3

Linear Regression (Doğrusal Regresyon)

Bu bölüm, denetimli öğrenme için çok basit bir yaklaşım olan doğrusal regresyon hakkındadır.

Özellikle, doğrusal regresyon, nicel bir yanıtı tahmin etmek için yararlı bir araçtır.

Bu kitabın sonraki bölümlerinde açıklanan daha modern istatistiksel öğrenme yaklaşımlarına göre biraz sıkıcı görünse de, doğrusal regresyon hala yararlı ve yaygın olarak kullanılan istatistiksel öğrenme yöntemidir.

Dahası, daha yeni yaklaşımlar için **iyi bir başlangıç noktası** olarak hizmet eder: sonraki bölümlerde göreceğimiz gibi, birçok süslü istatistiksel öğrenme yaklaşımı, **doğrusal regresyonun genelleştirmeleri veya uzantıları** olarak görülebilir.

Sonuç olarak, daha karmaşık öğrenme yöntemlerini incelemeden önce doğrusal regresyonu iyi anlamanın önemi abartılmış olmaz.

Bu bölümde, doğrusal regresyon modelinin altında yatan **bazı temel fikirlerin** yanı sıra bu modele en sık uyan **en küçük kareler** yaklaşımını gözden geçiriyoruz.

Bölüm 2'deki reklam verilerini hatırlayın. **Şekil 2.1**, TV, radyo ve gazete için reklam bütçelerinin bir fonksiyonu olarak belirli bir ürün için satışları gösterir.

İstatistik danışmanları burada, bu verilere dayanarak, gelecek yıl için yüksek ürün satışlarıyla sonuçlanacak bir pazarlama planı önermemizin istendiğini varsayalım.

Böyle bir tavsiyede bulunmak için hangi bilgiler faydalı olacaktır? İşte ele almaya çalışabileceğimiz birkaç önemli soru :

1. Reklam bütçesi ile satış arasında bir ilişki var mı?

İlk hedefimiz, verilerin, reklam harcamaları ile satışlar arasında bir ilişkiye dair kanıt sağlayıp- sağlamadığını belirlemek olmalıdır.

Kanıt zayıfsa, reklama para harcanmaması gerektiği tartışılabilir!

***

2. Reklam bütçesi ile satışlar arasındaki ilişki ne kadar güçlü?

Reklam ve satış arasında bir ilişki olduğunu varsayarsak, bu ilişkinin gücünü bilmek isteriz. Başka bir deyişle, belirli bir reklam bütçesi verildiğinde, satışları yüksek bir doğruluk düzeyinde tahmin edebilir miyiz? **Bu güçlü bir ilişki olur.**

Yoksa reklam harcamalarına dayalı bir satış tahmini, **rastgele bir tahminden yalnızca biraz daha mı iyidir?** **Bu zayıf bir ilişki olur.**

***

3. Hangi medya satışlara katkı sağlar?

Üç medya aracı - TV, radyo ve gazete - satışlara katkıda bulunuyor mu, yoksa medyanın yalnızca bir veya ikisi mi katkıda bulunuyor?

Bu soruyu cevaplamak için, üç medyaya da para harcadığımızda her bir medyanın bireysel etkilerini ayırmanın bir yolunu bulmalıyız.

***

4. Her bir ortamın satışlar üzerindeki etkisini ne kadar doğru tahmin edebiliriz?

Belirli bir ortamda reklama harcanan her dolar için, satışlar ne kadar artacak?

Bu artış miktarını ne kadar doğru tahmin edebiliriz?

***

5. Gelecekteki satışları ne kadar doğru tahmin edebiliriz?

Herhangi bir televizyon, radyo veya gazete reklamı düzeyi için satış tahminimiz nedir ve bu tahminin doğruluğu nedir?

***

6. İlişki doğrusal mı?

Çeşitli medyadaki reklam harcamaları ile satışlar arasında yaklaşık olarak doğrusal bir ilişki varsa, doğrusal regresyon uygun bir araçtır. Değilse, tahmin ediciyi veya yanıtı dönüştürmek yine de mümkün olabilir, böylece doğrusal regresyon kullanılabilir.

***

7. Reklam medyası arasında sinerji var mı?

Belki televizyon reklamcılığına 50.000 dolar ve radyo reklamcılığına 50.000 dolar harcamak, 100.000 doları televizyona veya radyoya ayrı ayrı ayırmaktan daha fazla satışla sonuçlanır. Pazarlamada bu sinerji etkisi olarak bilinir, istatistikte ise etkileşim etkisi olarak adlandırılır.

Doğrusal regresyonun bu soruların her birini yanıtlamak için kullanılabileceği ortaya çıktı. Önce tüm bu soruları genel bir bağlamda tartışacağız ve sonra **Bölüm 3.4**'te bu özel bağlamda buraya geri döneceğiz.

3.1

Simple Linear Regression (Basit Doğrusal Regresyon)

Basit doğrusal regresyon kendi adıyla yaşar : Tek bir tahmin değişkeni **X** temelinde nicel bir yanıt **Y**'yi tahmin etmek için kullanılan çok basit bir yaklaşımdır.

X ve Y arasında yaklaşık olarak doğrusal bir ilişki olduğunu varsayar. Matematiksel olarak bu doğrusal ilişkiyi şu şekilde yazabiliriz:

$$Y \approx \beta 0 + \beta 1X. \qquad (3.1)$$

"≈" simgesi "yaklaşık olarak modellenmiştir.

Bazen (3.1) 'i X üzerinde Y'yi (veya Y'yi X'e) gerilediğimizi söyleyerek tanımlayacağız. Örneğin, X TV reklamcılığını ve Y satışları temsil edebilir. O zaman modeli uydurarak satışları TV'ye çekebiliriz.

$$\text{sales} \approx \beta 0 + \beta 1 \times \text{TV}.$$

Denklem 3.1'de, **β0** ve **β1** doğrusal modelde **kesişim** ve **eğim** terimlerini temsil eden **iki bilinmeyen sabittir**.

B0 ve β1 birlikte model katsayıları veya parametreleri olarak bilinir.

Tahmini model katsayılarını (**β^0 ve β^1**) bulmak için eğitim setimizi kullandıktan sonra,

$$\hat{y} = \hat{\beta}0 + \hat{\beta}1x, \qquad (3.2)$$

eşitliğini hesaplayarak, TV reklamcılığının belirli bir değeri temelinde gelecekteki satışları tahmin edebiliriz. ^y, X = x temelinde **Y**'nin bir tahminini gösterir.

Burada, bilinmeyen bir parametre veya katsayı için tahmini değeri belirtmek veya yanıtın tahmin edilen değerini belirtmek için şapka sembolü ^ kullanıyoruz.

3.1.1

Estimating the Coefficients (Katsayıların Tahmini)

Pratikte **β0** ve **β1** bilinmemektedir. Dolayısıyla, tahmin yapmak için (3.1) 'i kullanmadan önce, katsayıları tahmin etmek için verileri kullanmalıyız.

Her biri bir X ölçümü ve bir Y ölçümünden oluşan n gözlem çiftini temsil edelim.

$$(x1, y1), (x2, y2),..., (xn, yn)$$

Reklam örneğinde bu veri seti, n = 200 farklı pazardaki TV reklam bütçesi ve ürün satışlarından oluşmaktadır.

Amacımız, doğrusal modelin (3.1) mevcut verilere iyi uyması için **β^0** ve **β^1** katsayı tahminlerini elde etmektir - yani **i = 1, ..., n** için **yi ≈ β^0 + β^1xi**.

Başka bir deyişle, elde edilen doğrunun n = 200 veri noktasına mümkün olduğu kadar yakın olmasını sağlayacak şekilde bir **β^0 kesişme noktası** ve **β^1 eğimi** bulmak istiyoruz.

Yakınlığı ölçmenin birkaç yolu vardır.

Bununla birlikte, en yaygın yaklaşım **en küçük kareler** kriterini en aza indirmeyi içerir.

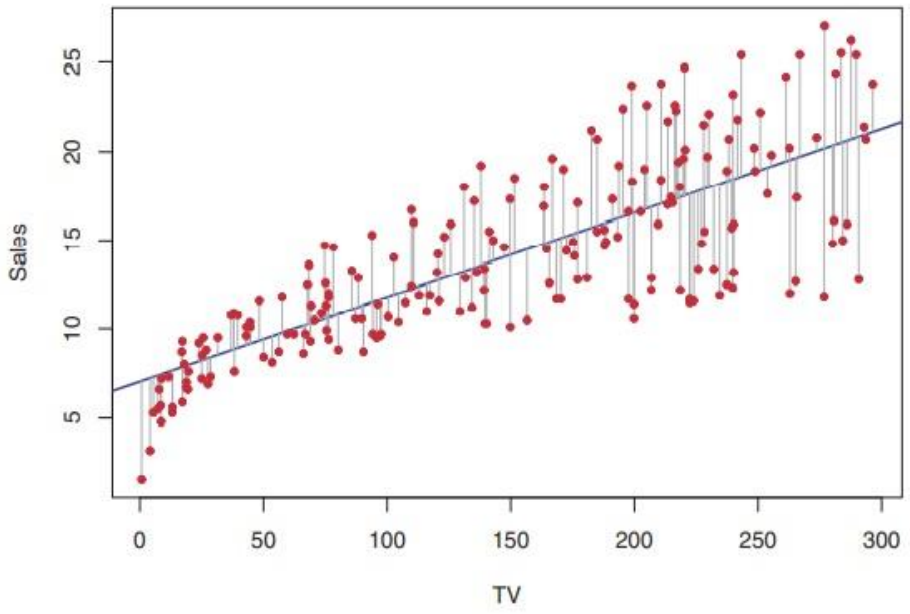

FIGURE 3.1.

Reklam verileri için, satışların TV'ye gerilemesi için en küçük kareler gösterilmektedir.

Uyum –fit-**, hata kareler toplamı** (**SSE**) en aza indirerek bulunur.

Her gri çizgi parçası bir hatayı temsil eder ve uyum –fit-, karelerinin ortalamasını alarak bir uzlaşma sağlar.

Bu durumda, çizginin solunda bir şekilde eksiklik olmasına rağmen, doğrusal uyum (linear fit) ilişkinin özünü yakalar. Kendime not: Yani gidişatı anlıyor.

yi = β^0 + β^1xi, X' in değerine bağlı olarak Y için tahmin olsun. O halde **ei = yi −y^i** , **i. artığı** temsil eder **- bu, artık i. gözlemlenen yanıt değeri ile doğrusal modelimiz tarafından tahmin edilen i. yanıt değeri arasındaki farktır.**

Kalan karelerin toplamını (RSS) şu şekilde tanımlıyoruz:

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

veya eşdeğer olarak şu şekilde tanımlarız:

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2. \quad (3.3)$$

En küçük kareler yaklaşımı, RSS'yi en aza indirmek için βˆ0 ve βˆ1'i seçer.

Bazı hesaplamalar kullanılarak, küçültücülerin (minimizer)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad (3.4)$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

olduğu gösterilebilir. Burada :

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ ve $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ örnek ortalamalarıdır.

Diğer bir deyişle, (3.4) basit doğrusal regresyon için en küçük kareler katsayı tahminlerini tanımlar.

Şekil 3.1, Reklam verilerine uyan basit doğrusal regresyonu gösterir;

burada βˆ0 = 7.03 ve βˆ1 = 0.0475. Başka bir deyişle, bu yaklaşıma göre, TV reklamcılığına harcanan ek 1.000 $, ürünün yaklaşık 47.5 ek biriminin satılmasıyla ilişkilidir.

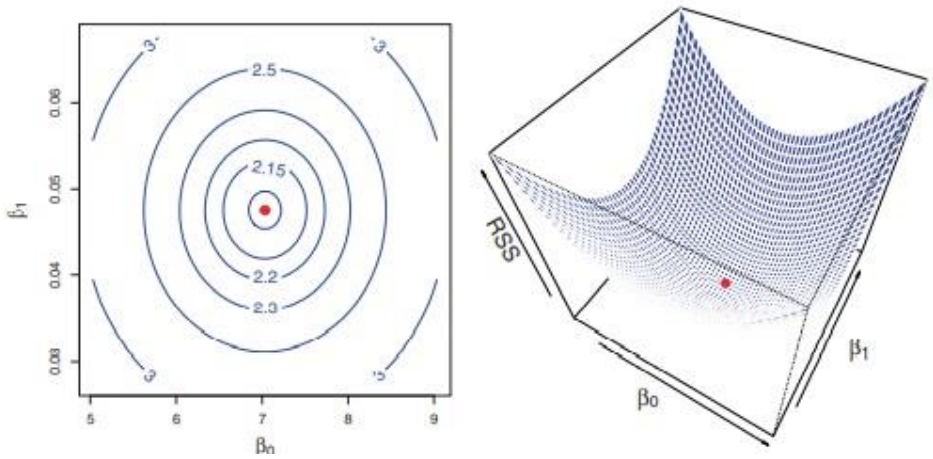

FIGURE 3.2. Contour and three-dimensional plots of the RSS on the Advertising data, using sales as the response and TV as the predictor.

Kırmızı noktalar, (3.4) ile verilen en küçük kareler tahminleri olan βˆ0 ve βˆ1'e karşılık gelir.

Şekil 3.2'de, yanıt olarak satış ve öngörücü olarak TV ile reklam verilerini kullanarak, β0 ve β1 değerleri için RSS'yi hesapladık. Her grafikte kırmızı nokta, (3.4) ile verilen en küçük kareler tahmin çiftini (βˆ0, βˆ1) temsil eder. Bu değerler RSS'yi açıkça en aza indirir

3.1.2

Assessing the Accuracy of the Coefficient Estimates

(Katsayı Tahminlerinin Doğruluğunun Değerlendirilmesi)

(2.1) 'den, X ve Y arasındaki gerçek ilişkinin, ortalama sıfır rastgele hata terimi olan bazı bilinmeyen f fonksiyonları için $Y = f(X) + \epsilon$ biçimini aldığını varsaydığımızı hatırlayın.

Eğer f doğrusal bir fonksiyonla yaklaşacaksa, bu ilişkiyi şöyle yazabiliriz :

$$Y = \beta_0 + \beta_1 X + \epsilon. \qquad (3.5)$$

Burada **β0 kesişme terimidir** - **yani X = 0 olduğunda Y'nin beklenen değeri** ve **β1 eğimdir** - **X'teki bir birimlik artışla ilişkili Y'deki ortalama artış**.

Hata terimi, bu basit modelde gözden kaçırdığımız her şeyi kapsayan bir kavramdır: gerçek ilişki muhtemelen doğrusal değildir, Y'de varyasyona neden olan başka değişkenler ve ölçüm hatası olabilir. Genellikle hata teriminin X'ten bağımsız olduğunu varsayıyoruz.

(3.5) ile verilen model, X ve Y arasındaki gerçek ilişkiye en iyi doğrusal yaklaşım olan popülasyon regresyon çizgisini tanımlar.(1) En küçük kareler regresyon katsayısı tahminleri (3.4), en küçük kareler çizgisini (3.2) karakterize eder.

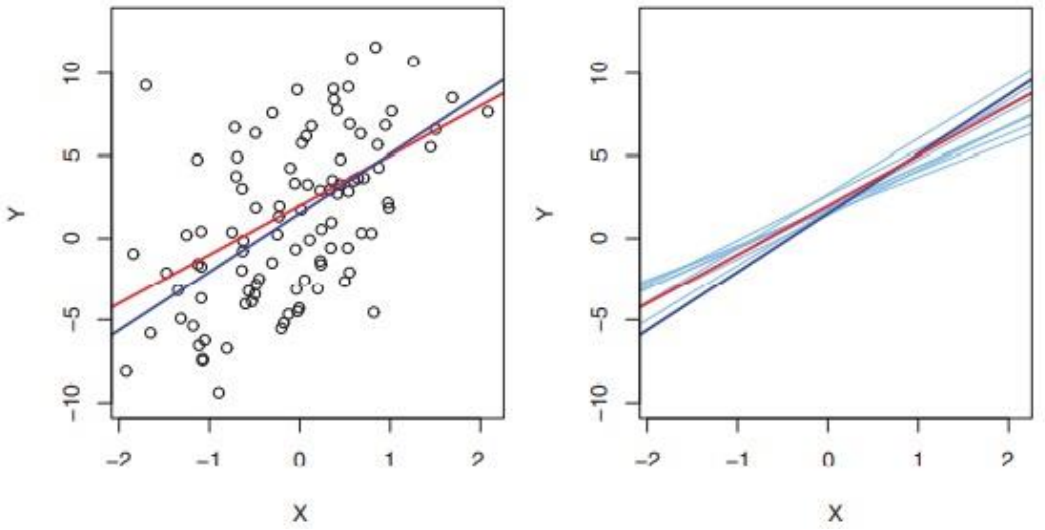

FIGURE 3.3. Simüle edilmiş bir veri seti. **Sol**: Kırmızı çizgi, popülasyon regresyon çizgisi olarak bilinen gerçek ilişkiyi, **f (X) = 2 + 3X**'i temsil eder. Mavi çizgi, en küçük kareler çizgisidir; siyahla gösterilen, gözlenen verilere dayalı olarak **f (X)** için en küçük kareler tahminidir.

Sağda: Nüfus regresyon çizgisi yine kırmızı ile gösterilir ve en küçük kareler koyu mavi ile gösterilir. Açık mavi renkte, her biri ayrı bir rasgele gözlem setine göre hesaplanan en küçük on kare çizgi gösterilir. Her en küçük kareler çizgisi farklıdır, ancak ortalama olarak en küçük kareler çizgileri, nüfus regresyon çizgisine oldukça yakındır.

Şekil 3.3'ün sol paneli, bu iki çizgiyi basit bir simüle edilmiş örnekte gösterir. 100 rastgele X oluşturduk ve **e**'nin ortalama sıfır ile normal bir dağılımdan üretildiği modelden 100 karşılık gelen Y üretti:

$$Y = 2 + 3X + \epsilon, \qquad (3.6)$$

Şekil 3.3'ün sol tarafındaki paneldeki kırmızı çizgi, gerçek ilişkiyi, **f (X) = 2 + 3X** gösterirken, mavi çizgi, gözlemlenen verilere dayalı en küçük kareler tahminidir.

Gerçek veriler için gerçek ilişki genellikle bilinmemektedir, ancak en küçük kareler doğrusu her zaman (3.4) 'te verilen katsayı tahminleri kullanılarak hesaplanabilir.

Başka bir deyişle, gerçek uygulamalarda, en küçük kareler doğrusunu hesaplayabileceğimiz bir dizi gözleme erişimimiz var; ancak, popülasyon regresyon çizgisi gözlemlenmemiştir.

Şekil 3.3'ün sağ panelinde (3.6) ile verilen modelden on farklı veri seti oluşturduk ve karşılık gelen en küçük on kare çizgisini çizdik.

Aynı gerçek modelden üretilen farklı veri kümelerinin biraz farklı en küçük kareler çizgileriyle sonuçlandığına, ancak gözlemlenmemiş popülasyon regresyon çizgisinin değişmediğine dikkat edin.

İlk bakışta, **popülasyon regresyon çizgisi** ile **en küçük kareler çizgisi** arasındaki fark ince ve kafa karıştırıcı görünebilir.

Yalnızca bir veri kümemiz var ve bu durumda iki farklı satırın tahmin edici ve yanıt arasındaki ilişkiyi tanımlaması ne anlama geliyor?

Temel olarak, bu iki çizginin kavramı, standart istatistiğin doğal bir uzantısıdır. Büyük bir popülasyonun özelliklerini tahmin etmek için bir örnekten alınan bilgileri kullanma yaklaşımı.

Çrneğin, bazı rasgele değişken Y'nin popülasyon ortalamasını μ bilmekle ilgilendiğimizi varsayalım. Maalesef **μ** bilinmiyor, ancak Y'den gelen **n** adet gözlemlere erişimimiz var, bunları **y1, ..., yn** olarak yazabiliriz ve **μ**'yu tahmin etmek için kullanabiliriz. Makul bir tahmin $\hat{\mu} = \bar{y}$,dir, burada $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ örnek ortalamadır.

Örneklem ortalaması ve popülasyon ortalaması farklıdır, ancak genel olarak örneklem ortalaması, popülasyon ortalamasının iyi bir tahminini sağlayacaktır.

Aynı şekilde, doğrusal regresyondaki bilinmeyen katsayılar β0 ve β1, popülasyon regresyon çizgisini tanımlar. Bu bilinmeyen katsayıları (3.4) 'te verilen βˆ0 ve βˆ1 kullanarak tahmin etmeye çalışıyoruz.

Bu katsayı tahminleri en küçük kareler çizgisini tanımlar.

Doğrusal regresyon ve rastgele bir değişkenin ortalamasının tahmini arasındaki benzerlik, önyargı kavramına dayanan uygun bir analojidir.

If we use the bias sample mean ˆμ to estimate μ, this estimate is unbiased, in the sense that unbiased on average, we expect ˆμ to equal μ.

Bu tam olarak ne anlama geliyor?

Bu, belirli bir gözlem setine göre **y1, ..., yn,** **ˆμ**'nin **μ** değerini fazla tahmin edebileceği ve başka bir gözlem setine göre ˆμ'nin μ değerini olduğundan daha az tahmin edebileceği anlamına gelir.

Ancak, çok sayıda gözlem setinden elde edilen çok sayıda μ tahmininin ortalamasını alabilirsek, bu ortalama tam olarak μ'ye eşit olacaktır.

Bu nedenle, tarafsız bir tahminci, gerçek parametreyi sistematik olarak fazla veya az tahmin etmez.

Tarafsızlık özelliği, (3.4) ile verilen en küçük kareler katsayısı tahminleri için de geçerlidir: belirli bir veri setine dayanarak β0 ve β1'i tahmin edersek, tahminlerimiz tam olarak β0 ve β1'e eşit olmayacaktır.

Ancak, çok sayıda veri seti üzerinden elde edilen tahminlerin ortalamasını alabilirsek, bu tahminlerin ortalaması yerinde olacaktır!

Yani, tek bir yerden alma her yerden al diyo. :D

Şekil 3.3'ün sağdaki panelinden, her biri ayrı bir veri kümesinden tahmin edilen birçok en küçük kare çizgisinin ortalamasının, gerçek popülasyon regresyon çizgisine oldukça yakın olduğunu görebiliriz.

Rastgele bir değişken Y'nin popülasyon ortalaması μ tahminiyle analojiye devam ediyoruz.

A natural question is as follows: how accurate is the sample mean ˆμ as an estimate of μ?

Birçok veri seti üzerindeki ˆμ'nin ortalamasının μ'ye çok yakın olacağını, ancak tek bir ˆμ'nin, μ'nin önemli ölçüde eksik veya fazla tahmini olabileceğini tespit ettik.

Bu tek ˆμ tahmini ne kadar uzakta olacak?

Genel olarak bu soruyu SE (ˆμ) olarak yazılan standart ˆμ hatasını hesaplayarak yanıtlıyoruz.

İyi bilinen formül:

$$\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}, \qquad (3.7)$$

burada σ, Y'nin her bir **yi** gerçekleştirmesinin standart sapmasıdır.

Kabaca konuşursak, **standart hata bize bu ˆμ'nin gerçek μ değerinden farklı olduğu ortalama miktarı söyler**. Eşitlik 3.7 ayrıca bize bu sapmanın n ile nasıl küçüldüğünü anlatır - **ne kadar çok gözlemimiz olursa, ˆμ'nin standart hatası o kadar küçük olur**.

Benzer şekilde, βˆ0 ve βˆ1'in gerçek β0 ve β1 değerlerine ne kadar yakın olduğunu merak edebiliriz. Bˆ0 ve βˆ1 ile ilişkili standart hataları hesaplamak için aşağıdaki formülleri kullanıyoruz:

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad (3.8)$$

burada **σ2 = Var(e). (2 burada kare içindir. omega kare)**

Bu formüllerin kesin olarak geçerli olması için, her gözlem için i hatalarının ortak varyans σ2 ile ilintisiz olduğunu varsaymamız gerekir. Bu, Şekil 3.1'de açıkça doğru değildir, ancak formül yine de iyi bir yaklaşımdır. Formülde, xi daha fazla yayıldığında SE (βˆ1) daha küçük olduğuna dikkat edin; sezgisel olarak, bu durumda bir eğimi tahmin etmek için daha fazla kaldıracımız var. Ayrıca ¯x sıfır olsaydı SE (βˆ0) ile SE (ˆμ) aynı olurdu (bu durumda βˆ0 , ¯y'ye eşit olurdu).

Genel olarak, σ2 bilinmemektedir, ancak verilerden tahmin edilebilir. σ tahmini, artık standart hata olarak bilinir ve aşağıdaki formülle verilir:

$\mathrm{RSE} = \sqrt{\mathrm{RSS}/(n-2)}$ Verilerden σ2 tahmin edildiğinde, bir tahminin yapıldığını belirtmek için SE (βˆ1) yazmalıyız, ancak gösterimin basitliği için bu fazladan "şapkayı" çıkaracağız.

Güven aralıklarını hesaplamak için standart hatalar kullanılabilir. % 95 güven aralığı,% 95 olasılıkla aralığın parametrenin gerçek bilinmeyen değerini içerecek şekilde bir değerler aralığı olarak tanımlanır. Aralık, veri örneğinden hesaplanan alt ve üst sınırlar cinsinden tanımlanır. Doğrusal regresyon için, β1 için% 95 güven aralığı yaklaşık olarak

βˆ1 ± 2 · SE (βˆ1) biçimini alır.

Yani, **[βˆ1 - 2 • SE (βˆ1), βˆ1 + 2 • SE (βˆ1)]** aralığının β1'in gerçek değerini içermesi yaklaşık% 95 olasılıktır. Benzer şekilde, β0 için bir güven aralığı yaklaşık olarak **βˆ0 ± 2 · SE (βˆ0)** biçimini alır.

Reklam verisi durumunda, β0 için% 95 güven aralığı [6.130, 7.935] ve β1 için% 95 güven aralığı [0.042, 0.053] 'tür. Bu nedenle, herhangi bir reklamın yokluğunda satışların ortalama 6.130 ile 7.940 birim arasında bir yere düşeceği sonucuna varabiliriz. Ayrıca, televizyon reklamcılığındaki her 1000 $ 'lık artış için ortalama bir 42 ile 53 adet arasında artış olacaktır.

Standart hatalar, hipotez katsayıları üzerinde hipotez testleri yapmak için de kullanılabilir.

En yaygın hipotez testi, boş hipotezin test edilmesini içerir.

H0: X ve Y arasında bir ilişki yok alternatif hipoteze karşı

Ha: X ve Y arasında bir miktar ilişki vardır.

Matematiksel olarak bu :

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0,$$

karşılık gelir ve, β1 = 0 ise, model (3.5) **Y = β0 + e'ye** indirgenir ve X, Y ile ilişkili değildir.

Yokluk hipotezini test etmek için, β1 için tahminimiz olan βˆ1'in sıfırdan yeterince uzak olup olmadığını belirlememiz gerekir ki, β1'in sıfır olmadığından emin olabiliriz.

Yeterince ne kadar uzak?

Bu tabii ki βˆ1'in doğruluğuna bağlıdır - yani SE'ye (βˆ1) bağlıdır. SE (βˆ1) küçükse, o zaman nispeten küçük βˆ1 değerleri bile **β1 ! = 0** olduğuna dair güçlü kanıtlar sağlayabilir ve dolayısıyla X ve Y arasında bir ilişki vardır. Aksine, SE (βˆ1) büyükse, yokluk hipotezini reddetmemiz için βˆ1 mutlak değerde büyük olmalıdır. Uygulamada, aşağıdaki gibi bir t-istatistiği hesaplıyoruz:

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}, \qquad (3.14)$$

βˆ1'in 0'dan uzak olduğu standart sapmaların sayısını ölçer. X ve Y arasında gerçekten bir ilişki yoksa, o zaman (3.14) 'ün n - 2 serbestlik dereceli bir t dağılımına sahip olmasını bekliyoruz. **T dağılımı çan şeklindedir ve yaklaşık 30'dan büyük n değerleri için normal dağılıma oldukça benzerdir.** Sonuç olarak, **β1 = 0** varsayılarak, **|t|**'e eşit veya daha büyük herhangi bir sayıyı mutlak değerde gözlemleme olasılığını hesaplamak basit bir meseledir. Bu olasılığı **p-değeri (p-value)** olarak adlandırıyoruz.

Kabaca p değerini şu şekilde yorumluyoruz: küçük bir p-değeri, tahminci ile yanıt arasında herhangi bir gerçek ilişki olmadığında, tahminci ile şansa bağlı yanıt arasında bu kadar önemli bir ilişki gözlemlemenin olası olmadığını gösterir.

Dolayısıyla, küçük bir p değeri görürsek, tahmin edici ile yanıt arasında bir ilişki olduğu sonucuna varabiliriz.

Eğer p değeri yeterince küçükse, yokluk hipotezini reddederiz - yani X ve Y arasında bir ilişki olduğunu beyan ederiz Yokluk hipotezini reddetmek için tipik p-değeri kesintileri % 1 veya 5'tir. N = 30 olduğunda, bunlar sırasıyla 2 ve 2.75 civarında t-istatistiklerine (3.14) karşılık gelir.

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

TABLE 3.1 Reklam verileri için, TV reklam bütçesinde satılan birim sayısının regresyonu için en küçük kareler modelinin katsayıları. TV reklam bütçesindeki 1000 $ 'lık artış, satışlarda yaklaşık 50 birim artışla ilişkilidir (Satış değişkeninin binlerce birim olduğunu ve TV değişkeninin binlerce dolar olduğunu hatırlayın).

Tablo 3.1, Reklam verileri için TV reklam bütçesinde satılan birim sayısının regresyonu için en küçük kareler modelinin ayrıntılarını vermektedir. Bˆ0 ve βˆ1 için katsayıların standart hatalarına göre çok büyük olduğuna dikkat edin, bu nedenle t-istatistikleri de büyüktür; H0 doğruysa bu tür değerleri görme olasılıkları neredeyse sıfırdır.

Dolayısıyla şu sonuca varabiliriz: **β0! = 0** ve **β1! = 0.**

3.1.3

Assessing the Accuracy of the Model

(Modelin Doğruluğunun Değerlendirilmesi)

Yokluk hipotezini (3.12) alternatif hipotez (3.13) lehine reddettiğimizde, modelin verilere ne ölçüde uyduğunu ölçmek doğaldır. Doğrusal bir regresyon uyumunun kalitesi, tipik olarak iki ilgili büyüklük kullanılarak değerlendirilir: artık standart hata (residual Standard error) **(RSE)** ve **R2** (R kare) istatistiği.

Tablo 3.2, TV reklam bütçesinde satılan birim sayısının doğrusal regresyonu için RSE, R2 istatistiği ve F-istatistiğini (Bölüm 3.2.2'de açıklanacaktır) göstermektedir.

Residual Standard Error

(3.5)teki modelden hatırlanacağı üzere bir hata terimi **e** 'dir. Bu hata terimlerinin varlığından dolayı, gerçek regresyon çizgisini bilsek bile (yani β0 ve β 1 bilinse bile), X'ten Y'yi mükemmel bir şekilde tahmin edemeyiz. RSE, **e** 'nin standart sapmasının bir tahminidir.

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

TABLE 3.2 Reklam verileri için, TV reklam bütçesinde satılan birim sayısının regresyonu için en küçük kareler modeli hakkında daha fazla bilgi. Kabaca, yanıtın gerçek regresyon çizgisinden sapacağı ortalama miktardır. Aşağıdaki formül kullanılarak hesaplanır:

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \qquad (3.15)$$

RSS'nin Bölüm 3.1.1'de tanımlandığını ve aşağıdaki formülle verildiğini unutmayın.

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \qquad (3.16)$$

Reklam verileri söz konusu olduğunda, Tablo 3.2'deki doğrusal regresyon çıktısından RSE'nin **3.26** olduğunu görüyoruz. Diğer bir deyişle, her bir pazardaki gerçek satışlar, gerçek regresyon çizgisinden ortalama olarak yaklaşık **3.260** birim sapmaktadır.

Bunu düşünmenin başka bir yolu da, model doğruydu ve bilinmeyen katsayılar B0 ve B1'in gerçek değerleri tam olarak biliniyordu, TV reklamcılığına dayalı herhangi bir satış tahmini yine de ortalama **3,260** birim kadar kapalı olacaktır.

Tabii ki, **3.260** birimin kabul edilebilir bir tahmin hatası olup olmadığı, problemin içeriğine bağlıdır. Reklam veri setinde, tüm pazarlarda satışların ortalama değeri yaklaşık **14.000** birimdir ve bu nedenle yüzde hatası **3.260 / 14.000 =% 23**'tür.

RSE, modelin (3.5) verilere uyumsuzluğunun bir ölçüsü olarak kabul edilir.

Model kullanılarak elde edilen tahminler gerçek sonuç değerlerine çok yakınsa, yani **i = 1, ..., n** için **ˆyi ≈ yi** ise, o zaman (3.15) küçük olduğunu ve modelin verilere çok iyi uyduğu sonucuna varabiliriz.

Öte yandan, bir veya daha fazla gözlem için **ˆyi yi**'den çok uzaksa, **RSE** oldukça büyük olabilir ve bu da modelin verilere tam olarak uymadığını gösterir.

R^2 Statistic

RSE, modelin (3.5) verilere uyumsuzluğunun mutlak bir ölçüsünü sağlar. Ancak Y birimleriyle ölçüldüğünden, iyi bir RSE'yi neyin oluşturduğu her zaman net değildir. R^2 istatistiği, alternatif bir uyum ölçüsü sağlar. **Bir oran biçimini** (açıklanan varyans oranı) alır ve bu nedenle her zaman **0 ile 1** arasında bir değer alır ve **Y ölçeğinden bağımsızdır**.

R^2'yi hesaplamak için aşağıdaki formülü kullanıyoruz :

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \qquad (3.17)$$

burada $\text{TSS} = \sum(y_i - \bar{y})^2$ A, karelerin toplamıdır ve RSS (3.16) 'da toplamı olarak tanımlanır. **TSS, Y yanıtındaki toplam varyansı ölçer ve kareler, regresyon gerçekleştirilmeden önce yanıta özgü değişkenlik miktarı olarak düşünülebilir**.

Bunun aksine, RSS regresyonu gerçekleştirdikten sonra açıklanamayan kalan değişkenlik miktarını ölçer. Dolayısıyla, TSS - RSS regresyonu gerçekleştirerek açıklanan veya kaldırılan yanıttaki değişkenlik miktarını ölçer ve R^2, X kullanılarak açıklanabilen Y'deki değişkenlik oranını ölçer.

1'e yakın bir R^2 istatistiği, tepkideki değişkenliğin büyük bir oranının regresyon ile açıklandığını gösterir. 0'a yakın bir sayı, regresyonun yanıttaki değişkenliğin çoğunu açıklamadığını gösterir; bu, doğrusal modelin yanlış olması veya doğal hata σ2'nin yüksek olması veya her ikisinin birden olması nedeniyle meydana gelebilir.

Tablo 3.2'de R^2 0.61 idi ve bu nedenle satışlardaki değişkenliğin üçte ikisinden biraz azı TV'deki doğrusal bir gerileme ile açıklanıyor.

R^2 istatistiği (3.17), RSE'ye (3.15) göre yorumlama avantajına sahiptir, çünkü RSE'den farklı olarak, her zaman 0 ile 1 arasındadır. Bununla birlikte, iyi bir R^2 değerinin ne olduğunu belirlemek yine de zor olabilir ve genel olarak bu, uygulamaya bağlı olacaktır. Örneğin, fizikteki bazı problemlerde, verilerin gerçekten küçük bir artık hata ile doğrusal bir modelden geldiğini bilebiliriz. Bu durumda, 1'e oldukça yakın bir R2 değeri görmeyi bekleriz ve önemli ölçüde daha küçük bir R2 değeri, verilerin üretildiği deneyde ciddi bir sorun olduğunu gösterebilir.

Öte yandan, biyoloji, psikoloji, pazarlama ve diğer alanlardaki tipik uygulamalarda, doğrusal model (3.5), verilere en iyi ihtimalle son derece kabaca bir yaklaşımdır ve diğer ölçülmemiş faktörlerden kaynaklanan artık hatalar genellikle çok büyüktür. Bu ortamda, yanıttaki varyansın çok küçük bir oranının tahminci tarafından açıklanmasını beklerdik ve 0.1'in çok altında bir R^2 değeri daha gerçekçi olabilir!

R2 istatistiği, X ve Y arasındaki doğrusal ilişkinin bir ölçüsüdür.

Aşağıda tanımlanan korelasyonun, X ve Y arasındaki doğrusal ilişkinin bir ölçüsü olduğunu hatırlayın :

$$\mathrm{Cor}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}, \qquad (3.18)$$

Bu, doğrusal modelin uyumunu değerlendirmek için R^2 yerine **r = Cor (X, Y)** kullanabileceğimizi göstermektedir. Aslında basit doğrusal regresyon ayarında R^2 = r2 olduğu gösterilebilir. Başka bir deyişle, kare korelasyon ve R^2 istatistiği aynıdır.

Bununla birlikte, bir sonraki bölümde yanıtı tahmin etmek için aynı anda birkaç öngörücü kullandığımız çoklu doğrusal regresyon problemini tartışacağız.

Yordayıcılar ve yanıt arasındaki korelasyon kavramı, otomatik olarak bu ayara kadar uzanmaz, çünkü korelasyon, daha fazla sayıda değişken arasındaki ilişkiden ziyade tek bir değişken çifti arasındaki ilişkiyi nicelleştirir. R2'nin bu rolü doldurduğunu göreceğiz.