📊 Çoklu Regresyon

📈 Basit Doğrusal Regresyon

Dr. Kubra Atalay Kabasakal

Bahar 2023

Basit Regresyon

- Bilindiği üzere, t-testi, varyans analizi gibi ortalama farkları ile ilgili **hipotez testleri değişkenler arasındaki ilişkiye dair herhangi bir bilgi vermemektedir.**
- Oysa serpilme diyagramlarına bakıldığında değişkenler arasında bir **ilişki** olabileceği hissedilebilmekte fakat bu tür analizlerle bu **ilişkiler ortaya koyulamamaktadır.**
- Dolayısıyla değişkenler arasındaki **ilişkinin şeklini, yönünü ve kuvvetini** belirleyebilmemiz için yeni metotlara ihtiyaç vardır. Bu metotlar ise genel olarak **regresyon (eğri uydurma) ve korelasyon analizi** olarak adlandırılır.

Regresyon Kullanım Alanları

- Tarımda belli ürünlerin verimi etkileyen toprak türü, tohum, sulama v.b. faktörlerin saptanması ve bunlar yardımıyla belli şartlarda alınacak ürün miktarının kestirilmesi tarımın önemli konusudur.

- Bir değişkenin değerlerinin ilgili başka değişkenler yardımıyla kestirilmesi, günlük yaşamımızın, ticaretin ekonominin, doğa ve sosyal bilimlerin önemli konularını içendedir.

- günlük yaşamımızın, ticaretin ekonominin, doğa ve sosyal bilimlerin pek çok alanındaki çalışmalarda iki ya da daha çok değişken arasında fonksiyonel ilişkiler vardır. **Bu ilişkiler matematiksel bir denklem yazılabilir.**

Örneğin taksi hizmeti ödenen $\ddot{\mathbf{u}}cret = a + bx$
a: sabit (taksimetre açılış ücreti)
b: her kilometrede artan ücret

Regresyon Kullanım Alanları

- Regresyon çözümlemenin temel amacı; **bağımlı değişken ile bağımsız değişken(ler) arasındaki ilişkiyi matematiksel modelle açıklayarak bağlantılar bulmak ve bağımsız değişken(ler) yardımıyla bağımlı değişkenli kestirmek şeklinde özetlenebilir.**

- Sosyal bilimlerde değişkenler arasındaki ilişkiler bir dereceye kadar fonksiyoneldir. (taksimetre örneği kadar net değildir!) Bu ilişkiye **probabilisitik** ilişki denir.

Regresyon Kullanım Alanları

- Sosyal bilimlerde değişkenler arasındaki ilişkilerin **matematiksel olarak kesin ifadelerle yazılamaması,** bu **değişkenlere ait önceki bilgiler yardımıyla elde edilmesi ve matematiksel ifadelerin bu bilgilere dayanılarak yazılması yolunu açmıştır.**

- Regresyon terimi 19. yüzyılda İngiliz istatistikçisi **Francis Galton** tarafından bir biyolojik inceleme için ortaya atılmıştır. Bu incelemenin ana konusu kalıtım olup, aile içinde baba ve annenin boyu ile çocukların boyu arasındaki bağlantıyı araştırmakta ve çocukların boylarının bir nesil içinde eski ata nesillerinin ortalamasına geri döndüklerini yani bir nesil içinde ortalamaya geri dönüş olduğu inceleme konusudur.

Basit Doğrusal Regresyon

- Bir bağımsız X değişkeninin değerlerinden ona bağlı değişkeninin değerlerinin kestirilmesini sağlayan **denkleme Y'in X'e göre regresyonu** denir.

$$Y = bx + a$$

- Regresyon denkleminde
- b doğrunun eğimidir => X'in 1 puanlık değişimine karşılık Y'nin ne kadar değişeceğini belirtir. (buna **regresyon katsayısı** denir)
- a ise Y kesişim noktasıdır => X sıfıra eşit olduğunda Y'nin alacağı değerdir (buna **regresyon sabiti** denir)

Basit Doğrusal Regresyon Uygulama

- **Lise matematik puanlarından** yararlanarak **üniversite genel matematik puanlarını** kestirme amacıyla üniversite genel matematik dersini alan öğrencilerden uygun bir örneklem alınmıştır.

```
lise_not <- c(18,35,53,24,64,58,32,39,64,82,32,49,48,70,57)
uni_not  <- c(33,46,47,21,73,55,74,32,56,68,43,46,68,84,61)
veri <- data.frame(lise_not, uni_not)
```

Basit Doğrusal Regresyon Uygulama

- Regresyon analizi yapmadan önce **saçılım diagramı** incelenmelidir. Puanlar saçılım grafiğinde **tek bir doğru oluşturmamaktadır.** Ancak **doğru oluşturma eğilimleri vardır.**

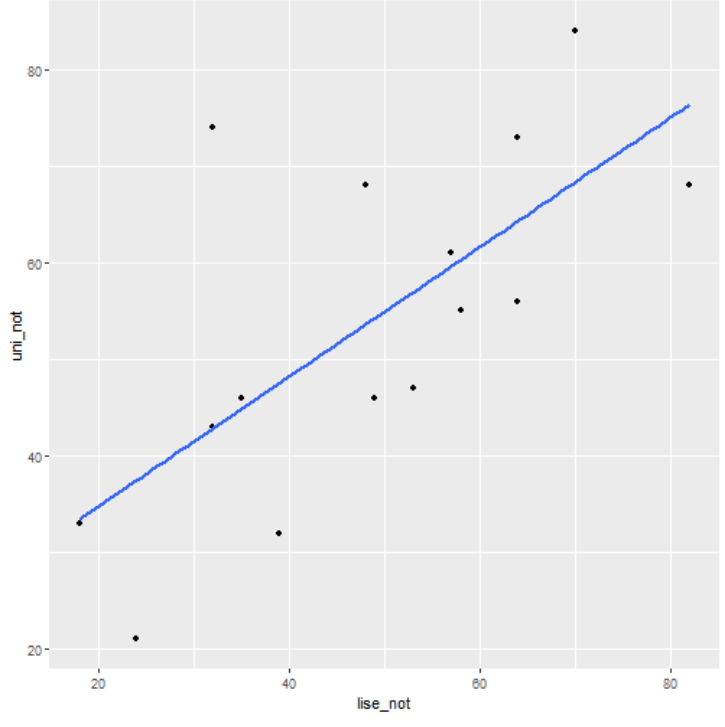
- Noktalardan olabildiğince yakın geçecek bir doğru çizilebilirse bu doğrudan yararlanarak X puanı bilinen öğrencilerin Y puanları kestirilebilir.

Basit Doğrusal Regresyon Uygulama

```
veri
```

```
##    lise_not uni_not
## 1        18       33
## 2        35       46
## 3        53       47
## 4        24       21
## 5        64       73
## 6        58       55
## 7        32       74
## 8        39       32
## 9        64       56
## 10       82       68
## 11       32       43
## 12       49       46
## 13       48       68
## 14       70       84
## 15       57       61
```

```
ggplot2::ggplot(veri,
aes(x = lise_not, y = uni_not)) +
geom_point() +
geom_smooth(method = "lm", se = F)
```



Basit Doğrusal Regresyon Uygulama

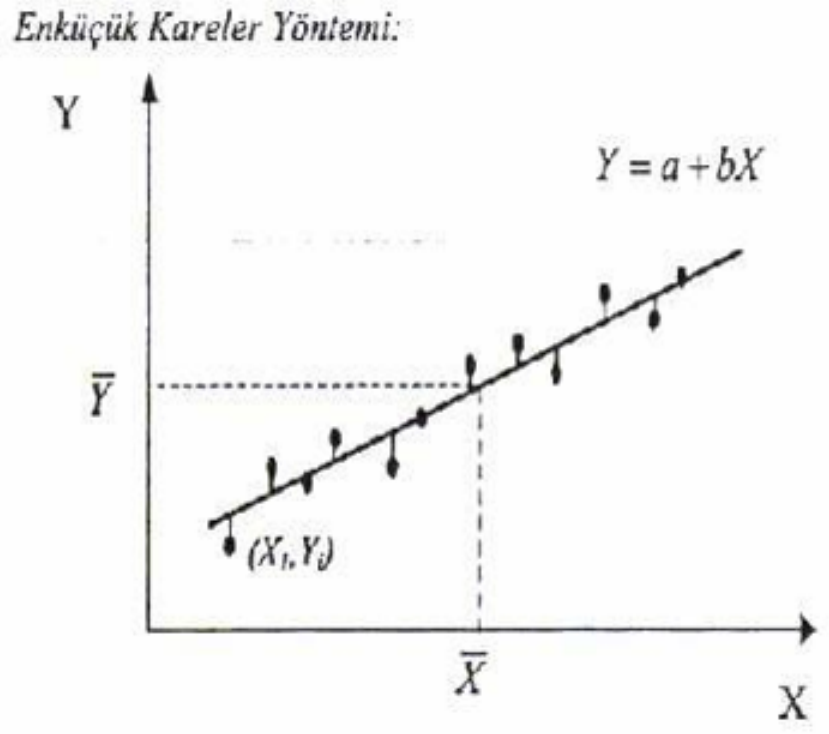
```
basitreg <- lm(uni_not ~ lise_not , veri)
summary(basitreg)
```

```
##
## Call:
## lm(formula = uni_not ~ lise_not, data = veri)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.475  -8.349  -0.449   5.037  31.158
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   21.373     10.196    2.10   0.0562 .
## lise_not       0.671      0.198    3.38   0.0049 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.4 on 13 degrees of freedom
## Multiple R-squared:  0.468,    Adjusted R-squared:  0.427
## F-statistic: 11.4 on 1 and 13 DF,  p-value: 0.0049
```

En küçük kareler yöntemi

- Bu yönteme göre **a** ve **b** öyle bir belirlenmelidir ki dağılımdaki noktaların, doğrunun etrafındaki değişkenliği en aza indirgenmiş olmalıdır.

- Regresyon doğrusu, **noktalar ile regresyon doğrusu arasındaki sapmaların kareler toplamı en az olacak şekilde,** saçılım grafiğindeki noktalar kümesine en uygun yere çizildiğinden bu ölçüte **en küçük kareler ölçütü** adı verilir.



En küçük kareler yöntemi

- Y değeri ve regresyon doğrusundaki Y' arasındaki farkın en küçük olacak şekilde yerleştirilir.
- $\sum(Y - Y')^2$ en küçük olacak şekilde yerleştirir.
- $b_{yx} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$
- $a_{yx} = \frac{n\sum Y - b_{YX}\sum X}{n}$

En küçük kareler yöntemi

- b_{yx} hesaplama
- $b_{yx} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$

```
n <- length(lise_not)
byx = (n*sum(lise_not*uni_not)-sum(lise_not)*sum(uni_not))/
  (n*sum(lise_not^2) - sum(lise_not)^2);byx
```

```
## [1] 0.671
```

- Regresyon doğrusunun eğimi, değişkenlerin standart sapmalarının oranlarıyla bunlar arasındaki korelasyonun çarpımına eşittir.

```
(sd(uni_not)/sd(lise_not))*cor(lise_not,uni_not)
```

```
## [1] 0.671
```

En küçük kareler yöntemi

- a_{yx} hesaplama
- $a_{yx} = \frac{n \sum Y - b_{yx} \sum X}{n}$

```
attach(veri)
ayx = (sum(uni_not) - byx*sum(lise_not))/15
ayx
```

```
## [1] 21.4
```

Kestirimin Standart Hatası

- Kestirim sonunda Y **değişkeninin gözlenen değerleri** ile **regresyon değerleri** Y' arasında fark olmaması veya bu farkın olabildiği kadar küçük olması istenir.

- **Gözlenen** Y ve **kestirilen** Y' değerleri arasındaki farklar **kestirimdeki hatalardır.** Bu farkların karelerinin ortalamasının kare köküne **kestirimin standart hatası** adı verilir.

- $$S_{yx} = \sqrt{\sum \frac{(Y - Y')^2}{n - 2}}$$

- $$S_{yx} = \sqrt{\frac{\sum Y^2 - a\sum Y - b\sum XY}{n - 2}}$$

Kestirimin Standart Hatası

- Ortak dağılımın için **kestirimin standart hatası** tek değişkenli dağılımın standart sapmasına benzer.

- Standart sapma tek değişkenli dağılımın ortalamadan farkının standart bir ölçüsü olduğu gibi, **kestirimin standart hatası da noktaların standart regresyon çizgisinden farkının ölçüsüdür.**

- Bu nedenle kestirimin standart hatası **verilen X değeri için kestirilen Y değerinin standart sapması** şeklinde okunabilen S_{yx} sembolü ile gösterilir.

Kestirimin Standart Hatası

X değerlerinden kestirlen Y' 'lerin standart hatası

```
sqrt((sum(uni_not^2)-ayx*sum(uni_not)-
       byx*(sum(uni_not*lise_not)))/13)
```

```
## [1] 13.4
```

```
res <- basitreg$residuals
sd(res)
```

```
## [1] 13
```

```
sqrt(sum((res - mean(res)) ^ 2 / (length(res)-2)))
```

```
## [1] 13.4
```

Basit Doğrusal Regresyon Uygulama

```
basitreg <- lm(uni_not ~ lise_not , veri)
library(broom)
glance(basitreg)
```

```
## # A tibble: 1 × 12
##   r.squared adj.r.s…¹ sigma stati…² p.value    df logLik   AIC   BIC devia…³
##       <dbl>     <dbl> <dbl>   <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1     0.468     0.427  13.4    11.4 0.00490     1  -59.2  124.  127.   2350.
## # … with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
## #   variable names ¹adj.r.squared, ²statistic, ³deviance
```

- R İki değişken arasında pearson korelasyon katsayısı
- **R-Square:** Determinasyon katsayısı/bağımsız değişkenin bağımlı değişken üzerindeki açıklama oranı
- **Adjusted R Square**: Düzeltmiş determinasyon katsayısı, şans eseri açıklanan değişimin neden olduğu hatanın arındırılmış hali.
- **Standart Kestirimin Hatası:** Hata teriminin standart sapmasıdır.

Basit Doğrusal Regresyon Uygulama

```
basitreg <- lm(uni_not ~ lise_not , veri)
library(broom)
glance(basitreg) %>% kable()
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.468	0.427	13.4	11.4	0.005	1	-59.2	124	127	2350	13	15

- Tablodaki p değeri regresyon modelindeki yordanan ve yordayan değişkenler arasındaki ilişki için hesaplanan değerin anlamlı olup olmadığını göstermektedir.

Basit Doğrusal Regresyon Uygulama

```
glance(basitreg)[,c(1,2,4,6,5)]
```

```
## # A tibble: 1 × 5
##   r.squared adj.r.squared statistic    df p.value
##       <dbl>         <dbl>     <dbl> <dbl>   <dbl>
## 1     0.468         0.427      11.4     1 0.00490
```

- Yani regresyon modelinde lise matematik puanları ile genel matematik puanları arasında doğrusal ilişki anlamlı düzeydedir. Regresyon modelindeki $\mathbf{df}$ 1 olması nedeni,regresyon modelindeki sabit ve eğimi katsayı olarak almasıdır. 2-1

Basit Doğrusal Regresyon Uygulama

```
tidy(basitreg)
```

```
## # A tibble: 2 × 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)   21.4      10.2       2.10 0.0562
## 2 lise_not       0.671     0.198     3.38 0.00490
```

- p değerleri sabitin ve yordayıcı değişkenin katsayısının anlamlılık testi sonuçları

teşekkürler