Doğrusal Regresyon

A. Talha Yalta
TOBB Ekonomi ve Teknoloji Üniversitesi

İKT-457 Ekonomi ve Finans İçin Yapay Zeka 1
Sürüm 0,93

Açık Lisans Bilgisi

A. Talha Yalta
TOBB Ekonomi ve Teknoloji Üniversitesi
2020 – 2021 

Ders Planı

1. Basit Doğrusal Regresyon
   - Katsayıların tahmini ve yorumu
   - Katsayıların ve modelin kesinliği
2. Çoklu Doğrusal Regresyon
   - Katsayıların tahmini ve yorumu
   - Katsayıların ve modelin kesinliği
   - Nitel değişkenler
   - Çoklu regresyonun uzantıları
3. Uygulamada Karşılaşılan Sorunlar
   - Modelleme sorunu
   - Hata teriminde korelasyon
   - Hata teriminde farklıserpilimsellik
   - Dışadüşenler
   - Çoklueşdoğrusallık
4. K-Enyakın Komşu Regresyonu

Doğrusal Regresyon Yöntemi

- Bu bölümde denetimli öğrenmeye temel olan doğrusal regresyon çözümlemesini ele alacağız.
- Bu yöntem, daha modern araçların yanında biraz sıkıcı görünse de kolay yorumlanabilirliği nedeniyle özellikle çıkarsamada en sık kullanılan yaklaşım olmayı sürdürmektedir.
- Ayrıca diğer birçok istatistiksel öğrenme yaklaşımını anlayabilmek için de iyi bir başlangıç noktası oluşturmaktadır.
- Kement, ridge, özyetinim, karar ağaçları gibi birçok modern araç regresyonu kullanır.
- Dolayısıyla, daha ileri yöntemlere geçmeden önce doğrusal regresyon konusunda sağlam bir altyapıya sahip olmak önemlidir.

Reklam Veri Seti

Bir önceki bölümde reklam veri setini kullanarak TV, radyo ve gazete reklamlarının satışlar üzerindeki etkisini incelemiştik.

Böyle bir çözümlemede regresyon kullanarak aşağıdakiler gibi birçok soruyu ele alabiliriz:

1. Reklam bütçesi ile satışlar arasında ilişki var mıdır?
2. İlişki varsa bu yararlı bir çözümleme yapacak kadar güçlü müdür?
3. Hangi medya aracılığıyla yapılan reklamlar satışlara katkı sağlar?
4. Her bir medyanın katkısını ne kadar kesinlikle tahmin edebiliriz?
5. Geleceğe yönelik ne kadar sağlıklı tahminler yapabiliriz?
6. Reklam ile satışlar arasındaki ilişkinin yapısı nedir?
7. Farklı reklam araçları birlikte daha etkili midir?

Regresyon yöntemini kullanarak bu yedi sorunun yanıtını bulabiliriz.

Ders Planı

Anakütle Regresyon Fonksiyonu

- İlk olarak, en yalın ve kolay durum olan basit doğrusal regresyon (simple linear regresyon) modelini ele alalım.
- Bu çözümlemede nicel bir Y değişkeninin tek bir X değişkenine verdiği doğrusal tepkiyi inceleriz.
- Önceki bölümde X ve Y arasındaki ilişkinin $Y = f(X) + \epsilon$ şeklinde olduğunu söylemiştik. Basit doğrusal regresyon için f şudur:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Yukarıdaki fonksiyona anakütle regresyon fonksiyonu (population regression function) denir.
- Burada β_0 ve β_1 bir doğruyu tanımlayan katsayılardır. β_0, doğrunun y-eksenini kestiği nokta, β_1 ise doğrunun eğimini gösterir.

Örneklem Regresyon Fonksiyonu

- Uygulamada β_0 ve β_1'in gerçek değerlerini bilemediğimiz için bunları tahmin ederiz:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\epsilon}$$

- Yukarıdaki fonksiyona da örneklem regresyon fonksiyonu (sample regression function) adı verilir.
- Burada ^ya da şapka (hat) işareti tahmin anlamına gelmektedir.
- Sonuç olarak, elimizde bulunan $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ şeklindeki eğitim verilerini kullanarak örneklem regresyon fonksiyonunu hesaplamak istiyoruz.
- $\hat{\beta}_0$ ve $\hat{\beta}_1$ değerlerini bulursak herhangi bir x_i gözlemi için $\hat{y}_i$ tahmini yapabiliriz:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Sıradan Enküçük Kareler (1)

- Örnek olarak, TV reklam harcamalarının satış üzerindeki etkisine bakmak istediğimizi düşünelim. Reklam veri setinde buna yönelik $n = 200$ adet gözlem bulunmaktadır.
- Bunları kullanarak $\hat{\beta}_0$ ve $\hat{\beta}_1$'yı öyle hesaplamalıyız ki bulduğumuz doğru elimizdeki 200 noktaya olabildiğince yakın olsun.
- Bir noktanın doğruya yakınlığını ölçmenin çeşitli yolları vardır.
- Ancak uygulamada açık ara en yaygın olan yöntem sıradan enküçük kareler (ordinary least squares) ya da kısaca SEK (OLS) yaklaşımıdır.

Sıradan Enküçük Kareler (2)

- SEK yöntemini anlamak için herhangi iki $\hat{\beta}_0$ ve $\hat{\beta}_1$ değeri alalım. Bunları kullanarak $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ noktalarını tahmin etmiş olalım.
- Burada her bir *i* gözlemi için yaptığımız hata şu olur:

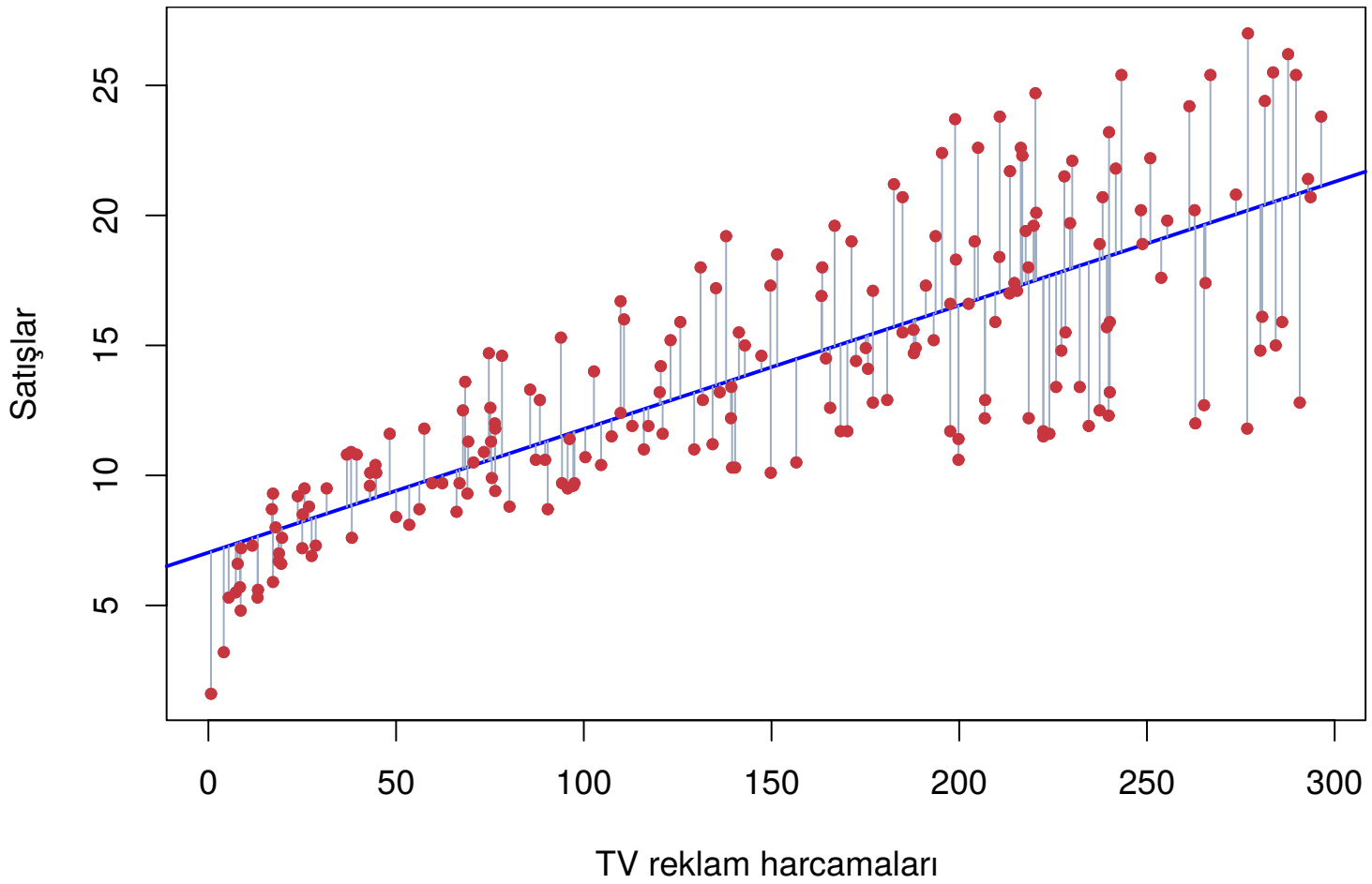
$$\epsilon_i = y_i - \hat{y}_i$$

- Yukarıdaki ϵ (epsilon) harfine hata (error) ya da kalıntı (residual) denir. Gözlenen *y* ile tahmin edilen $\hat{y}$ arasındaki farktır.
- Tüm gözlemlere ait ϵ'ları kullanarak kalıntı kareleri toplamı (residual sum of squares) ya da kısaca KKT (RSS) değerini buluruz:

$$\text{KKT} = \epsilon_1^2 + \epsilon_2^2 + \cdots + \epsilon_n^2$$

- İşte, SEK yöntemi elimizdeki verilere en iyi yakışan doğruyu bulmak için KKT değerini minimize eder.
- Bu minimizasyon işlemi doğrusal cebir ve kalkülüs kullanılarak yapılır ve bu dersin konusu dışındadır. $\hat{\beta}_0$ ve $\hat{\beta}_1$ tahminleri günümüzde bilgisayarlar tarafından kolayca hesaplanmaktadır.

Sıradan Enküçük Kareler (3)



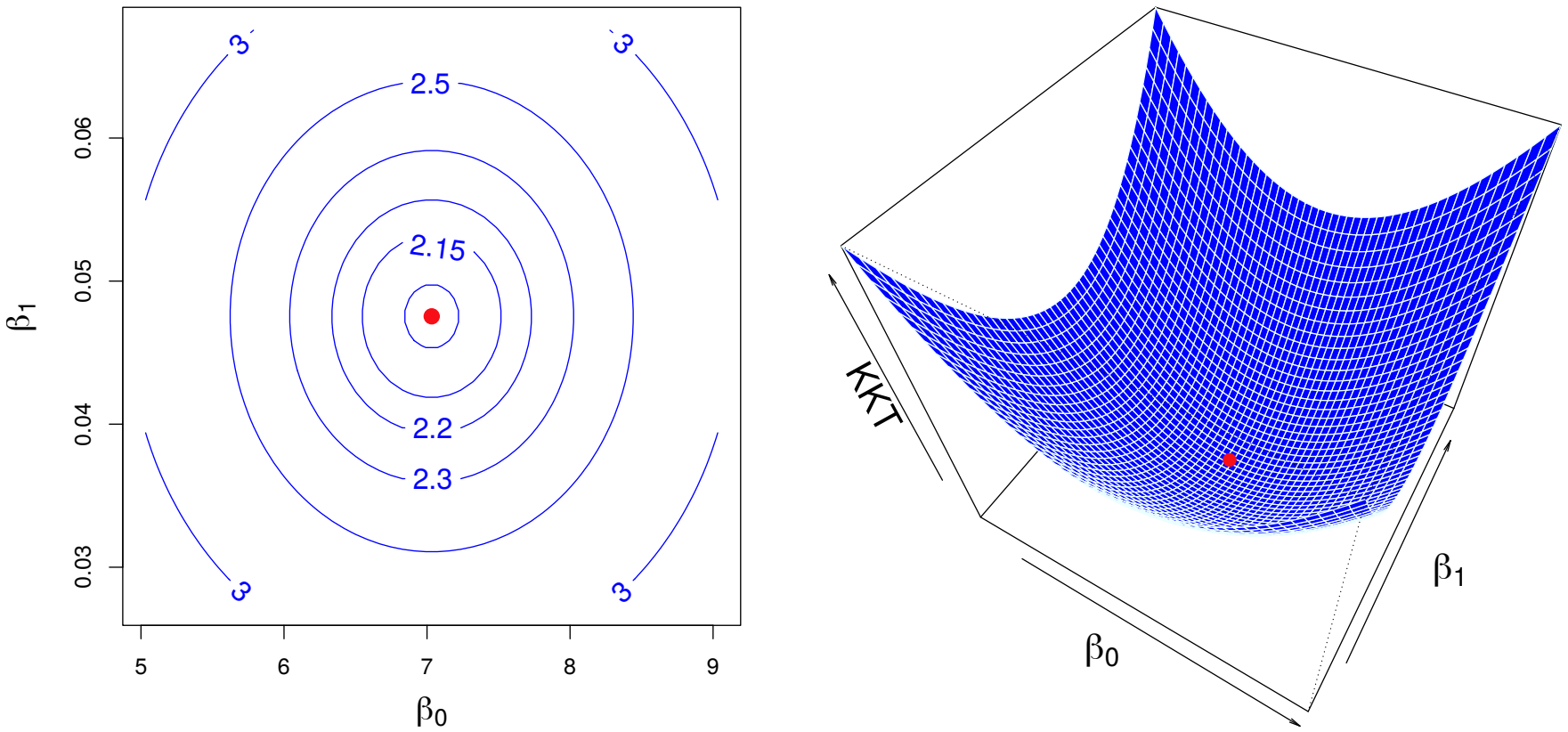
Şekil 1: Satışlar ile TV reklam harcamalarının ikili SEK regresyonu

Sıradan Enküçük Kareler (4)

- TV reklam harcamalarının satışlar üzerindeki etkisine yönelik basit doğrusal regresyon tahmini Şekil 1'de gösterilmiştir.
- Şekildeki kırmızı noktalar gözlemler, mavi çizgi ise tahmin edilen regresyon doğrusudur.
- SEK yöntemiyle hesaplanan doğrunun dikey kesme noktası $\hat{\beta}_0 = 7{,}03$ ve eğimi de $\hat{\beta}_1 = 0{,}0475$ olarak bulunmuştur.
- Bu sonuçları değişkenlerin birimine bakarak yorumlamalıyız. Veri setinde harcamalar 1000 dolar, satışlar ise 1000 adet şeklindedir.
- Dolayısıyla $\hat{\beta}_1$ katsayısını şöyle yorumlarız: TV reklamı için yapılan her 1000 dolarlık harcama satışları yaklaşık $0{,}0475 \times 1000 = 47{,}5$ adet artırmaktadır.
- $\hat{\beta}_0$'ı ise şöyle yorumlarız: Hiç reklam yapılmaması durumda yaklaşık $7{,}03 \times 1000 = 7030$ adet satış beklenmektedir.
- Farklı $\hat{\beta}_0$ ve $\hat{\beta}_1$ kombinasyonları için hesaplanan KKT değerleri Şekil 2'de verilmiştir.

Sıradan Enküçük Kareler (5)

- Şekilde hem sol hem de sağ paneli incelediğimizde $\hat{\beta}_0 = 7{,}03$ ve $\hat{\beta}_1 = 0{,}0475$ değerlerinin KKT'yi minimize ettiğini görülmektedir.



Şekil 2: KKT'yi belirleyen farklı $\hat{\beta}_0$ ve $\hat{\beta}_1$ tahmin değerleri

Ders Planı

Katsayıların Kesinliğinin Ölçülmesi (1)

- Yukarıda hesapladığımız $\hat{\beta}_0$ ve $\hat{\beta}_1$ değerlerini yorumlayabilmek için bunların ne kadar kesin olduğunu bilmek önemlidir.
- Bunlar birer tahmin olduğuna göre bunların gerçek değeri için belli bir aralık söyleyebilmek zorundayız.
- Bunun için kullandığımız standart yöntem eldeki eğitim verilerini kullanarak anakütleye ilişkin çıkarsama yapmaktır.
- Bunu basit bir örnekle açıklayalım: Bir zar atma deneyi düşünelim. Burada zar 1 ile 6 arasında her değeri alabilir. Öte yandan defalarca zar atarsak bunların ortalaması 3,5'e yakınsayacaktır.
- Yalnızca bir ya da iki zar atarak 3,5 değerini bulamayabiliriz. Ancak yeterince büyük bir örneklem alırsak, söz gelimi 30 kez zar atarsak 3,5'e çok yakın değerler elde ederiz.
- Elde edeceğimiz bu değerler yansız (unbiased) tahminlerdir. Diğer bir deyişle, 3,5'ten biraz farklı çıkabilirler ama gerçek değerden hep daha düşük ya da hep daha yüksek olmazlar.
- Ayrıca örneklem büyüdükçe de gerçek değere yakınsarlar.

Katsayıların Kesinliğinin Ölçülmesi (2)

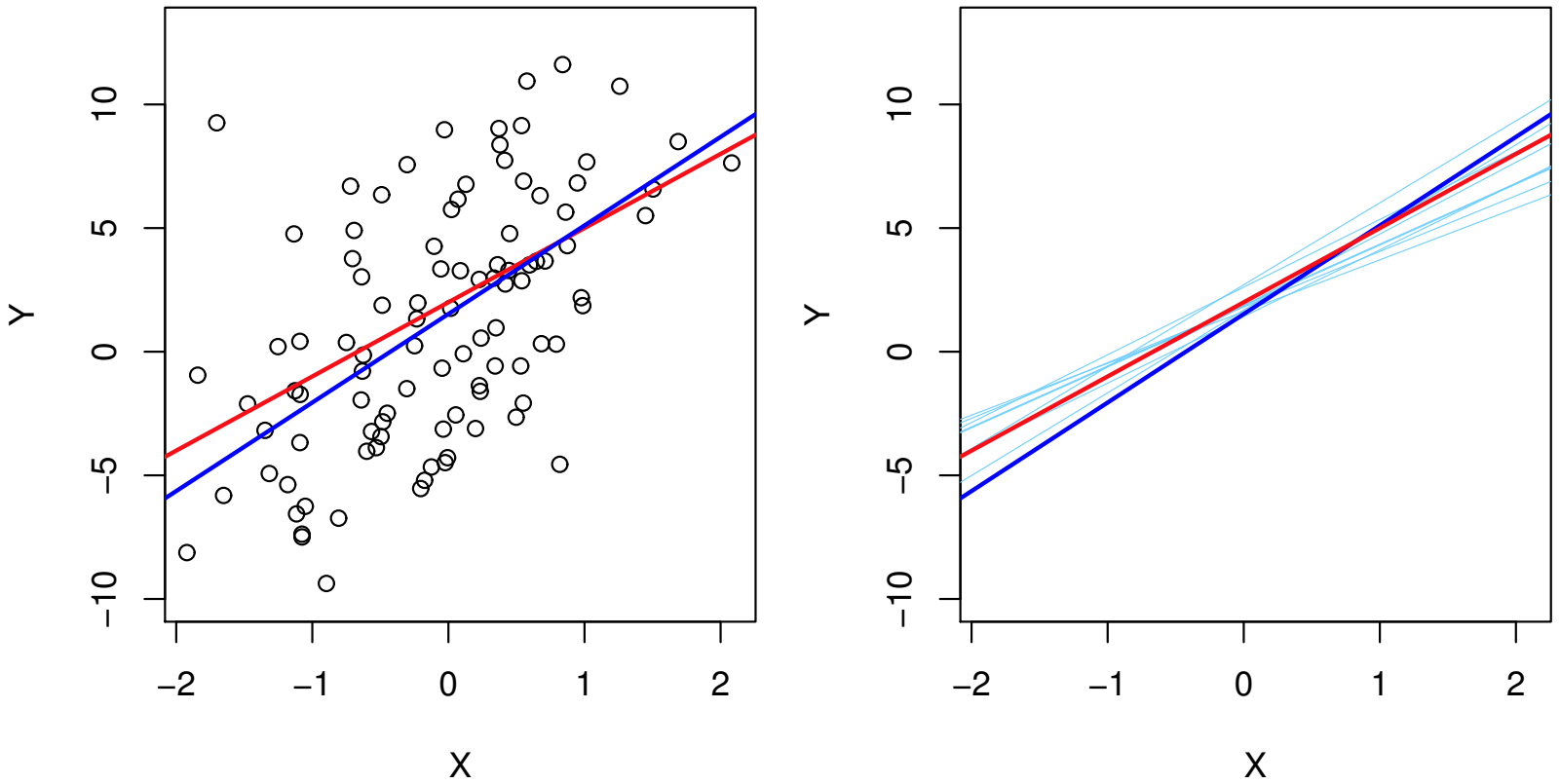
- Örneklemden gelen yansız tahminlerin anakütle değerine yakınsaması olgusu regresyon bağlamında da geçerlidir.
- Gerçekte f'yi bilemeyiz. Elimizde yalnızca bir tahmin olan örneklem regresyon fonksiyonu vardır.
- Ancak tahminimizin ne kadar kesin olduğunu hem analitik hem de deneysel olarak inceleyebiliriz.
- Deneysel bir örnek olarak aşağıdaki fonksiyonu ele alalım:

$$Y = 2 + 3X + \epsilon$$

- Yukarıdaki formülü kullanarak farklı rastsal X değerlerine karşılık Y'leri hesaplayabiliriz. Böylece, deneysel veri setleri üretebiliriz.
- Bu şekilde elde edilen örneklemlerden hesaplayacağımız yansız tahminlerde $\hat{\beta}_0 \approx 2$ ve $\hat{\beta}_1 \approx 3$ çıkar.
- Anakütle regresyon doğrusu ile 10 farklı simülasyon veri setinden elde edilen örneklem tahminleri Şekil 3'te verilmiştir.

Katsayıların Kesinliğinin Ölçülmesi (3)

- Sol panelde anakütle regresyon fonksiyonu kırmızı, örneklem regresyon fonksiyonu ise mavi renkle çizilmiştir. Sağdaki 10 farklı örneklem regresyon fonksiyonu da benzer tahminler üretmektedir.



Şekil 3: Anakütle regresyonu ile yansız örneklem regresyonları

Ölçünlü Hata

- Zar örneğine geri dönelim. Diyelim ki 30 defa zar attık ve ortalama 3,78 çıktı. Elimizdeki bu yansız tahminin kesinlik derecesi nedir?
- Bunu öğrenmek için bulduğumuz tahmine ait ölçünlü hata (standard error) ya da kısaca ÖH (SE) değerini hesaplarız:

$$\text{ÖH}(\hat{\mu})^2 = \text{var}(\hat{\mu}) = \frac{\sigma^2}{n}$$

- Yukarıda var, varyans ve $\hat{\mu}$ da tahmin edilen ortalama demektir.
- En sağdaki σ ise her bir gözlemin ortalama değerden ne kadar saptığını gösteren ölçünlü sapma (standard deviation) değeridir.
- Örneklem büyüklüğü (n) arttıkça ÖH'nin azaldığına dikkat ediniz.
- ÖH değeri bize ortalama bir tahminin gerçek değerden kaç birim saptığı bilgisini verir. Bu yaklaşımı kullanarak regresyon için $\text{ÖH}(\hat{\beta}_0)$ ve $\text{ÖH}(\hat{\beta}_1)$ değerlerini de kolayca bulabiliriz.
- Gerçekte anakütleye ait σ değeri bilinmez, ancak bu da tahmin edilebilir. Böylece, elimizde $\widehat{\text{ÖH}}(\hat{\beta}_0)$ ve $\widehat{\text{ÖH}}(\hat{\beta}_1)$ tahminleri olur.

Güven Aralıkları

- Ölçünlü hataları kullanarak bir katsayı tahminine ait güven aralığı (confidence interval) oluşturabiliriz.
- Örnek olarak, β_0 ve β_1 için %95 güven aralığı yaklaşık şöyledir:

$$\beta_0 \approx \hat{\beta}_0 \pm 2 \times \widehat{\text{ÖH}}(\hat{\beta}_0) \quad \text{ve} \quad \beta_1 \approx \hat{\beta}_1 \pm 2 \times \widehat{\text{ÖH}}(\hat{\beta}_1)$$

- TV reklamı örneğimize dönelim. Bu regresyonda $\widehat{\text{ÖH}}(\hat{\beta}_0)$=0,4578 ve $\widehat{\text{ÖH}}(\hat{\beta}_1)$=0,0027 bulunmuştur.
- Buna göre %95 güven aralıkları aşağıdaki gibi hesaplanır:

$$[6{,}130 \leq \beta_0 \leq 7{,}935] \quad \text{ve} \quad [0{,}042 \leq \beta_1 \leq 0{,}053]$$

- Bu güven aralıklarının yorumu şöyledir: Eğer farklı test veri setleri oluşturur ve regressionu tekrar tekrar hesaplayacak olursak bulacağımız 100 tahminden 95'inin bu aralıkta olmasını bekleriz.
- Dolayısıyla gerçek değeri de %95 olasılıkla bu aralıkta bekliyoruz.
- Uygulamada en çok %95 güven aralıkları kullanılır ancak %90 ve %99 aralıkları da yaygındır. Bunların yorumu da benzerdir.

Önsav Sınamaları (1)

- Ölçünlü hataları kullanarak herhangi bir β katsayısı üzerinde önsav sınamaları da yapabiliriz.
- Bunun için öncelikle bir H_0 sıfır önsavı (null hypothesis) ile H_1 almaşık önsav (alternative hypothesis) belirtiriz. Örnek olarak:

$$H_0 : \beta = \beta^* \quad \text{ve} \quad H_1 : \beta \neq \beta^*$$

- Bu sınamada amacımız β için yaptığımız $\hat{\beta}$ tahmininin β^*'dan anlamlı (significant) derecede uzak olup olmadığı bulmaktır. Diğer bir deyişle, $\beta = \beta^*$ olmadığını güvenle söyleyebilir miyiz?
- Bunun için aşağıdaki test istatistiğini hesaplarız:

$$t = \frac{\hat{\beta} - \beta^*}{\widehat{\text{ÖH}}(\hat{\beta})}$$

- H_0'ın geçerli olması durumunda yukarıdaki test istatistiği $n - 2$ serbestlik derecesi (degree of freedom) ile t dağılımına uyar.
- Bulunan değerin ilgili dağılımdan gelme olasılığını bilgisayar ile hesaplayabilir ve böylece, H_0'ı ret edebilir ya da etmeyebiliriz.

Önsav Sınamaları (2)

- Örnek olarak, TV reklamı örneğimizdeki $\hat{\beta}_1$'nın sıfırdan anlamlı derecede uzak olup olmadığını sınamak istediğimizi düşünelim:
$$H_0 : \beta_1 = 0 \quad \text{ve} \quad H_1 : \beta_1 \neq 0$$
- $\hat{\beta}_1 = 0{,}0475$ ve $\widehat{\text{ÖH}}(\hat{\beta}_1) = 0{,}0027$ bulduğumuzu daha önce söylemiştik. Bu durumda test istatistiği şudur:
$$t = \frac{0{,}0475 - 0}{0{,}0027} = 17{,}59$$
- Bilgisayar bize 17,59 değerinin ilgili t dağılımından gelme olasılığının onbinde birden küçük olduğunu söyleyecektir.
- Dolayısıyla anakütledeki gerçek β_1 değerinin 0 olmadığı konusunda yüksek bir kesinlikle çıkarımda bulunabiliriz.
- Burada yaptığımız şey tek bir β parametresine ilişkin t sınamasıdır. Bunun dışında birçok farklı önsav sınaması vardır. Bunları daha sonra yeri geldikçe tartışacağız.

İkili Regresyon Bilgisayar Çıktısı

- Yukarıda gördüğümüz regresyon katsayı tahminleri ve önsav sınamaları bilgisayarlar tarafından kolayca hesaplanmaktadır.
- TV reklamı regresyonumuza ait bilgisayar çıktısı şöyledir:

Değişken	**Katsayı**	**Ölçünlü hata**	**t-istatistiği**	**p-değeri**
Sabit terim	7,0325	0,4578	15,36	< 0,0001
TV reklamı	0,0475	0,0027	17,59	< 0,0001

- Çizelgede sağdaki ilk iki sütunda $\hat{\beta}_0$ ve $\hat{\beta}_1$ katsayıları ile bunların ölçünlü hataları görülmektedir.
- Son iki sütunda ise az önce β_1 için yaptığımız $H_0 : \beta = 0$ ve $H_1 : \beta \neq 0$ şeklindeki t-sınaması sonuçları ile bunların p-değerleri verilmiştir. TV reklamı için 17.59 değerini biz de hesaplamıştık.
- Tipik bir regresyon çıktısında bu sınama otomatik yapılarak her bir katsayının sıfırdan anlamlı derecede uzak olup olmadığı ölçülür.
- Kesinlik değerlendirmesi için genellikle yüzde 95 güven düzeyi kullanılır. Bu düzeyde anlamlılık kararı verebilmek için p-değerinin 0,05'ten küçük olmasına bakılır.

Modelin Kesinliğinin Ölçülmesi

- Model katsayılarının nasıl yorumladığını ve kesinliklerinin nasıl ölçüldüğünü yukarıda gördük.
- Katsayıları değerlendirdikten sonra tahmin ettiğimiz modelin genel olarak verilere ne derece yakıştığını da bilmek isteriz.
- Bu amaçla kullandığımız temel ölçütler kalıntı ölçünlü hatası (residual standard error), R^2 istatistiği ve F-istatistiğidir.
- TV reklamları örneğimiz için bu değerler Çizelge 1'deki gibidir:

İstatistik	**Değer**
Kalıntı ölçünlü hatası	3,260
R^2	0,612
F-istatistiği	312,100

Çizelge 1: Satışlar ve TV reklamları regresyonuna ait özet istatistikler

- Şimdi yukarıdakilerden ilk ikisine bakacağız. F-istatistiği'ni ise çoklu regresyon bölümünde tartışacağız.

Kalıntı Ölçünlü Hatası

- Kalıntı ölçünlü hatası (residual standard error) ya da kısaca KÖH (RSE), y_i tepki değerlerinin regresyon doğrusundan ortalama kaç birim saptığını ölçer.
- KÖH'ü bulmak için hata terimi ϵ'un ölçünlü sapmasını hesaplarız:

$$\text{KÖH} = \sqrt{\frac{1}{n-2}\text{KKT}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- Örnek olarak, TV reklamı örneğimizde her bir gözlemin regresyon doğrusundan ortalama 3,260 adet saptığını görüyoruz.
- Bunun kabul edilebilir bir değer olup olmadığı duruma göre değişir. Bu veri setinde ortalama satış 14,000 adet olduğu için modelimizdeki gözlemler $3{,}260/14{,}000 = \%23$ sapma göstermiştir.
- Görüldüğü gibi, KÖH aslında yakışmanın iyiliğini değil, yakışma eksikliğini göstermektedir ve birim ölçeğindedir.

Belirleme Katsayısı

- Yakışmanın iyiliğine yönelik ikinci ölçüt belirleme katsayısı (coefficient of determination) ya da kısaca R^2 istatistiğidir.
- R^2'yi yorumlamak daha kolaydır çünkü yakışmayı 0 ve 1 aralığında bir oran olarak ölçer:

$$R^2 = \frac{\text{TKT} - \text{KKT}}{\text{TKT}} = 1 - \frac{\text{KKT}}{\text{TKT}}$$

- Burada TKT (TSS), toplam kareleri toplamı (total sum of squares) anlamındadır. Tepki değişkeni Y'deki ortalama değişkenliği verir:

$$\text{TKT} = \sum (y_i - \bar{y})^2$$

- KKT, kalıntılardan kaynaklanan ve regresyon tarafından açıklanamayan değişkenliktir. Dolayısıyla TKT−KKT de regresyonun açıkladığı değişkenliği anlatır. İşte, bunun toplama oranı da R^2 olur.
- Örneğimizde $R^2 = 0{,}61$ çıkmıştır. Bunu şöyle yorumlarız: Regresyon doğrusu Y'deki değişikliği yüzde 61 oranında açıklamaktadır. Geriye kalan yüzde 39 ise diğer etmenlerden kaynaklıdır.

Ders Planı

Çoklu Doğrusal Regresyon (1)

- Veri çözümlemesi uygulamalarında genellikle elimizde birden fazla açıklayıcı değişken olur.
- Örnek olarak, reklam veri setinde TV harcamalarının yanı sıra radyo ve gazete reklam harcamaları da bulunmaktadır.
- Bu değişkenleri de dikkate almak istediğimiz zaman tek tek ikili regresyonlar yapabiliriz. Ancak bu doğru bir yaklaşım değildir.
- Üç ayrı regresyonla tek bir kestirim elde edilemez. Ayrıca ikili regresyonlar diğer değişkenlerin etkisini dikkate almayarak eksik ve yanlı sonuçlar üretirler.
- TV, gazete ve radyo reklam harcamalarını ayrı ayrı inceleyen ikili regresyonlar Çizelge 2'de görülmektedir.

Çoklu Doğrusal Regresyon (2)

Değişken	Katsayı	Ölçünlü hata	t-istatistiği	p-değeri
Sabit terim	7,033	0,458	15,36	< 0,0001
TV reklamı	0,048	0,003	17,59	< 0,0001

Değişken	Katsayı	Ölçünlü hata	t-istatistiği	p-değeri
Sabit terim	9,312	0,563	16,54	< 0,0001
Radyo reklamı	0,203	0,020	9,92	< 0,0001

Değişken	Katsayı	Ölçünlü hata	t-istatistiği	p-değeri
Sabit terim	12,351	0,621	19,88	< 0,0001
Gazete reklamı	0,055	0,017	3,30	0,0012

Çizelge 2: Satışlar ile TV, radyo ve gazete reklamlarının ikili regresyonları

Çoklu Doğrusal Regresyon (3)

- Elimizde birden fazla X değişkeni olduğu zaman aşağıdaki gibi bir çoklu doğrusal regresyon (multiple linear regression) belirtiriz:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- Burada X_j'ler farklı açıklayıcı değişkenlerdir. Önlerindeki β_j katsayıları ise diğer tüm değişkenler sabitken ilgili X_j 1 birim arttığı zaman Y'nin kaç birim değiştiğini gösterir.
- Yukarıdaki anakütle regesyonunu eğitim verileri ile tahmin etmek için yine, örneklem regresyon fonksiyonunu kullanırız:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p + \hat{\epsilon}$$

- Çoklu regresyonda da $\hat{\beta}$ tahminleri için SEK yöntemi kullanılır. Buradaki karmaşık işlemler bilgisayarlar tarafından kolayca yapılmaktadır.

Çoklu Regresyon Bilgisayar Çıktısı

- Reklam örneğimize dönelim. Satışlar ile TV, radyo ve gazete reklam harcamalarına ilişkin çoklu regresyon modeli şöyledir:

$$\text{Satışlar} = \beta_0 + \beta_1\ \text{TV} + \beta_2\ \text{Radyo} + \beta_3\ \text{Gazete} + \epsilon$$

- Model tahminine ilişkin bilgisayar çıktısı Çizelge 3'te verilmiştir.

Değişken	**Katsayı**	**Ölçünlü hata**	**t-istatistiği**	**p-değeri**
Sabit terim	2,939	0,3119	9,42	< 0,0001
TV reklamı	0,046	0,0014	32,81	< 0,0001
Radyo reklamı	0,189	0,0086	21,89	< 0,0001
Gazete reklamı	−0,001	0,0059	−0,18	0,8599

Çizelge 3: Satışlar ile TV, radyo ve gazete reklamları çoklu regresyonu

Çoklu Regresyon Katsayılarının Yorumlanması

- Çizelgedeki katsayıları değişkenlerin birimine göre yorumlarız. Bu veri setinde harcamalar 1000 dolar, satışlar 1000 adet şeklindedir.
- Dolayısıyla TV reklamına ait $\hat{\beta}_1$=0,046 katsayısının yorumu şudur: Radyo ve gazete reklamı sabitken, TV reklamındaki 1 birim (1000 dolar) artış sonucunda satışlar 0,046 ($\times$1000) adet artmaktadır.
- Diğer bir deyişle TV reklamlarındaki her 1000 dolarlık harcama satışları yaklaşık 46 adet artırmaktadır.
- Radyo ve gazete katsayılarının yorumu da benzer şekildedir.
- Sabit terim $\hat{\beta}_0 = 2{,}939$ katsayısının yorumu ise şöyledir: Eğer TV, radyo ve gazete reklam harcamalarının hepsi birden sıfır olursa yaklaşık $2{,}939 \times 1000 = 2939$ adet satış beklenmektedir.
- Çoklu regresyondaki $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ katsayıları ikili regresyondakilere benzerdir. Öte yandan sabit terimin farklı olduğuna dikkat ediniz. Ayrıca burada gazete katsayısının p-değeri 0,8599'a yükselmiştir.
- Sonuç olarak, ikili ve çoklu regresyon birbirinden oldukça farklı sonuçlar verebilmektedir.

Karıştırıcı Değişken Etkisi

- Çoklu regresyonda gazete reklamı p-değerinin 0,8599 olduğuna dikkat ediniz.
- Bu, $H_0 : \beta_3 = 0$ önsav sınamasının sonuç istatistiğidir. Dolayısıyla, β_3'ün sıfırdan anlamlı derecede uzak olmadığını gösterir.
- Bunun nedeni ise bu örnekte radyo ve gazete reklamlarının yüksek korelasyona sahip olmasıdır. İkili regresyon bu ilişkiyi dikkate almadığı için daha önce gazete reklamları anlamlı çıkmıştı.
- Buna benzer durumlar uygulamada sıkça karşımıza çıkar.
- Tipik bir örnek olarak, yazın kumsalda dondurma satışları ile köpekbalığı saldırıları arasında güçlü ve anlamlı bir ilişki bulabiliriz.
- Bu hatalı sonucun nedeni hava sıcaklığının dikkate alınmamış olmasıdır. Burada hava sıcaklığına karıştırıcı değişken (confounding variable) denir. Bunu dikkate alınca sonuçlar düzelir.
- Karıştırıcı değişkenler modellemede son derece önemlidir. Yapay zeka bu konuda zorlanır.

Ders Planı

Çoklu Regresyonda Çıkarsama

- Bir regresyon modelini tahmin edip katsayıları yorumlamak kolay iştir. Asıl önemli olan, elde yorumlamaya değer bir sonuç olup olmadığını bilebilmektir.
- Bu doğrultuda aşağıdaki dört temel soruya yanıt ararız:
  1. Veriler modele ne kadar iyi yakışmıştır?
  2. Model bir bütün olarak olarak anlamlı mıdır?
  3. Y'yi açıklayan önemli X değişkenleri neleridir?
  4. Elde ettiğimiz kestirimler ne kadar güvenilirdir?
- Gelin, şimdi de bu konulara kısaca değinelim.

Çoklu Regresyonda Yakışmanın İyiliği

- En temel yakışmanın iyiliği ölçütleri olan R^2 ve kalıntı ölçünlü hatasından başta söz etmiştik.
- Bunların hesaplanışı ve yorumu ikili regresyondaki gibidir. Çoklu regresyon örneğimiz için aşağıdaki çizelgeye bakalım:

İstatistik	**Değer**
Kalıntı ölçünlü hatası	1,690
R^2	0,897
F-istatistiği	570,000

Çizelge 4: Satış ve reklamlar çoklu regresyonu, özet istatistikler

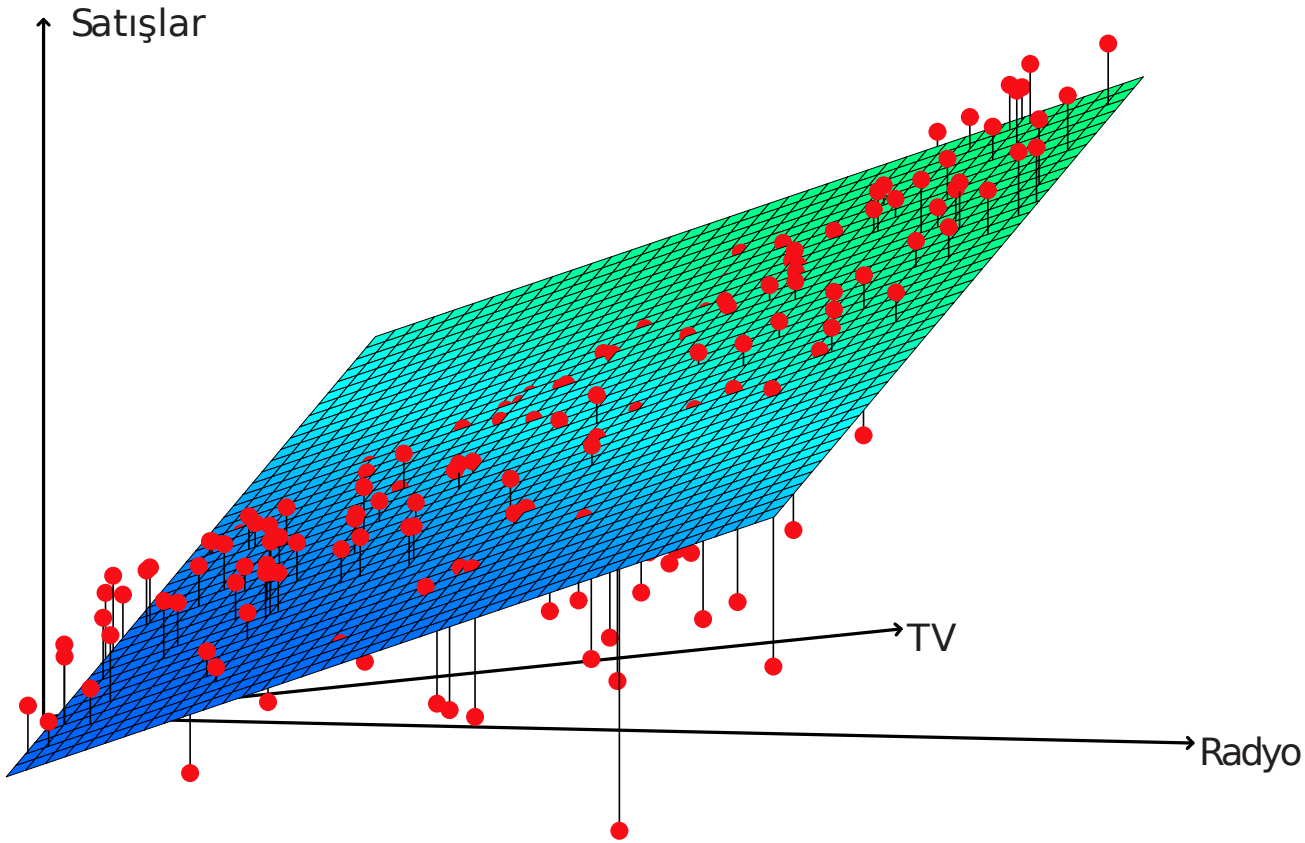
- İlk olarak, KÖH değerinin 1,690 çıktığını görüyoruz. Buna göre modelin tahmin ettiği satışlar ortalama olarak, gözlenen satışlardan 1,690 birim (1690 adet) sapmaktadır.
- TV ikili regresyonunda bu 3,260 idi. Burada yakışma iyileşmiştir.
- Ancak katsayısı anlamlı çıkmayan gazeteyi modelden atarsak KÖH, 1,681 olmaktadır. Buna göre gazetenin yakışmaya katkısı yoktur.

Ayarlamalı R-Kare

- R^2'ye de bakalım. Çoklu regresyonda $R^2 = 0{,}897$ bulunmuştur.
- Buna göre TV, radyo ve gazete reklam harcamalarını içeren model satışlardaki değişimi yüzde 89,7 düzeyinde açıklamaktadır.
- Baştaki ikili regresyonunda $R^2 = 0{,}612$ idi. Dolayısıyla, daha fazla değişkeni dikkate alan çoklu regresyonda yakışma artmıştır.
- Öte yandan R^2, modeldeki değişken sayısına karşı hassastır. Yeni değişken eklendikçe bunların açıklama gücü yoksa bile R^2 artar.
- Bu nedenle, tek açıklayıcı değişkeni olan baştaki model ile üç açıklayıcı değişkeni olan yeni modeldeki R^2'leri karşılaştıramayız.
- Böyle durumlarda karşılaştırılabilir olan istatistik ayarlamalı R-kare (adjusted R-squared) ya da kısaca $\bar{R}^2$ değeridir.
- Ayarlamalı R-kare modeldeki değişken sayısını dikkate aldığı için normal R-kareden düşük çıkar.
- Örnek olarak, çoklu regresyon örneğimizde $\bar{R}^2 = 0{,}88$'dir. Bunu 0,612 ile karşılaştırınca yakışmanın iyiliğinin arttığı görülüyor.

İki Değişkenli Kalıntı Çizimi

- Yakışmanın iyiliğini şekil üzerinde incelemek de yararlıdır. Burada, yalnızca TV ya da radyo reklamı yapıldığı zaman modelin satışları yüksek tahmin ettiği görülüyor. Doğrusal-dışı bir ilişki söz konusu.



Şekil 4: Çoklu regresyonda iki değişkenli kalıntı çizimi

Çoklu Regresyonda Bütünün Anlamlılığı

- Regresyonda eğer yakışma düzeyi düşükse bütünün anlamlılığı (overall significance) durumuna özellikle bakmak isteriz.
- Bu, aşağıdaki önsav sınamasını yapmak demektir:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \quad \text{ve} \quad H_1 : \text{En az bir } \beta_j \neq 0$$

- Görüldüğü gibi burada tüm katsayıların aynı anda sıfır olup olamayacağı sorgulanmaktadır. Bunun için şu F-istatistiği hesaplanır:

$$F = \frac{(\text{TKT} - \text{KKT})/p}{\text{KKT}/(n-p-1)}$$

- Eğer regresyon kalıntıları normal dağılımlıysa ve H_0 doğru ise yukarıdaki sınama istatistiği F dağılımına uyar. Bilgisayar bunu ve buna ait p-değerini kolayca hesaplar ve çıktı olarak verir.
- Bu aslında $H_0 : R^2 = 0$ sınamasıdır. Yakışmanın yokluğunu ölçer.
- Çizelge 4'e dönersek örneğimizde $F = 570$ olduğu görülüyor. İlgili F dağılımında bu değeri bulma p-değeri $< 0{,}0001$'tir. Dolayısıyla modelin bütün olarak anlamlı olmadığı sıfır önsavını reddederiz.

Genel F Sınaması

- Yukarıdaki standart F sınaması dışında isteğe göre kendi özel F sınamalarımızı da tasarlayabiliriz. Örnek olarak, şunu sınayalım:

$$H_0 : \beta_1 = 7, \beta_2 = \beta_3 \quad \text{ve} \quad H_1 : H_0 \text{ geçerli değil.}$$

- Bu sıfır önsavının geçerli olması durumunda baştaki reklam harcamaları modelimiz değişir ve aşağıdaki gibi olur:

$$Y - 7X_1 = \beta_0 + \beta_2(X_2 + X_3) + \epsilon$$

- Yukarıdaki β_0, β_2 ve ϵ değerleri artık ilk modeldekilerden farklıdır.
- Sınırladığımız modeli kullanarak şu F istatistiğini hesaplarız:

$$F = \frac{(\text{KKT}_s - \text{KKT}_{sz})/m}{\text{KKT}_{sz}/(n-p)}$$

- Burada s harfi "sınırlamalı", sz ise "sınırlamasız" demektir. Ayrıca m harfi sınırlama sayısıdır ve H_0'daki $=$ işareti sayısı ile aynıdır.
- Görüldüğü gibi, genel F sınaması için baştaki (sınırlamasız) model ile H_0 uygulanmış (sınırlamalı) modeli tahmin edip her ikisinin KKT değerlerini kullanırız. Tüm bunlar yine bilgisayarda yapılır.

Önemli Açıklayıcı Değişkenler

- Çeşitli t ve F sınamalarına bakınca bazı X'lerin anlamlı olmadığını bulabiliyoruz. Bu durumda doğal olarak, Y'yi açıklamada önemli olan değişkenlere karar vermek isteriz.
- Bunun için çok fazla sayıda modeli tek tek denemek gerekir.
- Ancak bunu yapmak zordur. Değişken sayısı p olan bir modelde 2^p adet farklı alt-model kombinasyonu söz konusudur.
- Seçim işini hızlı ve otomatik yapmak için üç klasik yaklaşım vardır:
- İleri seçim (forward selection): Yalnızca sabit terim içeren en basit modelle başlanır ve KÖH değerini en çok düşüren değişkenler sırayla eklenir. KÖH'ün fazla düşmediği belli bir noktada durulur.
- Geri seçim (backward selection): Başta tüm değişkenler modele eklenir ve p-değeri en yüksek olan değişkenler sırayla çıkartılır.
- Karma seçim (mixed selection): Değişkenler modele tek tek eklenir. Ancak işlem sırasında önceki bir değişkenin p-değeri belli bir eşikten fazla yükselirse bu değişken çıkartılır.
- Değişken seçimi konusunu 6. Bölümde ayrıntılı işleyeceğiz.

Kestirimlerin Güvenilirliği

- Bir model tahmin ederken önemli bir amacımız çeşitli $X_1, X_2, \ldots X_p$ değerlerine karşılık gelen Y değerini kestirmektir.
- Ancak bu kestirimle ilgili 3 farklı belirsizlik söz konusudur:

  1. $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ katsayı tahminlerindeki belirsizlik. Bunlar 2. Bölümde söz ettiğimiz azaltılabilir hatalar ile ilgilidir. Bu belirsizlik nedeniyle katsayı güven aralıkları hesaplarız.
  2. $f(X)$ fonksiyonundaki belirsizlik. Bu, model yanlılığı dediğimiz azaltılabilir hata ile ilgilidir. Burada şimdilik bunu yok sayalım.
  3. Hata terimi ϵ'dan kaynaklı azaltılamayan hata. Bununla ilgili olarak Y ile $\hat{Y}$'nın farkına yönelik kestirim aralıkları hesaplarız.

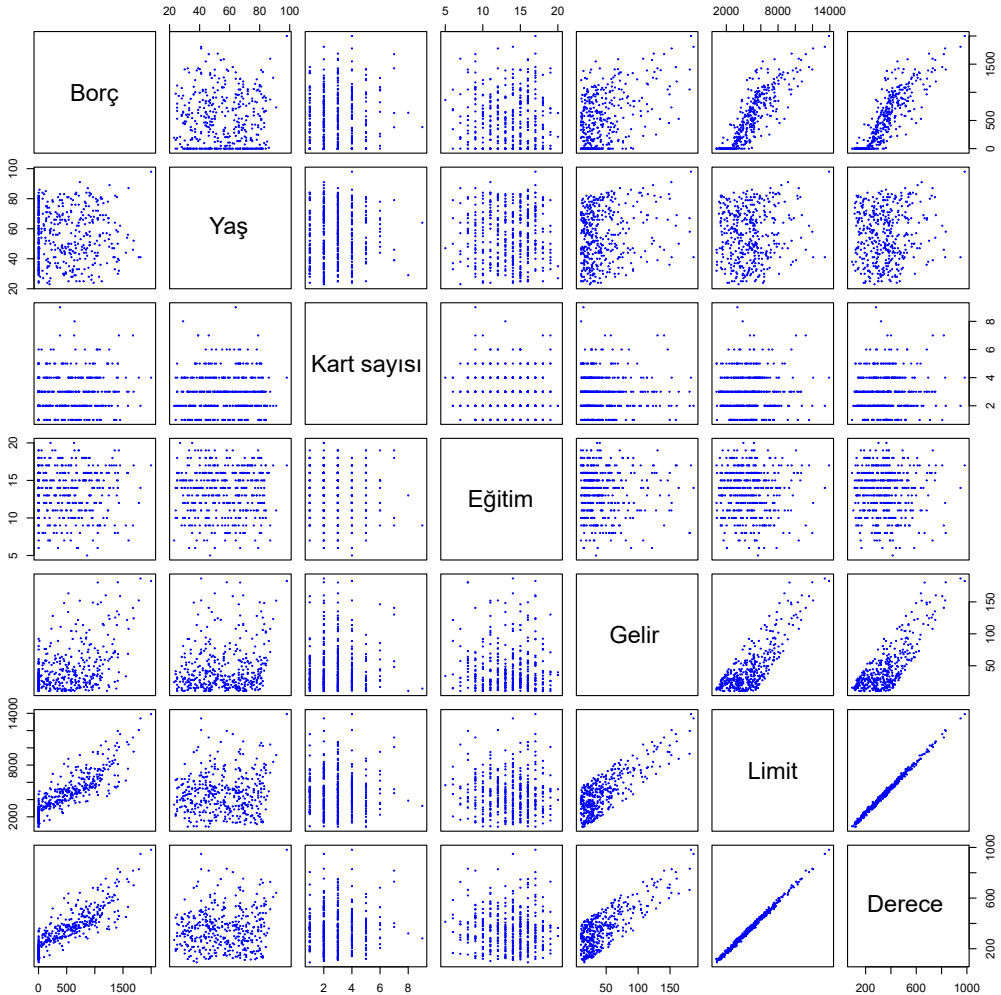
- Katsayı güven aralıkları, veri setindeki tüm Y değerlerine ilişkin ortalama belirsizliğe ilişkindir. Kestirim güven aralığı ise tek bir Y kestirim değerine ait belirsizliği gösterir.
- Bu yüzden kestirim aralıkları, güven aralıklarından daha geniştir. Örneklem ortalamasından uzaklaştıkça kestirim aralığı genişler.

Ders Planı

Nitel Değişkenler

- Bir nicelik yerine sınıflandırma gösteren nitel (qualitative) değişkenlerden önceki bölümde söz etmiştik.
- Regresyon çözümlemesi uygulamalarında X değişkenleri yalnızca nicel değil, nitel de olabilir.
- Örnek olarak, kredi kartı veri setini ele alalım. Bu veri setinde cinsiyet, ırk, medeni durum ve eğitim düzeyi şeklinde dört farklı nitel değişken bulunmaktadır.
- Bunların birbirleriyle ve diğer nicel değişkenlerle olan ilişkisi Şekil 5'te serpilim çizimi matriksi (scatter plot matrix) olarak verilmiştir.

Kredi Kartı Verileri



Şekil 5: Kredi kartı veri setindeki değişkenlere ait serpilim çizimi matriksi

İki Düzeyden Oluşan X Değişkeni

- Basit bir örnek olarak, erkek ve kadınlar arasındaki kredi kartı borcu farkını incelemek istediğimizi düşünelim.
- Bunun için yalnızca iki değer alabilen basit bir gösterge (indicator) değişkeni ya da kukla (dummy) değişken oluştururuz:

$$x_i = \begin{cases} 1 & \text{eğer } i\text{'inci kişi kadınsa} \\ 0 & \text{eğer } i\text{'inci kişi erkekse} \end{cases}$$

- Daha sonra bu değişkeni regresyonumuzda kullanırız:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{eğer } i\text{'inci kişi kadınsa} \\ \beta_0 + \epsilon_i & \text{eğer } i\text{'inci kişi erkekse} \end{cases}$$

- Bu modelde β_0 değeri erkeklerdeki ortalama kredi kartı borcunu gösterir. $\beta_0 + \beta_1$ ise kadınlar için ortalama borçtur.
- Dolayısıyla β_1 burada kadınların erkeklere göre borç farkı olur.
- Kime 0 ya da 1 değeri verdiğimiz sonucu değiştirmez. Eğer erkeklere 1 dersek bu sefer β_1 erkeklerin kadınlara göre farkını verir.

Kukla Değişkende -1, 1 Kodlaması

- Kukla değişkenlere 0 ve 1 değerleri vermek yerine -1 ve 1 değerlerini de kullanabiliriz:

$$x_i = \begin{cases} 1 & \text{eğer } i\text{'inci kişi kadınsa} \\ -1 & \text{eğer } i\text{'inci kişi erkekse} \end{cases}$$

- Yeni durumda model belirtimi aşağıdaki gibi olur:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{eğer } i\text{'inci kişi kadınsa} \\ \beta_0 - \beta_1 + \epsilon_i & \text{eğer } i\text{'inci kişi erkekse} \end{cases}$$

- Burada β_0 parametresi kadın/erkek ayrımı yapılmaksızın ortalama kart borcudur. β_1 ise kadınların bu ortalamanın ne kadar üstünde ve erkeklerin de ortalamanın ne kadar altında olduğunu verir.
- Bu modelin sonuçları önceki model ile aynı çıkar. Aradaki tek fark yorumdadır.
- Öte yandan, uygulamada kukla değişkenleri 0 ve 1 şeklinde kodlamak daha yaygındır.

İkiden Fazla Sınıftan Oluşan X Değişkeni

- Sınıf sayısı ikiden çoksa daha fazla kukla değişken kullanırız:

$$x_{i1} = \begin{cases} 1 & \text{eğer } i\text{'inci kişi Asyalıysa} \\ 0 & \text{eğer } i\text{'inci kişi Asyalı değilse} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{eğer } i\text{'inci kişi beyazsa} \\ 0 & \text{eğer } i\text{'inci kişi beyaz değilse} \end{cases}$$

- Böylece, model aşağıdaki gibi olur:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & i\text{'inci kişi Asyalıysa} \\ \beta_0 + \beta_2 + \epsilon_i & i\text{'inci kişi beyazsa} \\ \beta_0 + \epsilon_i & i\text{'inci kişi zenciyse} \end{cases}$$

- Her zaman sınıf sayısından bir eksik kukla değişken olmalıdır. Kuklası olmayan sınıfa temel ya da karşılaştırma sınıfı denir.
- Bu modelde β_1, Asyalıların zencilere göre borç farkını, β_2 ise beyazların yine zencilere göre borç farkını gösterir.

Kukla Değişkenlerin Yorumlanması

- Kredi kartı borçlarını etnik kökene göre inceleyen regresyon tahminleri aşağıdaki gibidir:

Değişken	Katsayı	Ölçünlü hata	t-istatistiği	p-değeri
Sabit terim	531,00	46,32	11,464	$<$ 0,0001
Asyalı	−18,69	65,02	−0,287	0,7740
Beyaz	−12,50	56,68	−0,221	0,8260

Çizelge 5: Kredi kartı borcu ile etnik köken çoklu regresyonu

- Çizelgede taban sınıf olan zenciler için ortalama kredi kartı borcu 531 dolardır. Bu miktar Asyalılar için 18,69 dolar, beyazlar için ise 12,50 dolar daha düşük bulunmuştur.
- Ancak Asyalılar ve beyazlara ait katsayıların p-değerleri yüksektir. Bu durumda üç grup arasında anlamlı bir fark yoktur diyebiliriz.
- Öte yandan katsayılar ve p-değerleri kuklaların nasıl belirlendiğine de bağlıdır. Dolayısıyla, bu konuda kesin karar vermek için $H_0 : \beta_1 = \beta_2 = 0$ şeklinde bir F sınaması yapmak uygun olur.

Ders Planı

Toplanırlık ve Doğrusallık Varsayımları

- Doğrusal regresyon modeli yorumlaması oldukça kolay ve çıkarsama için de yararlı sonuçlar üretir.
- Ancak bunu genellikle uygulamada geçerli olmayan iki kısıtlayıcı varsayım pahasına yapar:
  1. Y ve X'ler arasında toplanır (additive) bir ilişki vardır.
  2. Y ve X'ler arasında doğrusal (linear) bir ilişki vardır.
- Toplanır ilişki, belli bir X_j'nin Y üzerindeki etkisinin diğer X'lerden bağımsız olması anlamına gelir.
- Doğrusal ilişki ise X_j'deki bir birim değişikliğin Y'ye etkisinin hep sabit olması, X_j'nin büyüklüğünden etkilenmemesi demektir.
- Şimdi, bu iki varsayımı nasıl gevşetebileceğimizi kısaca tartışalım.

Etkileşim Terimi

- İlk olarak, toplanırlık varsayımını ele alalım. Aşağıdaki üç değişkenli modeli inceleyelim:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- Burada X_1 eğer 1 birim artarsa Y de ortalama β_1 birim artmaktadır. Öte yandan bu etki X_2'den bağımsızdır. X_2 sıfır da olsa, yüksek bir değer de alsa etki sabittir.
- Bu durum gerçek yaşamda geçerli olmayabilir. Örnek olarak, TV reklamının etkisi radyo reklamının varlığıyla güçlenebilir. Pazarlamada buna sinerji (synergy) etkisi denilmektedir.
- Bu etkiyi dikkate almanın bir yolu yeni bir değişken eklemektir:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

- Burada β_3'e etkileşim terimi (interaction term) denir. Adından da anlaşıldığın gibi bu terim X_1 ve X_2 arasındaki etkileşimi ölçer.
- Yeni değişkeni X_1 ile X_2'yi çarparak bizim oluşturduğumuza dikkat ediniz.

Etkileşim Teriminin Yorumu (1)

- Etkileşim terimini anlamak için reklam örneğimize geri dönelim:

$$\text{Satış} = \beta_0 + \beta_1 \text{TV} + \beta_2 \text{ Radyo} + \beta_3(\text{TV} \times \text{Radyo}) + \epsilon$$

- Yukarıdaki modeli yorumlamayı kolaylaştırmak amacıyla iki farklı şekilde yeniden yazabiliriz:

$$\text{Satış} = \beta_0 + (\beta_1 + \beta_3 \times \text{Radyo})\text{TV} + \beta_2 \text{ Radyo} + \epsilon$$
$$\text{Satış} = \beta_0 + (\beta_2 + \beta_3 \times \text{TV}) \text{ Radyo} + \beta_1 \text{ TV} + \epsilon$$

- Görüldüğü gibi, etkileşim terimi içeren modelde TV reklamının satışlara etkisi artık $\beta_1 + \beta_3 \times$ Radyo harcaması kadardır.
- Benzer şekilde radyo reklamının etkisi de $\beta_2 + \beta_3 \times$ TV reklam harcamasına bağlıdır.

Etkileşim Teriminin Yorumu (2)

- Modele ait regresyon tahmin sonuçları aşağıda verilmiştir:

Değişken	Katsayı	Ölçünlü hata	t-istatistiği	p-değeri
Sabit terim	6,7502	0,248	27,23	$<$ 0,0001
TV	0,0191	0,002	12,70	$<$ 0,0001
Radyo	0,0289	0,009	3,24	0,0014
TV$\times$Radyo	0,0011	0,000	20,73	$<$ 0,0001

Çizelge 6: Satışlar ile TV ve radyo reklamları etkileşimli regresyonu

- Yukarıda etkileşim teriminin anlamlı olduğu görülmektedir. Ayrıca etkileşimin eklenmesiyle R^2 de 0,897'den 0,968'e yükselmiştir.
- Burada artık bir reklam türünün etkisi diğerinin miktarına bağlıdır.
- Örnek olarak, radyo reklam harcaması 1000 dolar iken 1000 dolarlık TV reklamının satışlara etkisi $19{,}1 + 1{,}1 \times 1 = 20{,}2$ adettir.
- Radyo reklamı 5000 dolar olduğunda ise aynı 1000 dolarlık TV reklamının etkisi artarak $19{,}1 + 1{,}1 \times 5 = 24{,}6$'ya yükselir.
- Radyo reklamlarının etkisi de buna benzer şekilde hesaplanır.

Kukla Etkileşim Terimi (1)

- Etkileşim terimlerini kukla değişkenlerle de kolayca kullanabiliriz.
- Örnek olarak, kredi kartı borcunun gelire ve öğrenci olma niteliğine göre regresyonu etkileşim terimiyle birlikte şöyle modellenir:

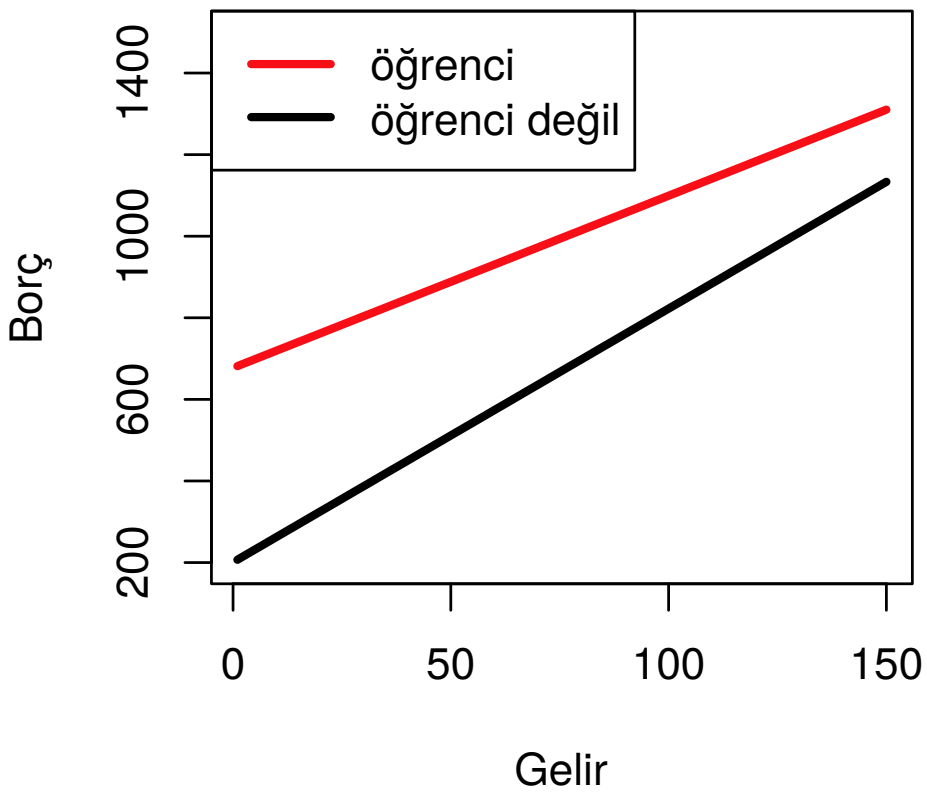
$$\text{Borç} = \beta_0 + \beta_1 \text{ Gelir} + \beta_2 \text{ Öğrenci} + \beta_3(\text{Gelir} \times \text{Öğrenci}) + \epsilon$$

- Bu durumda aşağıdaki regresyon tahmin edilmiş olur:

$$\text{Borç} = \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \text{ Gelir} & \text{eğer öğrenci ise} \\ \beta_0 + \beta_1 \text{ Gelir} & \text{eğer öğrenci değilse} \end{cases}$$

- Yukarıda β_2, ikinci doğrunun (öğrenci olmanın) sabit terim farkı olarak yorumlanır. β_3 ise ikinci doğrunun eğim farkı olur.
- Dolayısıyla aslında iki ayrı regresyon doğrusu tahmin ettiğimize dikkat ediniz. Bunlar Şekil 6'da gösterilmiştir.

Kukla Etkileşim Terimi (2)



Şekil 6: Kredi borcunun gelire ve öğrenci olma niteliğine göre regresyonları

Polinom Regresyonu (1)

- Doğrusallık varsayımının uygulamada regresyon modelleri için bir kısıtlama oluşturduğunu yukarıda söylemiştik.
- Doğrusal-dışı ilişkileri dikkate almanın basit bir yolu polinom regresyon (polynomial regression) modelidir.
- Örnek olarak, yakıt tüketimi ile motor gücünü aşağıdaki gibi ikinci derece bir polinom regresyonuna yakıştırabiliriz:

$$\text{Yakıt tüketimi} = \beta_0 + \beta_1\ \text{Güç} + \beta_2\ \text{Güç}^2 + \epsilon$$

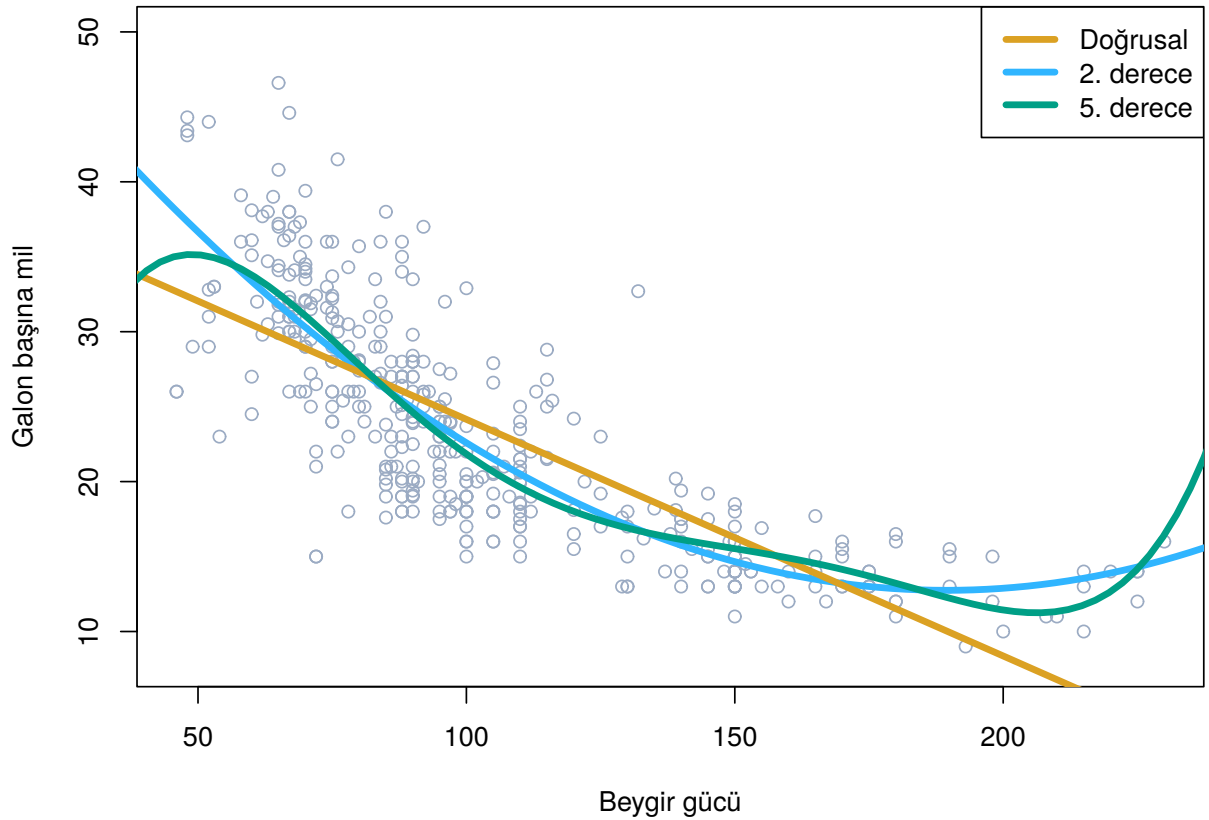
- Daha yüksek derece polinom regresyonları da buna benzerdir:

$$\text{Yakıt tüketimi} = \beta_0 + \beta_1\ \text{Güç} + \beta_2\ \text{Güç}^2 + \cdots + \beta_p\ \text{Güç}^p + \epsilon$$

- Bu modellerde regresyon çizgisi bir doğru değil, eğri şeklindedir. Dolayısıyla X'in Y'ye etkisi X'in büyüklüğüne göre değişir.
- Otomobil veri seti kullanılarak tahmin edilmiş doğrusal model ile 2. derece ve 5. derece polinom modelleri Şekil 7'deki gibidir.

Polinom Regresyonu (2)

- Şekilde 2. derece polinom regresyonunun verilere iyi yakıştığı, 5. derece polinomun ise gereksiz derecede kıvrımlı olduğu görülmektedir. Dolayısıyla esneklik seçimi burada da önemlidir.



Şekil 7: Yakıt tüketimi ile motor gücüne ilişkin polinom regresyonlar

Ders Planı

Uygulamada Karşılaşılan Sorunlar

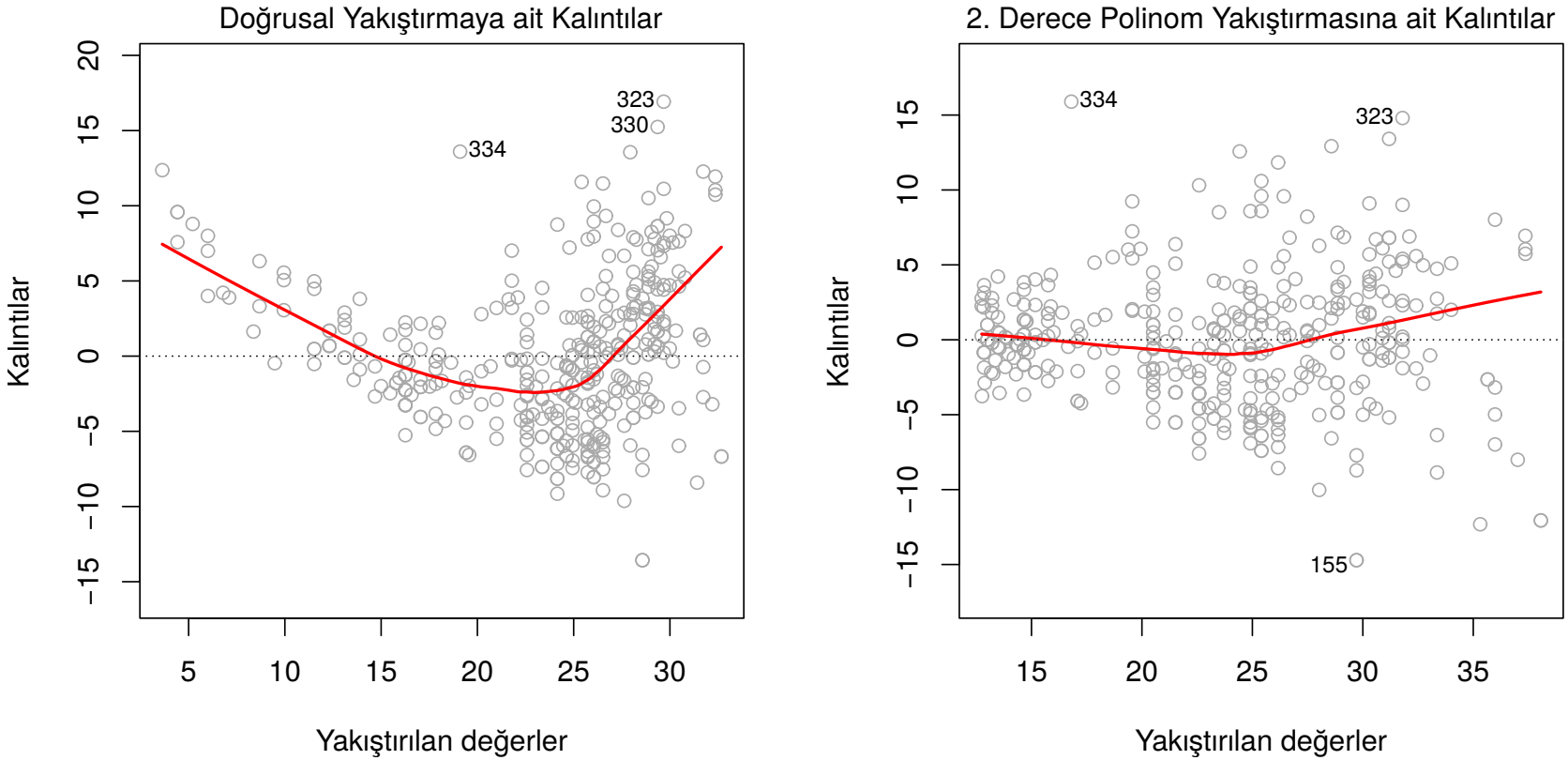
- Bir veri setine regresyon modeli yakıştırdığımız zaman çeşitli sorunlarla karşılaşabiliriz. Bunların başlıcaları şunlardır:
  1. Modelleme hatası
  2. Hata teriminde korelasyon
  3. Hata teriminde farklıserpilimsellik
  4. Dışadüşenler
  5. Eşdoğrusallık
- Yukarıdaki hataları saptamak ve düzeltmek oldukça ayrıntılı başlıklardır. Bu konularda yazılmış birçok kitap bulunmaktadır.
- Burada biz uygulamada karşılaşılan olası sorunları kısa ve öz bir şekilde ele alacağız.
- Modelleme hatası ile başlayalım.

Modelleme Sorunu (1)

- Regresyon yönteminde modelleme sorunu genellikle doğrusal-dışı ilişkiler modelde dikkate alınmadığı zaman ortaya çıkar.
- Bu durumda tüm tahminler kuşkulu duruma düşer ve modelin kestirim gücü de ciddi oranda azalabilir.
- Doğrusal-dışı ilişkileri saptamanın iyi bir yolu değişkenleri ya da kalıntıları şekil üzerinde incelemektir.
- Bu amaçla regresyon kalıntıları ile yakıştırılan $\hat{y}_i$ değerlerinin çizimine bakılır. Ayrıca farklı X'lerin Y'ye karşı çizimleri de yararlıdır.
- Eğer görsel incelemede doğrusal-dışı ilişki bulunursa değişkenler üzerinde $\log(X)$, $\sqrt{X}$ gibi dönüştürmeler yapılabilir ya da X^2 gibi yeni değişkenler modele eklenebilir.
- Örnek olarak, otomobil veri setindeki yakıt tüketimi ve motor gücü regresyonuna ilişkin kalıntılar ile yakıştırılan değerler Şekil 8'de gösterilmiştir.

Modelleme Sorunu (2)

- Sol paneldeki doğrusal modele ait kalıntılar güçlü bir doğrusal-dışı örüntü göstermektedir. Modele X^2 eklenerek elde edilen sağ paneldeki polinom modelde ise sorun büyük oranda düzelmiştir.



Şekil 8: Doğrusal ve polinom regresyonlara ait kalıntı çizimleri

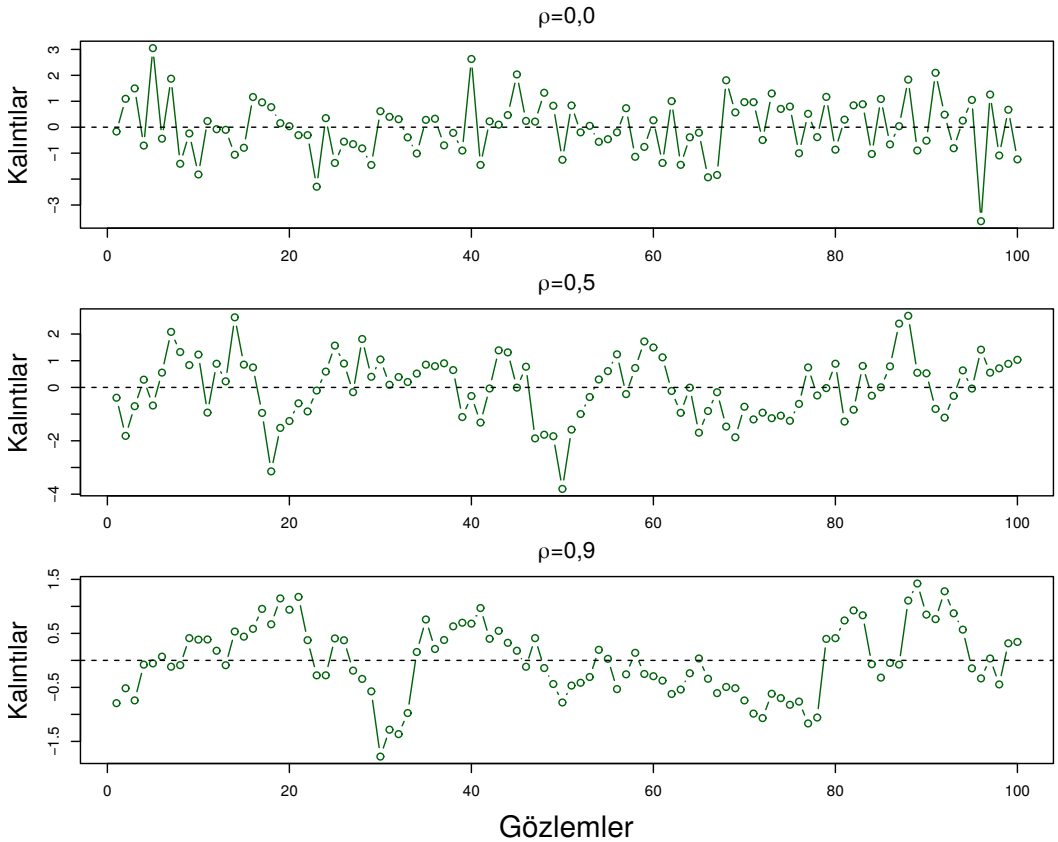
Ders Planı

Hata Teriminde Korelasyon (1)

- Doğrusal regresyon modelinin en önemli varsayımlarından biri hata terimleri $\epsilon_1, \epsilon_2, \ldots \epsilon_n$ arasında korelasyon olmamasıdır.
- Örnek olarak, ϵ_i'nin aldığı değer ϵ_{i+1}'i etkilememelidir.
- Eğer bu varsayım çiğnenirse tahmin edilen tüm güven/kestirim aralıkları ile *p*-değerleri hatalı ve yanıltıcı olur.
- Özilinti (autocorrelation) denilen bu olgu en çok zaman serilerinde görülür. Bunun önemli bir nedeni zaman içinde değişim yavaş gerçekleşirken gözlem değerlerinin ise belli aralıklarla ölçülmesidir.
- Bu sorunu saptamak için regresyon kalıntılarını inceleriz. Özilinti olması durumunda ardışık hata terimleri benzer ya da tam zıt değerler alır. Genellikle de art arda benzer değerler görülür.
- Özilinti olgusu zaman serileri dışında da görülen önemli bir konudur. Bu soruna yönelik birçok ileri yöntem geliştirilmiştir ancak biz burada ayrıntıya girmeyeceğiz.
- Farklı derecelerde özilintili regresyon kalıntıları Şekil 9'daki gibidir.

Hata Teriminde Korelasyon (2)

- Üst panelde kalıntılar arası korelasyon yoktur. Alt panelde ise korelasyon yüksek olduğu için art arda pozitif ya da negatif değerler görülmektedir. Orta panelde sorun daha az belirgindir.



Şekil 9: Farklı düzeylerde ardışık korelasyona sahip regresyon kalıntıları

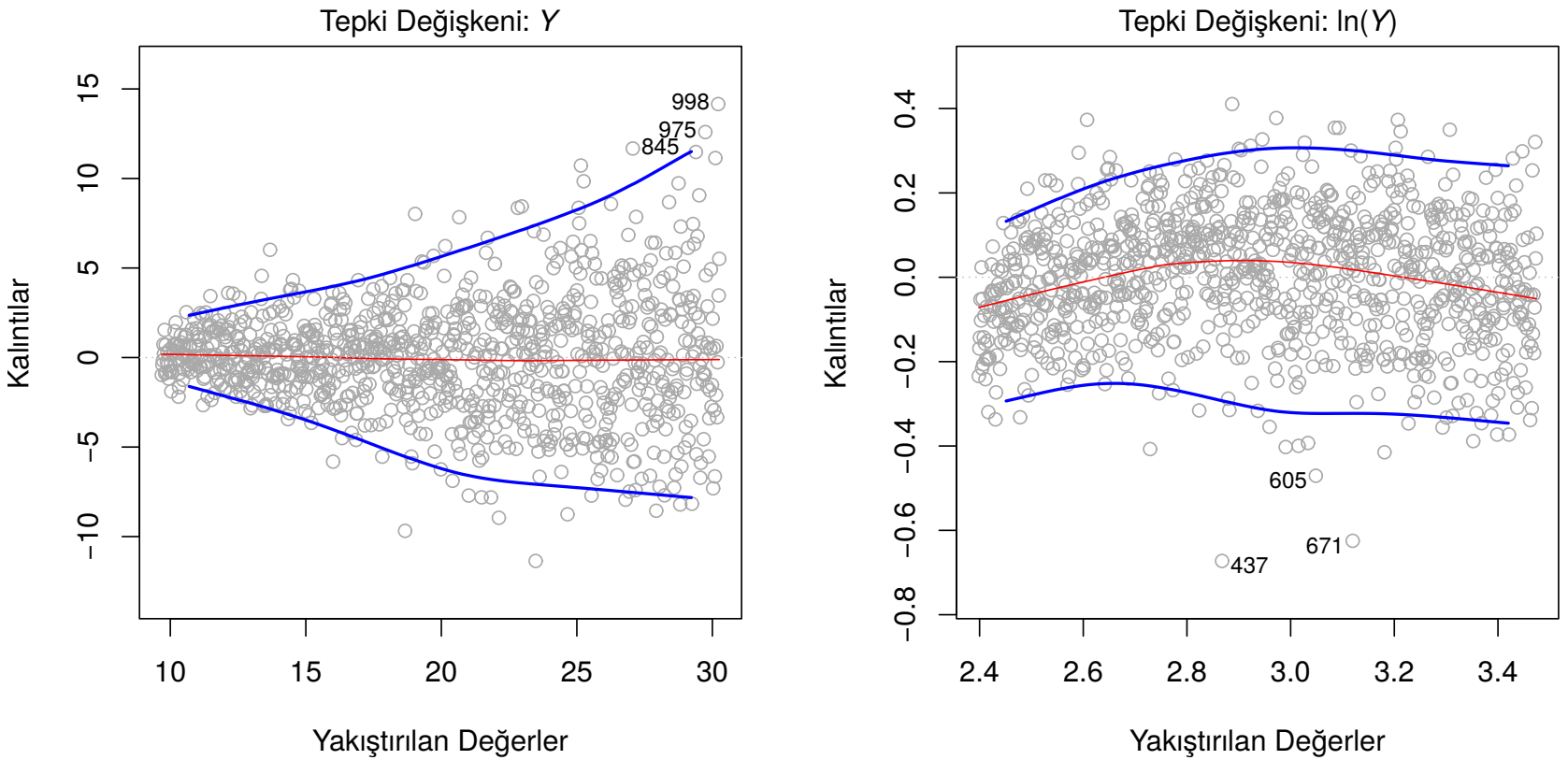
Ders Planı

Hata Teriminde Farklıserpilimsellik (1)

- Doğrusal regresyonun bir diğer varsayımı da hata terimlerinin σ^2 büyüklüğünde sabit varyansa sahip olmasıdır: $\text{var}(\epsilon_i) = \sigma^2$
- Model tahminine ilişkin ölçünlü hatalar, katsayı güven aralıkları ve önsav sınama sonuçları bu varsayımdan güç almaktadır.
- Hata teriminin varyansı sabit olmadığı zaman farklıserpilimsellik (heteroskedasticity) sorunu ortaya çıkar.
- Durumu saptamak amacıyla yine regresyon kalıntılarını inceleriz. Yakıştırılan $\hat{Y}$ değerleri arttıkça kalıntıların yayılımı genişliyor ya da daralıyorsa farklıserpilimsellik var demektir.
- Sorunu çözmenin basit bir yolu X ya da Y'lerin logaritmalarını alarak değişkenleri dönüştürmektir.
- Kalıntıların daha ayrıntılı örüntüler göstermesi durumunda gelişmiş sağlam (robust) ölçünlü hata hesaplamaları ya da ağırlıklı en-küçük kareler (weighted least squares) yöntemi kullanılabilir.
- Farklıserpilimsellik gösteren kalıntılar ve $\ln(Y)$ dönüştürmesi sonucunda elde edilen sorunsuz kalıntılar Şekil 10'da verilmiştir.

Hata Teriminde Farklıserpilimsellik (2)

- Sol panelde giderek genişleyen kalıntılar tipik bir farklıserpilimsellik göstergesidir. Y değişkeninin doğal logaritmasının alınması sonucunda sağ panelde sorun büyük oranda düzelmiştir.

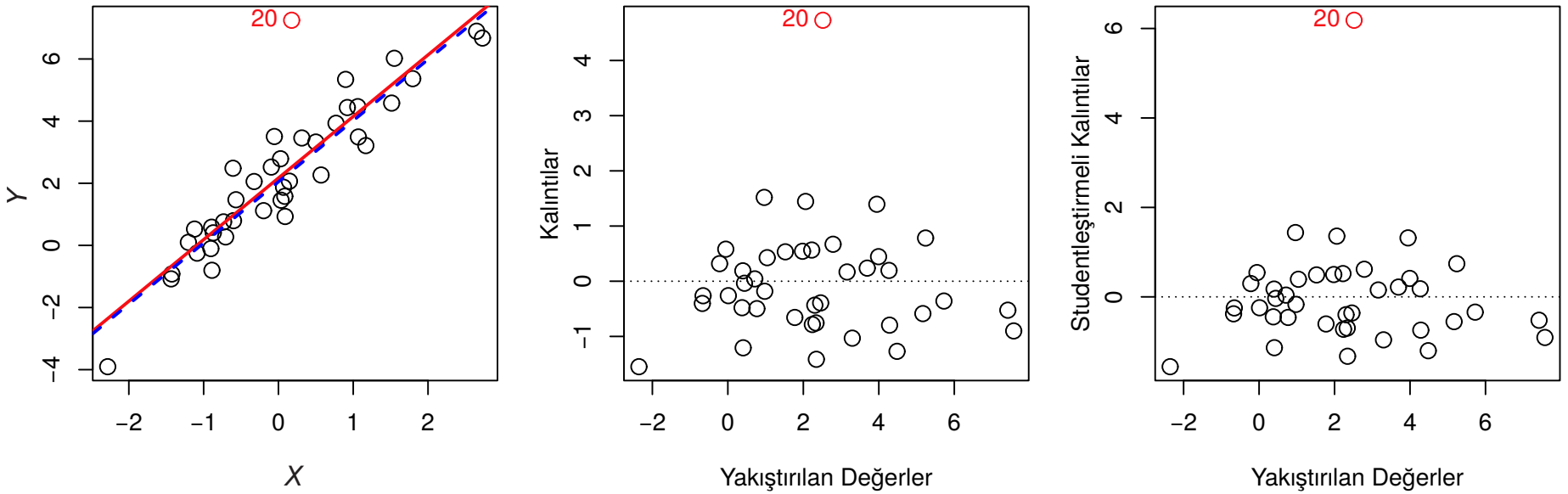


Şekil 10: Farklıserpilimsellik gösteren kalıntılar ve ln(Y) dönüştürmesi

Ders Planı

Dışadüşenler (1)

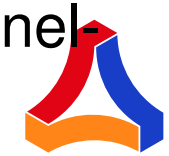
- Dışadüşen (outlier), model tarafından tahmin edilen $\hat{y}_i$'den uzak bir y_i değerine sahip olan gözleme denir. Aşağıda sağ paneldeki gözlem no 20 bir dışadüşendir.



Şekil 11: Dışadüşen gözleme ait X-Y ve kalıntı çizimleri

Dışadüşenler (2)

- Eğer dışadüşenin X değeri ortalama X'ten uzak değilse bu durumda yakıştırılan fonksiyon çok etkilenmez.
- Şekildeki mavi kesikli çizgi dışadüşen çıkartıldığı zaman tahmin edilen regresyon doğrusudur. Görüldüğü gibi kırmızı ve mavi çizgi arasında fazla bir fark yoktur.
- Ancak dışadüşen yine de sonuçları olumsuz etkiler. Örnek olarak, bu regresyonda tek bir dışadüşen nedeniyle KÖH değeri 0,77'den 1,09'a yükselmektedir. R^2 ise 0,892'den 0,805'e düşmektedir.
- Dolayısıyla dışadüşenleri saptamak önemlidir. Bunun için kalıntı çizimine başvururuz. Şekildeki orta panelde 20 no'lu gözleme ait kalıntının diğerlerinden uzak olduğu açıkça görülmektedir.
- Ancak dışadüşenlere karar vermek genellikle daha zordur. Bu yüzden her bir kalıntıyı KÖH'e bölerek şekilde sağ panelde görülen studentleştirmeli (studentized) kalıntıları da hesaplarız.
- Studentleştirmeli kalıntı değeri 3'ten büyük olan gözlemler genellikle dışadüşen kabul edilir.

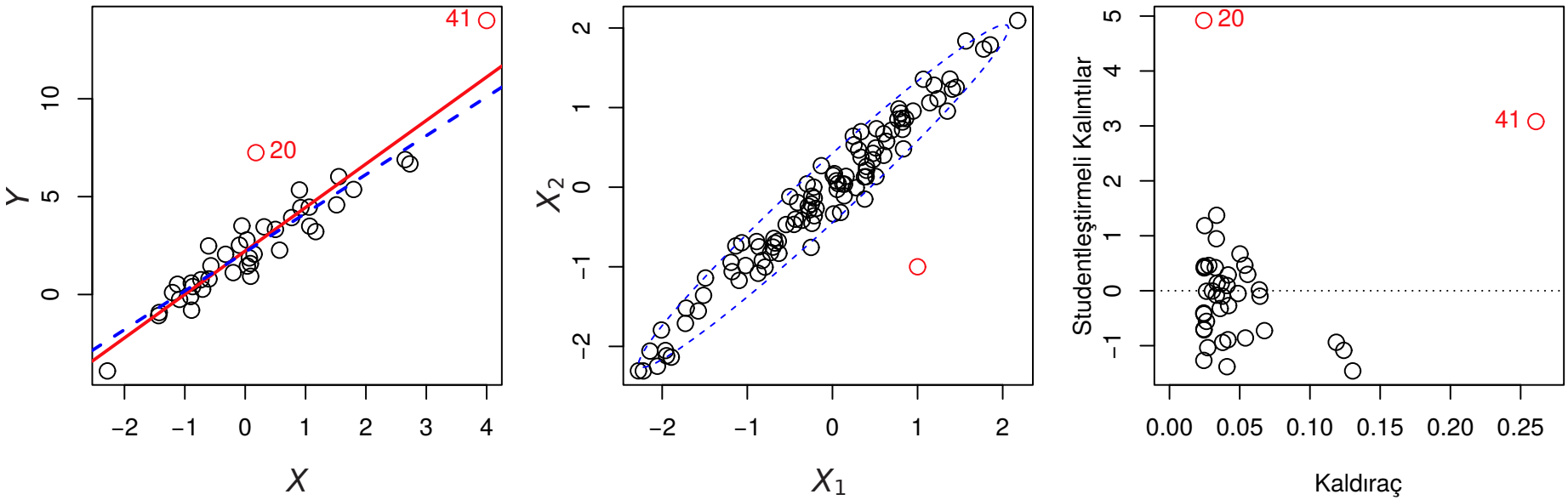


Kaldıraç Etkisi (1)

- Dışadüşenlerin olumsuz etkisini yok etmek için bunları regresyondan çıkartabiliriz. Ancak zorunlu olmadıkça bu doğru değildir çünkü bunlar incelediğimiz ilişki için önemli bilgiler veriyor olabilir.
- Zorunluluk yaratan durum genellikle bir dışadüşenin kaldıraç (leverage) etkisi yüksek olduğu zaman ortaya çıkar.
- Kaldıraç etkisi yüksek dışadüşenler farklı Y ya da X değişkeni kombinasyonlarında "olağanüstü" değerler alan gözlemlerdir.
- Bunlar yakışmanın iyiliğini bozmakla kalmaz, aynı zamanda tahmin edilen regresyon doğrusunu da önemli oranda değiştirirler.
- Kaldıraç etkisi yüksek gözlemleri bulmak için her bir X değişkenini diğer X'lere ve Y'ye karşı tek tek çizmek gereklidir.
- Çoklu regresyonda bu işlem çok zor olabileceği için bunun yerine kaldıraç istatistiği (leverage statistic) hesaplanır. Bulunan değer $(p+1)/n$'den ne kadar büyükse kaldıraç etkisi de o kadar ciddidir.
- Kaldıraç etkisi yüksek olan bir dışadüşen ve bunun yarattığı etkiler Şekil 12'de gösterilmiştir.

Kaldıraç Etkisi (2)

- Solda: Yüksek kaldıraça sahip 41 no'lu gözlem regresyon doğrusunu kendine çekmektedir. Ortada: Böyle sıradışı gözlemler herhangi iki değişkenden kaynaklanabilir. Sağda: Bu yüzden studentleştirmeli kalıntılara ek olarak kaldıraç etkisi de hesaplanır.

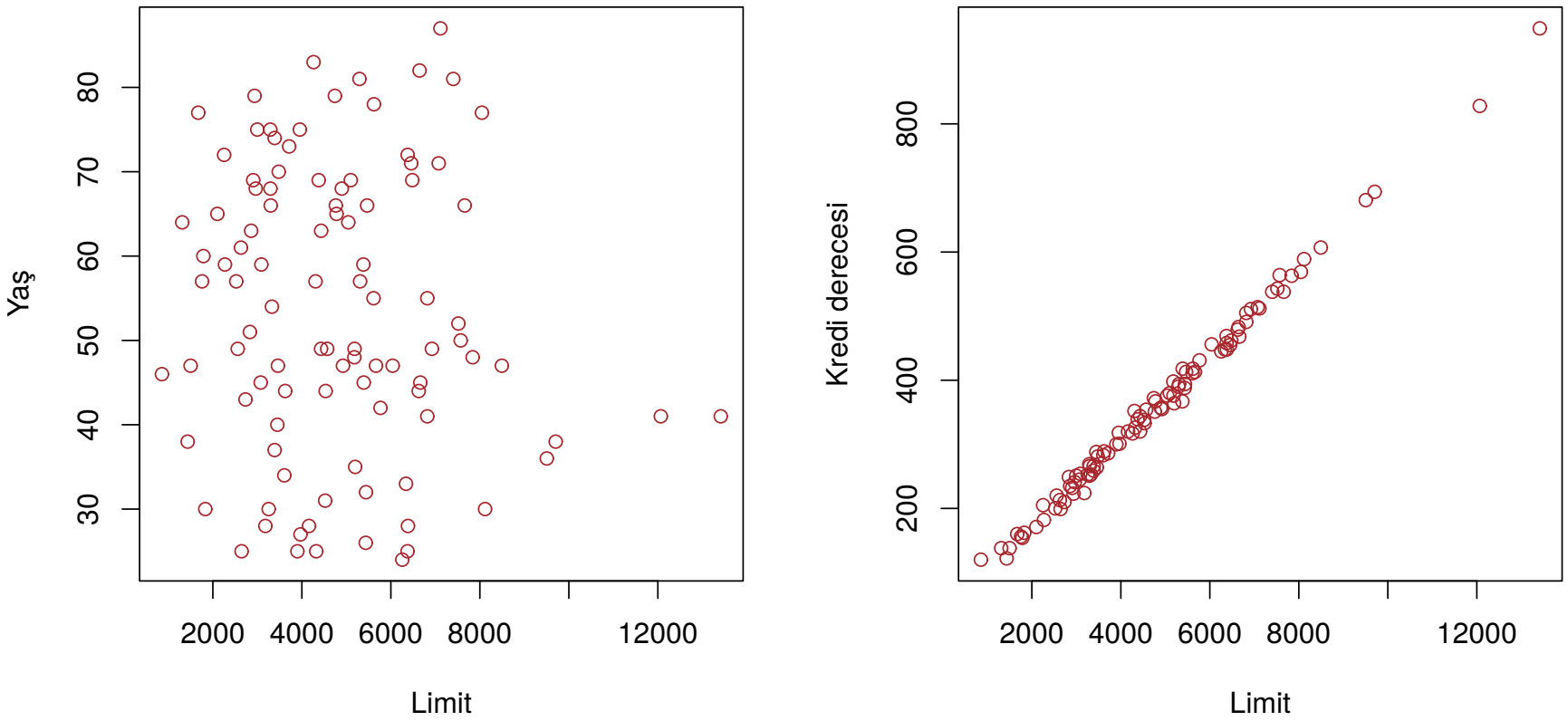


Şekil 12: Kaldıraç etkisi yüksek gözleme ait X-Y ve kalıntı çizimleri

Ders Planı

Eşdoğrusallık (1)

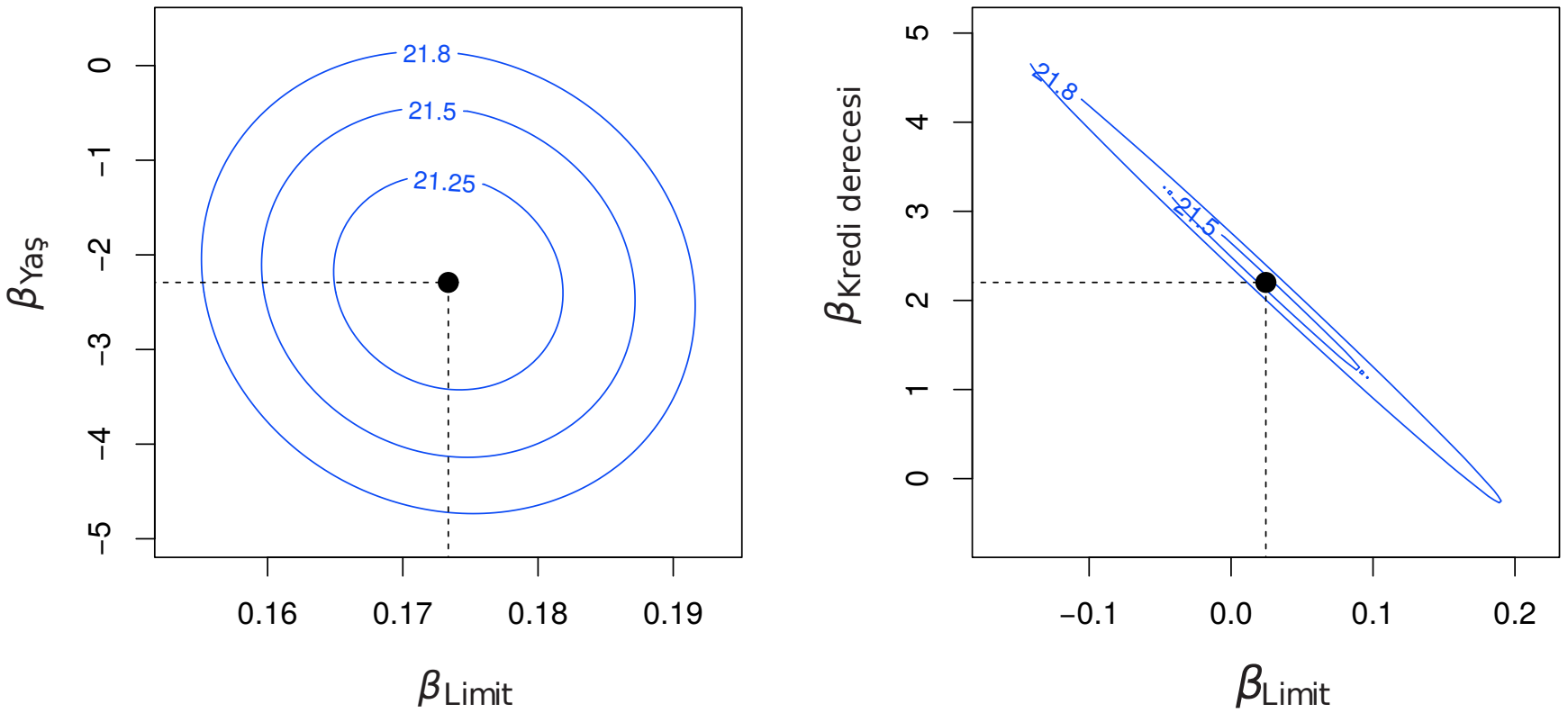
- Eşdoğrusallık (collinearity), iki değişkenin yakın ilişkili olması durumudur. Aşağıda sol panelde eşdoğrusallık yoktur. Sağ panelde ise limit ve kredi derecesi yüksek eşdoğrusallığa sahiptir.



Şekil 13: İki X değişkeni arasında düşük ve yüksek eşdoğrusallık durumu

Eşdoğrusallık (2)

- Regresyonda eşdoğrusallık olduğu zaman tahmin edilen katsayılar birbirine bağımlı hale gelir. Bu durum aşağıda sağ panelde görülmektedir.



Şekil 14: Düşük ve yüksek eşdoğrusallık altında katsayı tahmin kesinliği

Eşdoğrusallık ve Regresyon Tahminleri (1)

- Şekildeki farklı elipsler, aynı KÖH düzeyini veren $\hat{\beta}$ tahmin çiftleridir. Dolayısıyla çizimde farklı kesinlik düzeyleri için olabilecek tüm $\hat{\beta}$ kombinasyonları verilmiştir.
- Eşdoğrusallığı gösteren sağdaki panelde iki $\hat{\beta}$ katsayısının yüksek korelasyona sahip olduğu görülmektedir.
- Bu durumda eldeki eğitim verilerine bağlı olarak bir katsayının tahmini değiştiği zaman öteki de belli bir oranda değişir. Böylece, iki X değişkeninin Y üzerindeki etkilerini ayrıştırmak zorlaşır.
- Buna ek olarak, eşdoğrusallık durumunda ölçünlü hatalar da artar ve katsayı güven aralıkları genişler. Dolayısıyla regresyon tahminlerinin kesinliği düşer.
- Şekildeki sol panele göre β_{Limit} tahmini $[0{,}16; 0{,}20]$ aralığında bulunmuşken sağ panelde ise eşdoğrusallık nedeniyle bu aralığın $[-0{,}2; 0{,}2]$ olarak yaklaşık sekiz kat genişlediğine dikkat ediniz.

Eşdoğrusallık ve Regresyon Tahminleri (2)

- Eşdoğrusallık sorunu göstermeyen ve gösteren iki regresyon çıktısı aşağıda karşılaştırılmıştır.

Değişken	Katsayı	Ölçünlü hata	t-istatistiği	p-değeri
Sabit terim	−173,411	43,828	−3,957	$<$ 0,0001
Yaş	−2,292	0,672	−3,407	0,0007
Limit	0,173	0,005	34,496	$<$ 0,0001

Değişken	Katsayı	Ölçünlü hata	t-istatistiği	p-değeri
Sabit terim	−377,537	45,254	−8,343	$<$ 0,0001
Kredi derecesi	−2,202	0,952	2,312	0,0213
Limit	0,025	0,064	0,384	0,7012

Çizelge 7: Satışlar ile TV, radyo ve gazete reklamlarının ikili regresyonları

- İkinci modelde limit değişkenine ilişkin ölçünlü hatanın yaklaşık 12 kat büyüdüğüne dikkat ediniz. Bunun sonucunda katsayının güven aralığı ciddi oranda genişleyerek anlamlılığını yitirmiştir.

Eşdoğrusallığı Saptamak ve Düzeltmek

- Eşdoğrusallığın başlıca nedeni yüksek korelasyona sahip değişkenlerdir. Dolayısıyla, sorununu saptamanın basit bir yolu X değişkenlerine ait korelasyon matrisini incelemektir.
- Ancak bu tek başına yeterli olmaz çünkü üç ya da daha fazla değişken arasında da karmaşık bir ilişki söz konusu olabilmektedir. Bu duruma çoklueşdoğrusallık (multicollinearity) adı verilir.
- Dolayısıyla, uygulamada sorunu saptamak için regresyondaki herbir katsayıya ait varyans şişme çarpanı (variance inflation factor) ya da kısaca VŞÇ (VIF) değerlerini hesaplarız.
- VŞÇ eğer 10'dan yüksekse eşdoğrusallık sorunu ciddi demektir.
- Örnek olarak, yukarıdaki ikinci regresyonda kredi derecesi ve limit değişkenleri için VŞÇ sırasıyla 160,67 ve 160,59'dur.
- Eşdoğrusallığı çözmenin bir yolu değişenlerden birini çıkarmaktır.
- Almaşık olarak değişkenler birleştirilebilir. Örnek olarak, kredi derecesi ve limitin ortalamasını alarak *kredibilite* diye yeni bir değişken tanımlayabiliriz.

Ders Planı

Parametrik ve Parametrik-dışı Regresyon

- Yukarıda ele aldığımız sıradan enküçük kareler regresyonu parametrik bir istatistiksel öğrenme yöntemidir.
- Parametrik yöntemlerin başlıca üstünlükleri (1) verilere kolay yakıştırılması, (2) kolay yorumlanması, (3) az sayıda gözleme gerek duyması ve (4) istatistiksel sınamaların kolayca yapılabilmesidir.
- Ancak bu yaklaşımın önemli bir sakıncası vardır. O da bilinmeyen $f(X)$ fonksiyonuna yönelik güçlü varsayımlar yapmasıdır.
- Gerçek fonksiyon eğer regresyonda belirtilenden farklıysa parametrik yöntem başarısız sonuç verir.
- Dolayısıyla, amaç kestirim olduğu zaman daha esnek yöntemlere başvurmak yerinde olabilir.
- Şimdi, son olarak, SEK regresyonuna almaşık basit ve iyi bilinen parametrik-dışı yöntemlerden biri olan K-enyakın komşu regresyonu ya da kısaca K-EK regresyonu yöntemini tartışalım.
- Böylece, parametrik ve parametrik-dışı yöntemlerin hangi durumlarda daha iyi sonuç verdiğini de karşılaştırmış olacağız.

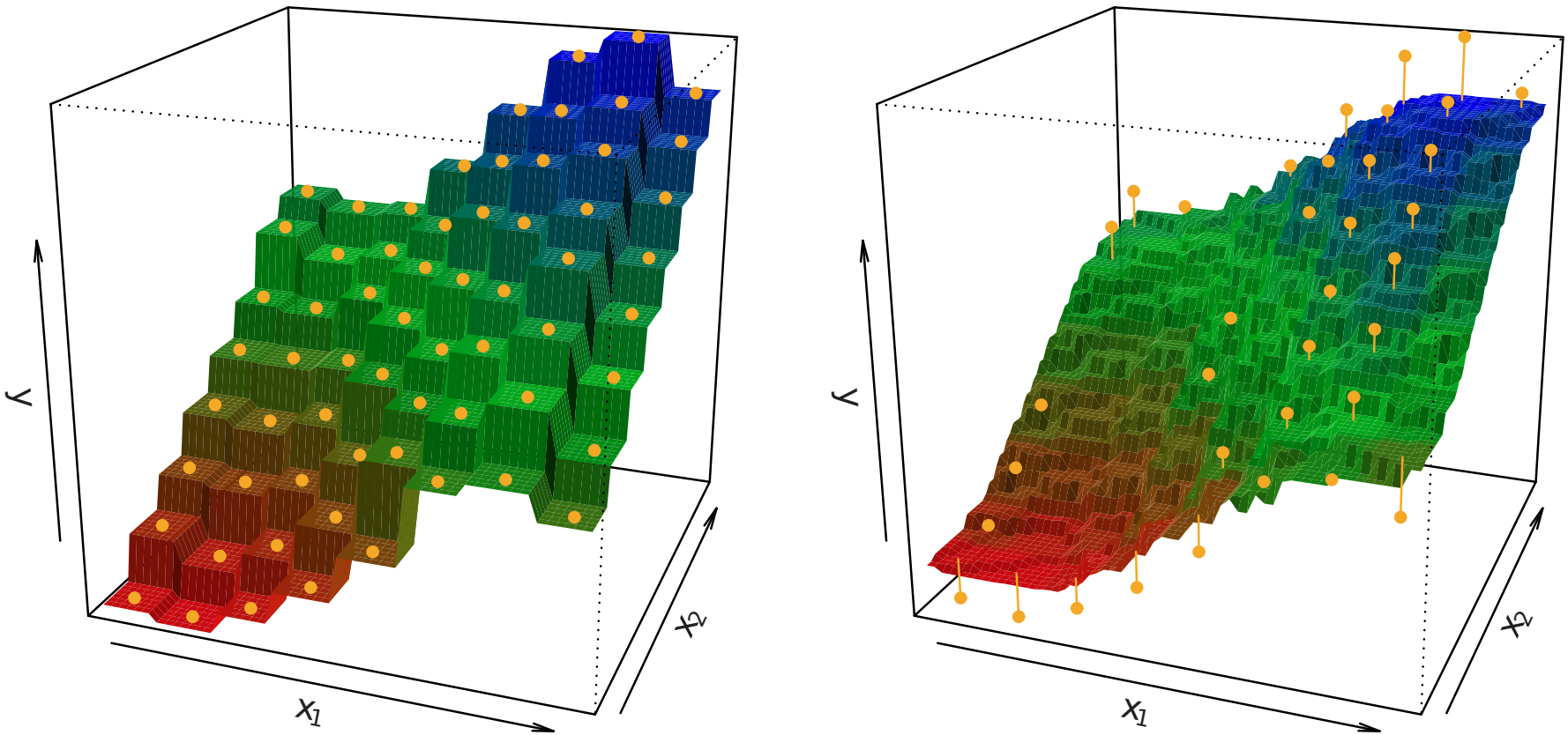
K-Enyakın Komşu Regresyonu (1)

- K-EK regresyonu, 2. Bölümde gördüğümüz K-EK sınıflandırma yöntemiyle yakın ilişkilidir.
- Bu yaklaşımda ilk önce bir K düzleştirme derecesi seçilir. Böylece, farklı x_0 noktaları için en yakın K adet gözlem belirlenir.
- Daha sonra, K-EK regresyonu N_0 ile gösterilen bu gözlem kümesindeki y'lerin x'lere tepkileri ortalamasını alarak $\hat{f}(x_0)$ fonksiyonunu tahmin eder:

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in N_0} y_i$$

- İki adet X açıklayıcı değişkeni içeren bir modele ait K-EK regresyonu sonuçları Şekil 15'te gösterilmiştir.

K-Enyakın Komşu Regresyonu (2)



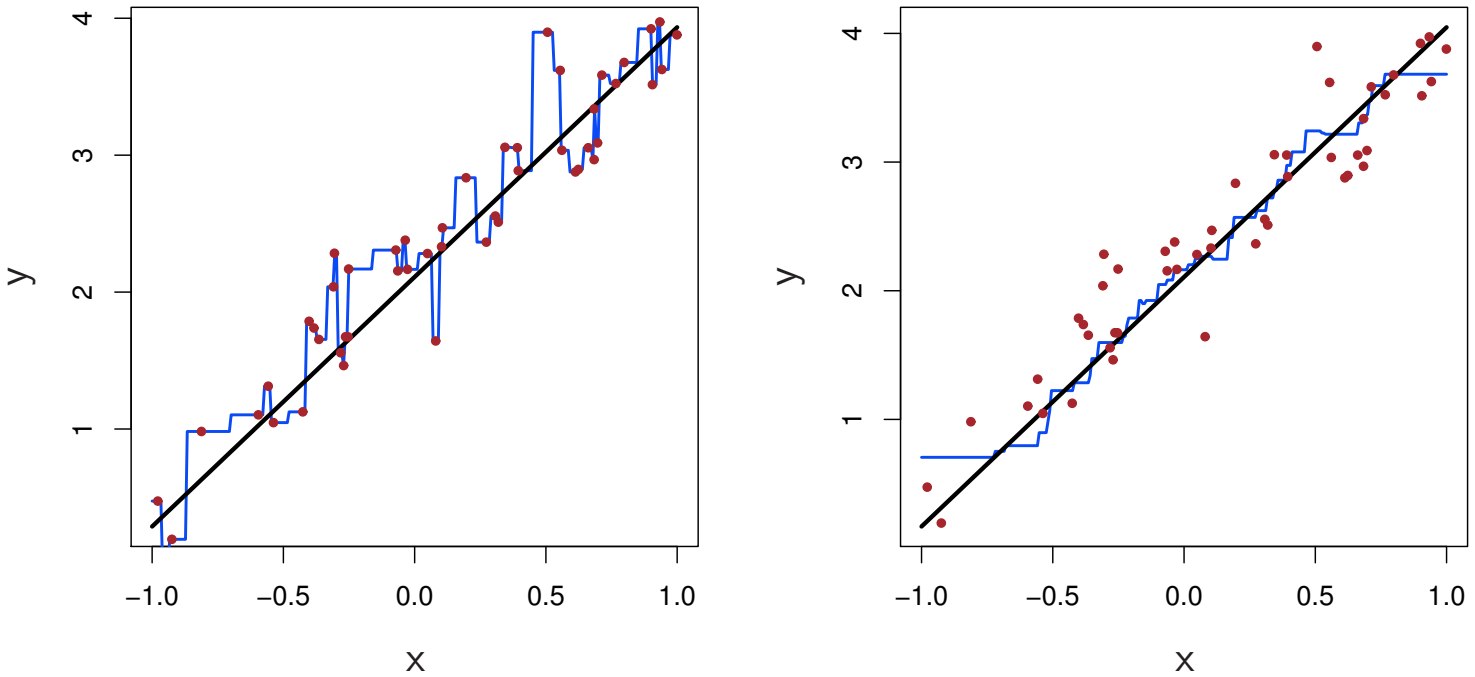
Şekil 15: $K = 1$ ve $K = 9$ için K-EK regresyonu örneği

K-EK ve SEK Regresyonları

- Şekilde solda $K = 1$ ve sağda $K = 9$ değerleri için tahmin edilen K-EK regresyon yüzeyleri görülmektedir.
- Yöntemin basamaklı bir fonksiyon verdiğine dikkat ediniz. Ancak K değeri arttıkça tahmin edilen yüzey de düzleşmektedir.
- K küçük olduğunda tek bir gözlemle sonuç değişeceği için yanlılık düşük, varyans yüksektir. Büyük olduğunda ise çok sayıda gözlemin ortalaması alındığı için varyans düşük, yanlılık yüksektir.
- Dolayısıyla K için en uygun değer de varyans-yanlılık ödünleşmesine bağlıdır. Eniyi K düzeyini belirlemeye yönelik çeşitli yöntemleri 5. Bölümde göreceğiz.
- Peki, parametrik bir yöntem olan SEK ile parametrik-dışı olan K-EK regresyonu arasındaki seçimi nasıl yapacağız?
- Bu, hangi yöntemin daha iyi sonuç vereceğine bağlıdır. Eğer gerçek f fonksiyonunu biliyorsak parametrik yöntemi seçmeliyiz.
- Bunu görmek için X ve Y arasındaki ilişkinin doğrusal olduğu daha basit bir örneği alalım ve aşağıdaki şekilleri inceleyelim.

İkili Doğrusal Modelde K-EK Yönteminin Başarımı

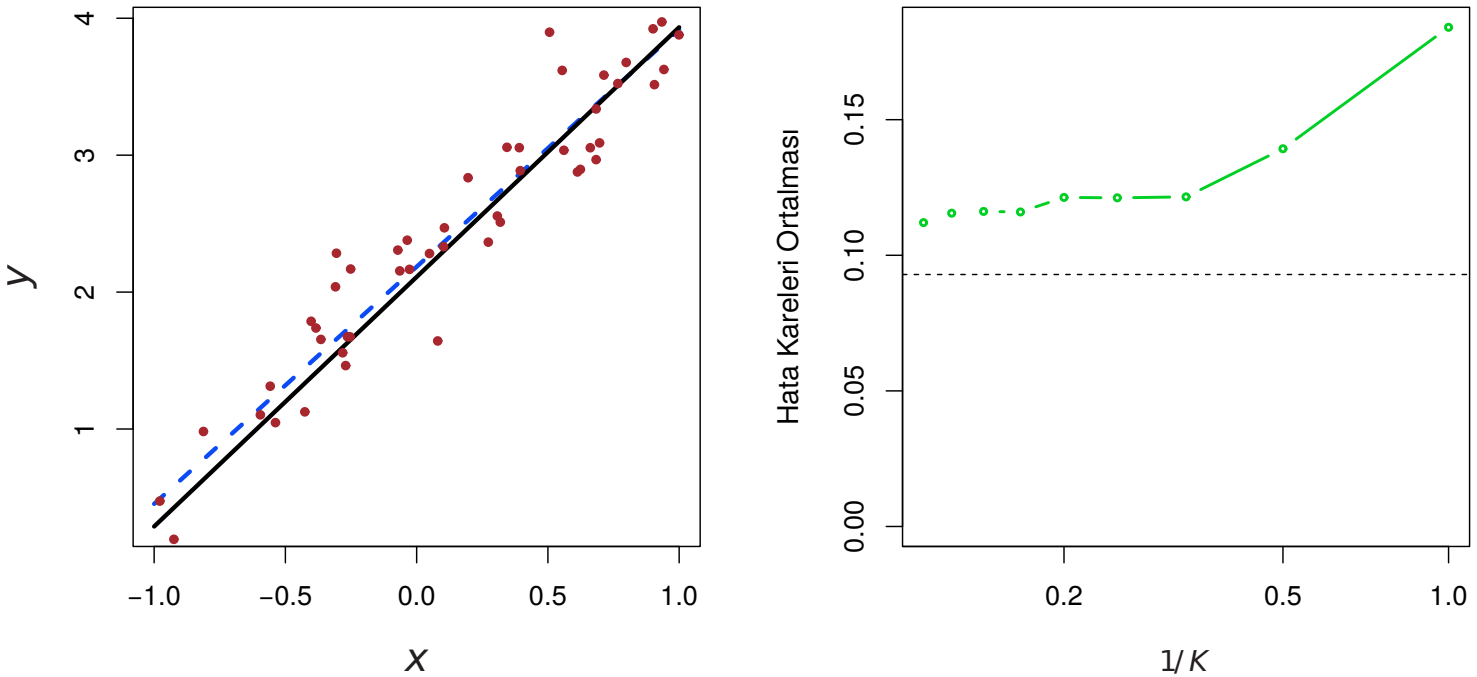
- Solda $K = 1$ iken yakıştırılan K-EK çizgisi verilerin üzerinden geçtiği için yüksek hata oranı vermektedir. Sağda $K = 9$ olduğunda ise yakışma daha düzgün ve doğrusala yakındır.



Şekil 16: $K = 1$ ve $K = 9$ için iki değişkenli K-EK regresyonu

İkili Doğrusal Modelde SEK Yönteminin Başarımı

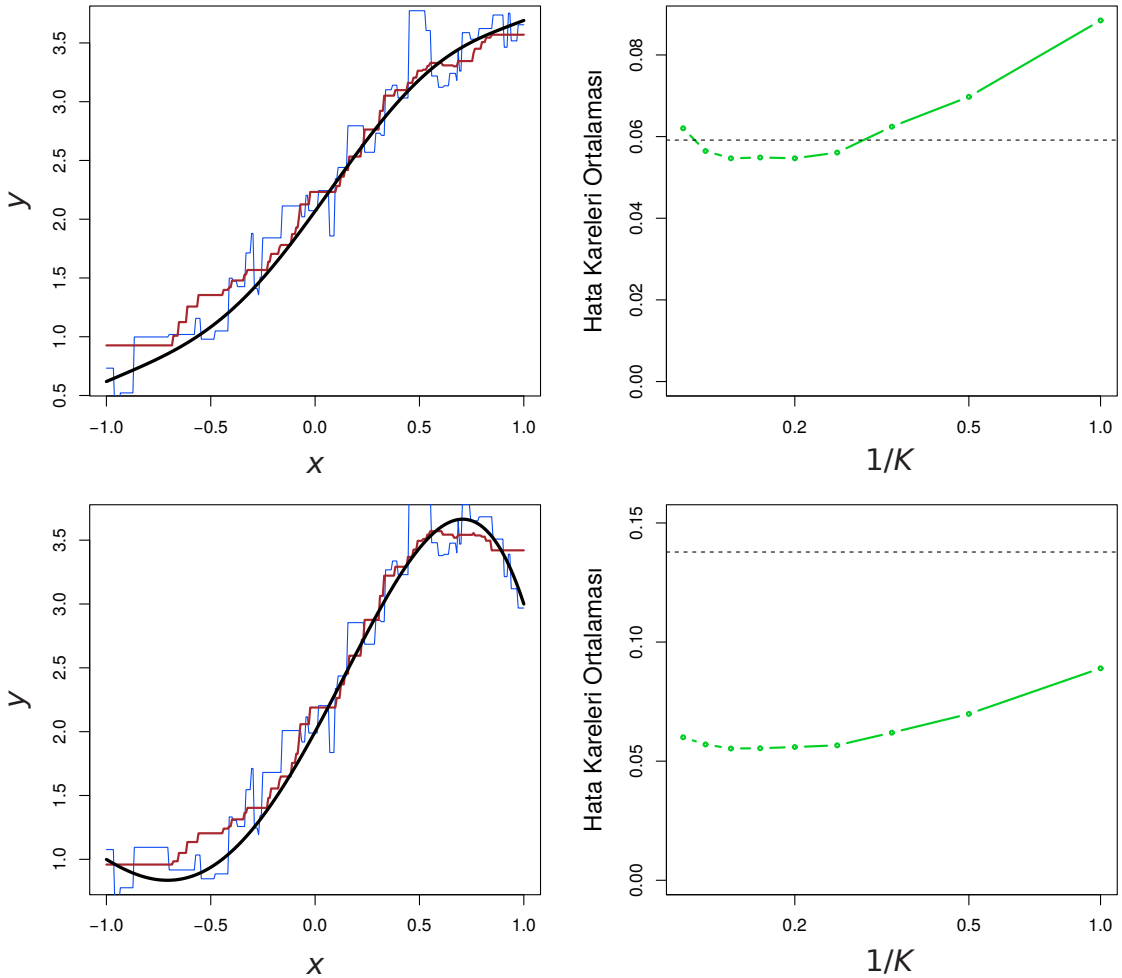
- Solda görülen SEK tahmini gerçek f'ye çok yakın sonuç vermektedir. Sağda ise K-EK yönteminin hata kareleri ortalaması K arttıkça azalmakla birlikte yine de SEK'ten oldukça yüksektir.



Şekil 17: İkili SEK yakıştırması ile SEK ve K-EK'e ait HKO değerleri

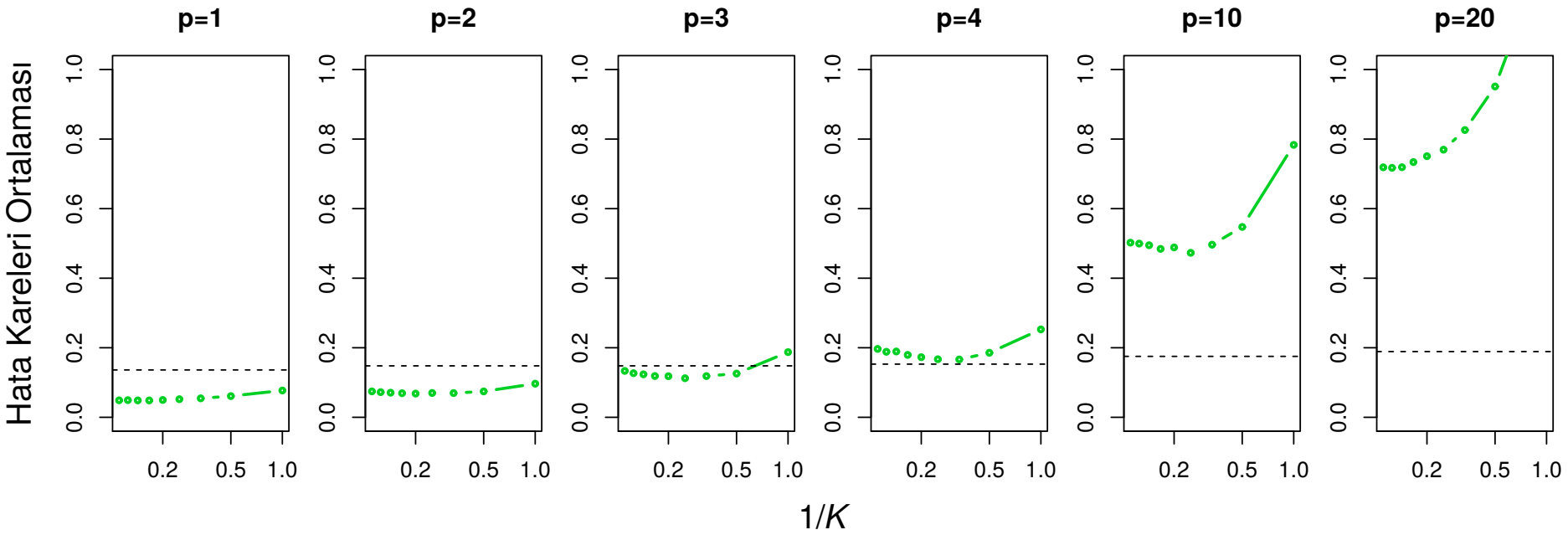
Doğrusal-dışı Modelde SEK ve K-EK Başarımları

- Gerçek yaşamda ilişkiler genellikle doğrusal-dışıdır. Sağ alt panelde görüldüğü gibi, özellikle de doğrusal-dışılık yüksek olduğu zaman K-EK hataları SEK'ten çok daha düşüktür.



Boyutsallık Laneti (1)

- Yukarıda gördüğümüz gibi, ikili regresyonda doğrusal-dışılık arttıkça K-EK, SEK'ten daha başarılı kestirimler üretmektedir.
- Ancak K-EK'in başarımı değişken sayısına da bağlıdır. Aşağıda gözlem sayısı 100 iken modele yeni değişkenler ekledikçe K-EK'e ait hata kareleri ortalamasının giderek arttığını görüyoruz.



Şekil 19: Boyutsallık laneti olgusu

Boyutsallık Laneti (2)

- Yukarıdaki şekilleri birlikte değerlendirdiğimiz zaman doğrusal-dışı ilişkilerde K-EK'in SEK'ten daha başarılı olduğunu görüyoruz.
- Ancak modele yeni değişken ekledikçe, diğer bir deyişle model-deki boyut sayısı arttıkça, bu üstünlük yok olmakta ve K-EK, SEK'e göre çok daha yüksek hata kareleri ortalamaları vermektedir.
- Bu duruma boyutsallık laneti (curse of dimensionality) denir. Sorunun asıl nedeni ise bu örnekte yalnızca 100 gözlem olmasıdır.
- Boyut sayısı arttıkça bir noktadan sonra çok sayıda x_0 kombinasyonuna ait yakın bir komşu bulmak giderek zorlaşır. Bunun sonucunda da K-EK kestiriminin kesinliği ciddi oranda düşer.
- Sonuç olarak, X değişkeni başına düşen gözlem sayısı azaldıkça SEK gibi parametrik yöntemler daha başarılı kestirimler üretir diyebiliriz. Ayrıca parametrik yöntem yorumlama kolaylığı da sağlar.
- Öte yandan, günümüzde verilerin giderek çoğaldığı da unutulmamalıdır. İnternet ölçeğinde verilerle yapılan birçok yapay zeka uygulamasında K-EK tercih edilecektir.

Önümüzdeki Dersin Konusu ve Ödev

Ödev

Kitaptan Bölüm 3 “Doğrusal Regresyon” okunacak.

Önümüzdeki Ders

Sınıflandırma Çözümlemesi