınlarından elde edilen rassal gözlemler olduğu varsayılır. $Y_i$ gözlemlerinin $\mathcal{E}(Y_i)$ kitle ortalamasından sapmalarını hesaba katmak için dahil edilen rassal hata $\epsilon_i$ ile istatistiksel model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i. \tag{1.2}$$

Alt simge $i$ belirli gözlem birimini gösterir, $i = 1, 2, \ldots, n$. $n$ gözlemli bağımsız değişken $X_i$'nin ölçme hatasız olduğu varsayılır. Böylece, $X$ 'in gözlemlenmiş değerlerinin bilinen sabitler seti olduğu varsayılır. $Y_i$ ve $X_i$ eşleştirilmiş gözlemlerdir, her ikisi de her gözlem biriminde ölçülür.

Rassal hataların $\epsilon_i$ sıfır ortalamaya sahip, ortak varyansı $\sigma^2$ olan ve birbirinden bağımsız hatalar olduğu varsayılır. Modeldeki tek rassal eleman $\epsilon_i$ olduğundan, bu varsayım $Y_i$'lerin de ortak varyansınının$\sigma^2$ olduğunu ve birbirinden bağımsız olduğunu ima eder. Anlamlılık testini yapmak için rassal hataların normal dağıldığı varsayılır ve bu $Y_i$'nin de normal dağıldığını ima eder. Rassal hata varsayımları çoğunlukla şu şekilde ifade edilir:

$$\epsilon_i \sim \ NID(0, \sigma^2), \tag{1.3}$$

burada NID "normal ve bağımsız dağılımı" temsil eder. Parantez içindeki nicelikler normal dağılımın sırasıyla ortalama ve varyansını göstermektedir.

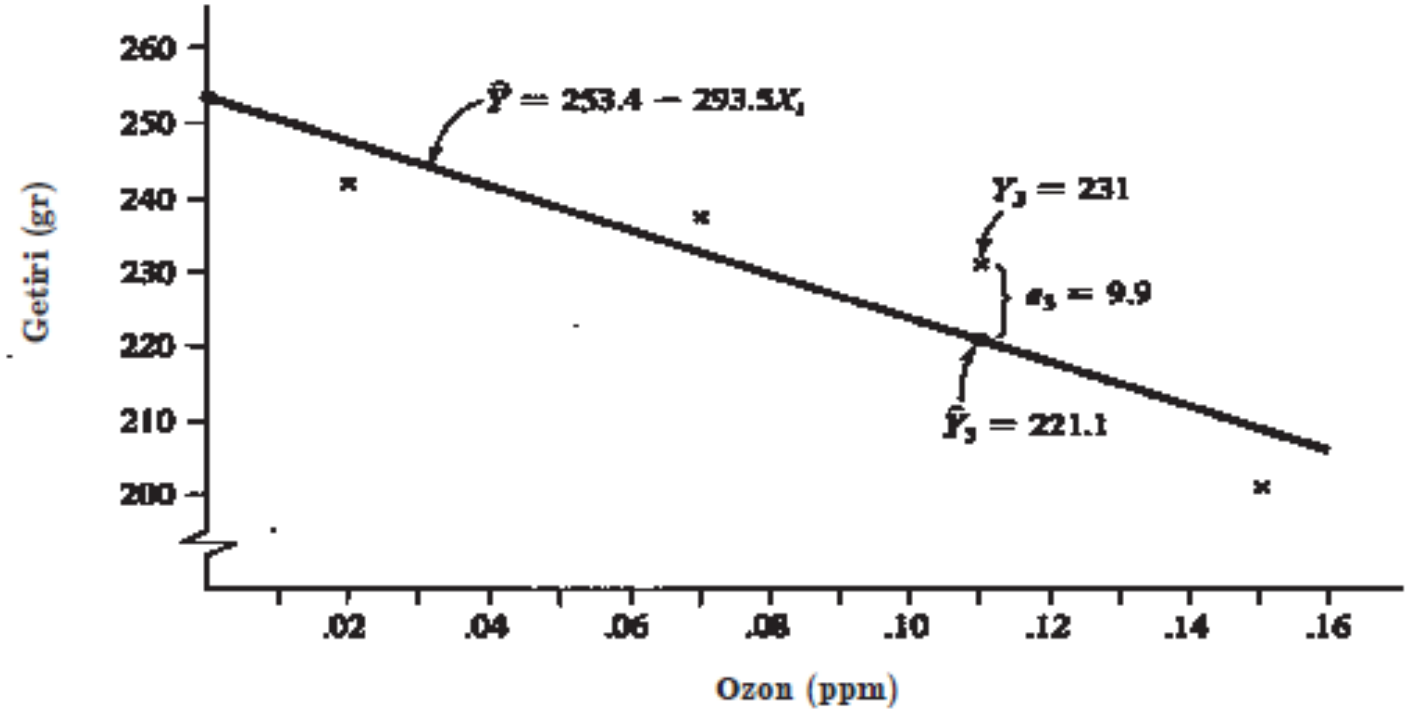
# EN KÜÇÜK KARELER TAHMİNİ

- Basit doğrusal modelin veriden tahmin edilen, ß0 ve ß1 olmak üzere iki parametresi vardır.
- Eğer ’ de rassal hata yoksa, her hangi iki veri ile parametre değerleri açıklıkla çözülebilir.
- Y’de ki rassal sapma gözlemlenen veri çiftlerinin farklı sonuçlar vermesine sebep olur. (Sadece gözlemlenen veri tam doğrunun üzerindeyse tüm tahminler eşit olabilir.)
- Bazı kriterler altında tüm bilgileri kullanarak tek “en iyi” sonucu verecek bir metoda ihtiyaç vardır.

# KESTİRİM DEĞERLERİ VE ARTIKLAR

- Örnek 1.1'deki regresyon eşitliği, bağımsız değişkenin seçilen düzeyi için **Y** bağımlı değişkenin ortalama tahminini elde etmek için kullanılabilir.

- Tabi ki bu gibi tahminlerin geçerliliği, modelin doğru kurulduğu varsayımına veya çalışmada gözlemlenen kirlilik dozlarının sınırları dahilin de en azından doğru modelin iyi bir kestirimi olmasına bağlıdır.




ŞEKİL 1.1. *Soya fasulyesi getirisinin ozon düzeyi üzerine regresyonu*

# BAĞIMLI DEĞİŞKENDEKİ DEĞİŞİMİN ANALİZİ

Eşitlik 1.13'teki gibi regresyon eşitliğinden elde edilen tahmin değerlerinin gözlem değerlerinden sapması artıklar olarak tanımlanmaktadır. Alternatif olarak, $Y_i$ bağımlı değişkenin her bir gözlem değeri, $X$ veri iken tahmin edilen $Y$'nin kitle ortalama değerleri ile bunlara karşılık gelen artıkların toplamı olarak yazılabilir:

$$Y_i = \widehat{Y}_i + e_i. \tag{1.14}$$

# TAHMİN KESİNLİĞİ

Rassal değişkenlerden hesaplanan herhangi niceliğin kendisi de bir rassal değişkendir. Böylelikle $\overline{Y}$, $\widehat{Y}$, $e$, $\widehat{\beta}_0$ ve $\widehat{\beta}_1$ $Y_i$'den hesaplanan rassal değişkenlerdir. Kesinliğin ölçüsü, varyans veya standart hata tahminleri, tahminlerin güvenirliliğine karar vermek için dayanak sağlar.

Hesaplanan regresyon katsayıları, $Y$ ve artıklar hepsi $Y_i$'nin doğrusal fonksiyonudur. Bunların varyansları, doğrusal bir fonksiyonun temel tanımı kullanılarak belirlenebilir. $U = \sum a_i Y_i$ rassal değişken $Y_i$'nin keyfi doğrusal bir fonksiyonu olsun, burada $a_i$'ler sabitlerdir. $U$'nun varyansının genel formülü

$$\mathrm{Var}(U) = \sum a_i^2 \mathrm{Var}(Y_i) + \sum \sum_{i \neq j} a_i a_j \mathrm{Cov}(Y_i, Y_j), \qquad (1.19)$$

burada çift toplam $i$ ve $j$'nin eşit olmadığı durumda, tüm $n(n-1)$ mümkün ikili terimler üzerindendir. $\mathrm{Cov}(\cdot,\cdot)$ parantez içinde verilen iki değişken arasındaki kovaryansı gösterir. (Kovaryans iki değişkenin birlikte azalış veya artış eğilimini ölçer.) Rassal değişkenler bağımsız olduğunda, klasik regresyon modelinde varsayıldığı gibi, tüm kovaryanslar sıfır olur ve çift toplam terimi yok olur. Bununla beraber, eğer yine klasik regresyon modelinde olduğu gibi Rassal değişkenlerin varyansları eşit ise $\mathrm{Var}(Y_i) = \sigma^2$ tüm $i$'ler için, doğrusal fonksiyonun varyansı şuna indirgenir

$$\mathrm{Var}(U) = (\sum a_i^2)\sigma^2. \tag{1.20}$$

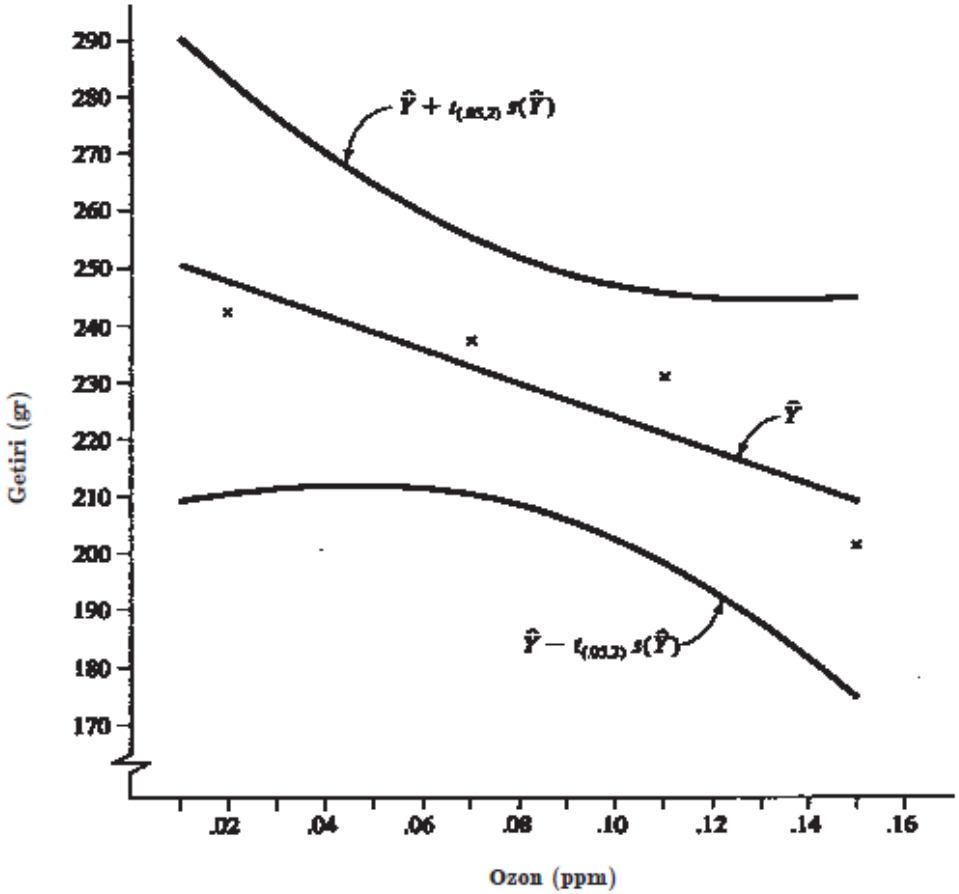
# ANLAMLILIK TESTLERİ VE GÜVEN ARALIKLARI

- Doğrusal regresyonda incelenen en yaygın hipotez doğrusal regresyon katsayısının gerçek değerinin, eğimin, sıfır olduğu hipotezidir.

- Bu şunu söylemektedir, bağımlı değişken , bağımsız değişkendeki değişmelerle ne doğrusal bir artış ne de azalış gösterir.

- Bazı durumlarda, problemin doğası gereği yokluk hipotezinde herhangi bir değere eşit olduğu iddia edilir.

- Rassal değişkenler olan hesaplanan regresyon katsayıları, hipotez doğru olsa dahi hiçbir zaman tam olarak hipotez değerine eşit olmayacaktır.




ŞEKİL 1.2. *Soya fasulyesi ortalama veriminin (tohum başına gram) ozon (ppm) üzerine regresyonundan elde edilen ortalama tepkinin bireysel güven aralık tahminleri.*

# ORİJİNDEN GEÇEN REGRESYON

Bazı durumlarda, regresyon çizgisinin orijinden geçmesi beklenmektedir. Bunun anlamı şudur; bağımsız değişken değeri sıfır olduğunda bağımlı değişkenin gerçek ortalamasının sıfır olması beklenir. Örneğin çoğu büyüme modelleri orijinden geçer. Katalizör olmadığında, katalizör ihtiyacı olan sistemin ürettiği kimyasal artık miktarı sıfır olacaktır. ß0 değeri sıfıra eşitlenerek doğrusal regresyon modeli orijinden geçirilir.
Doğrusal model şu hale gelir,

$$Y_i = \beta_1 X_i + \epsilon_i. \qquad (1.39)$$

TABLO 1.6. *Dokuz grup için toza maruz kalınmanın göreli riski. Toza maruz kalma parçacık /ft³/yıl olarak bildirilmiştir ve* $10^6$ *bölünerek ölçeklendirilmiştir.*

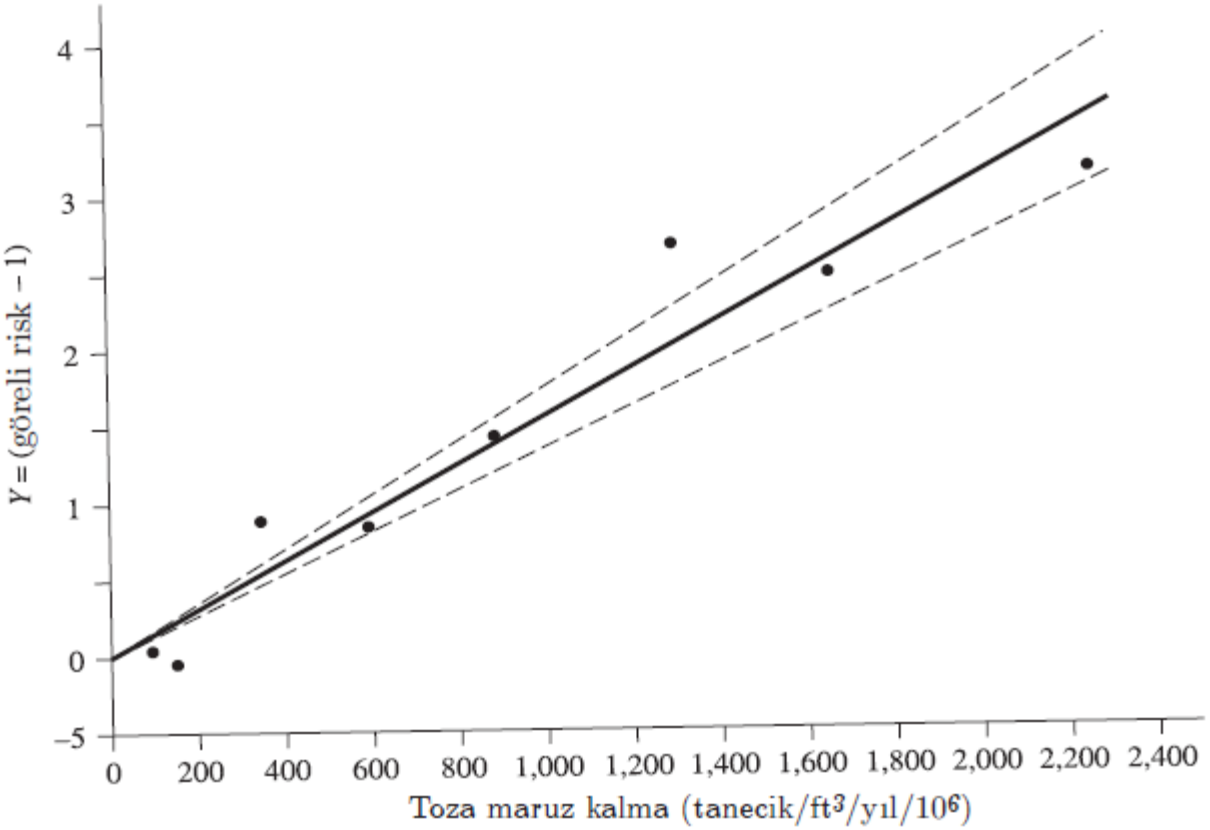
| $X =$ *Toza Maruz Kalma* | *Göreli Risk* | $Y =$ *Göreli risk* $- 1$ |
|---:|:---:|---:|
| 75 | 1.10 | .10 |
| 100 | 1.05 | .05 |
| 150 | .97 | $-.03$ |
| 350 | 1.90 | .90 |
| 600 | 1.83 | .83 |
| 900 | 2.45 | 1.45 |
| 1,300 | 3.70 | 2.70 |
| 1,650 | 3.52 | 2.52 |
| 2,250 | 4.16 | 3.16 |
| $\sum X_i = 7,375$ | | $\sum Y_i = 11.68$ |
| $\sum X_i^2 = 10,805,625$ | | $\sum Y_i^2 = 27.2408$ |
| | $\sum X_i Y_i = 16,904$ | |

TABLO 1.7. *Maruz kalma düzeyinin, artan göreli risk (Y = göreli risk − 1) üzerine orijinden geçen doğrusal regresyonundan elde edilen* $Y_i$, $\widehat{Y}_i$ *ve* $e_i$.

| $Y_i$ | $\widehat{Y}_i$ | $e_i$ |
|---:|---:|---:|
| .10 | .1173 | −.0173 |
| .05 | .1564 | −.1064 |
| −.03 | .2347 | −.2647 |
| .90 | .5475 | .3525 |
| .83 | .9386 | −.1086 |
| 1.45 | 1.4079 | .0421 |
| 2.70 | 2.0337 | .6663 |
| 2.52 | 2.5812 | −.0612 |
| 3.16 | 3.5198 | −.3598 |
| $\sum Y_i^2 = 27.2408$ | $\sum \widehat{Y}_i^2 = 26.4441$ | $\sum e_i^2 = .7967$ |

TABLO 1.8. *Toz parçacıklarına maruz kalma düzeyinin, artan göreli risk üzerine orijinden geçen regresyonun özet varyans analizi.*

| *Kaynak* | *s.d.* | *SS* | *MS* | $\mathcal{E}(MS)$ |
|---|---|---|---|---|
| $\text{Toplam}_{\text{düzeltilmemiş}}$ | $n$=9 | 27.2408 | | |
| Model kaynaklı | $p$=1 | 26.4441 | 26.4441 | $\sigma^2 + \beta_1^2(\sum X_i^2)$ |
| Artık | $n-p$=8 | .7967 | .0996 | $\sigma^2$ |




ŞEKİL 1.3. *Toz parçacıklarına maruz kalmanın göreli risk üzerine artan regresyonu ile orjinden geçirilen regresyon. Regresyon doğrusu üzerindeki bantlar, ortalamaların % 95 güven aralığı tahminlerinin limitleridir.*

# BİRKAÇ BAĞIMSIZ DEĞİŞKENLİ MODELLER

Çoğu model bağımlı değişkenin davranışını açıklamak için birden fazla bağımsız değişken kullanır. İlave olunan doğrusal model birkaç bağımsız değişken eklenerek genişletilebilir:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_p X_{ip} + \epsilon_i. \qquad (1.49)$$

Alt simge notasyonu her bir $X$ sayısını ve her bir bağımsız değişkeni tanımlayan regresyon katsayılarını eklemek için genişletilmiştir. $p$ tane bağımsız değişken ve $\beta_0$ dahil $p' = p+1$ tane tahmin edilen parametre vardır.

Bilinen en küçük kareler varsayımları uygulanır. $\epsilon_i$'nin bağımsız ve $\sigma^2$ ortak varyanslı olduğu varsayılır. Anlamlılık testlerinin kurulması veya güven aralıklarını oluşturulması için rassal hataların ayrıca normal dağılıma sahip olduğu varsayılır. Bağımsız değişkenlerin hatasız ölçüldüğü varsayılır.

Bu modele en küçük kareler yöntemi uygulanır ve

$$\begin{aligned} \text{SS(Res)} &= \sum (Y_i - \widehat{Y}_i)^2 \\ &= \sum (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{i1} - \widehat{\beta}_2 X_{i2} - \cdots - \widehat{\beta}_p X_{ip})^2 \end{aligned} \qquad (1.50)$$

ifadesi en küçüklenerek $p+1$ tane parametrenin tahmini bulunur.

# VARSAYIMLARIN İHLALİ

Kısım 1.1'de

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

rassal hata terimi $\epsilon_i$'ni, sıfır ortalamalı ve $\sigma^2$ sabit varyanslı normal dağılıma sahip bağımsız rassal değişken olduğu ve $X_i$'in ölçüm hatasız $n$ gözlemli bağımsız değişken olduğu varsayılmıştı. Bu varsayımlar altında en küçük kareler tahmin edicileri $\beta_0$ ve $\beta_1$ tüm mümkün yansız tahmin ediciler arasında en iyi (en küçük varyanslı) olanıdır. Bir önceki kısımda işlenen hipotez testi, kestirim ve güven aralıkları gibi istatistiksel çıkarımlar bu varsayımlar altında geçerlidir. Burada varsayımlardan ihlallerin, tahminler ve istatistiksel çıkarımlar üzerindeki etkilerini kısaca göstereceğiz. En küçük karelerdeki problemler ve mümkün düzeltme yollarının daha detaylı irdelenmesi Bölüm 10-14 arası gösterilecektir.