R ile Uygulamalı Ekonometri

Uraz Akgül

12/6/2022

Table of contents

Önsöz

R ile Uygulamalı Ekonometri isimli kitabım ile üniversite öğrencilerinden sektörde çalışan profesyonellere kadar geniş bir kitleye ulaşmayı hedefliyorum.

Kitabın konu başlıkları ile içeriğini zamanla kendini tamamlayacak ve gerekirse güncelleyecek şekilde tasarladım. Süreç ile ilgili bilgiler bir sonraki sayfada verilecektir.

Kitabın hazırlanması aşamasında ana kaynağım *Damodar Gujarati*'nin **Örneklerle Ekonometri (Econometrics by Example)** kitabı olacaktır. Bu kitabın yanında faydalandığım diğer birçok kaynağa *Yararlandığım Kaynaklar* bölümünden ulaşabilirsiniz.

R programlama dilini indirdikten sonra uzun bir süre kullanıp deneyim kazandığım **RStudio** veya geçiş yaptığım **Visual Studio Code** IDE'si ile ilerlemenizi ve bu dil ve tercih ettiğiniz IDE hakkında kendinizi yeterli hissedecek kadar bilgi sahibi olmanızı tavsiye ederim. Kitapta R programlama dili hakkında herhangi bir konu anlatımı olmayacaktır.

Ekonometri konularını akademik bir anlatımdan tam değil fakat olabildiğince uzak tutmaya çalışacağım ancak bu ezbere dayalı bir yol ile öğreneceğiniz anlamına gelmeyecektir. Her konu hakkında temel bir bilginiz olacak ve mutlaka en az bir uygulama yapmış olacaksınız.

Öğrenmek dinamik bir süreçtir. Bu kitap temel bilgileri verecek olmak ile beraber kendini tamamlama aşamasında ilke edindiği aşağıdaki söz ile ilerleyecektir.

> *İlim ve fennin yaşadığımız her dakikadaki safhalarının gelişmesini kavramak ve izlemek şarttır.* -Mustafa Kemal ATATÜRK

Süreç

06/12/2022:

- Projenin ilk paylaşımı yapıldı.

Veri Setleri

Kitap boyunca kullanılacak veri setlerine buradan ulaşabilirsiniz.

Kütüphaneler

```
library(readxl)
library(tidyverse)
library(magrittr)
library(HistData)
```

1 Doğrusal Regresyon Modeli

1.1 Genel

Temellerin atıldığı bir konu olduğu için bu bölümün çok önemli olduğunu düşünüyorum. William H. Greene, *"ekonometricinin alet çantasındaki en kullanışlı tek araç doğrusal regresyon modelidir."* der. Doğrusal Regresyon Modeli (DRM) ile ilgili herhangi bir soru işaretinizin kalmadığına emin olmalısınız.

DRM'yi şöyle yazalım:

$Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i$

Y : Bağımlı ya da açıklanan değişken.

X : Bağımsız ya da açıklayıcı değişken.

ϵ : Rassal veya stokastik hata terimi. Çeşitli nedenler ile modele konulamayan değişkenleri içerir. Stokastik sözcüğü hedef ya da hedefin göbeği anlamına gelir. Peter Kennedy, stokastik ve hata terimini, *"Stokastik bir ilişki, okun nadiren hedefi tam on ikiden vurması gibi, bağımlı değişken değerinin tam olarak öngörülmesi anlamında, her zaman hedefi vuramaz. Hata terimi açıkça bu hedeften sapmaların ya da hataların büyüklüklerini belirlemek amacı ile kullanılır."* örneği ile açıklar. Hata terimi gibi bir de kalıntı elde edeceğiz. İkisi aynı anlama gelmek ile beraber örneklem söz konusu olduğu zaman kullanılan ifade kalıntı olacaktır. Kennedy, hata terimi için *"ekonometristlerin kullandığı tahmin yöntemlerinin başarısı büyük bir oranda hata teriminin yapısına bağlıdır."* der.

β_1 : Kesme terimi ya sa sabit terim. Tüm X'ler sıfıra eşitlendiğinde Y'nin ortalama değerini gösterir deriz ancak daha net bir ifadeyle modelde bulunmayan (aşağıda açıklanan hata terimine bakın) bütün değişkenlerin Y üzerindeki ortalama etkisidir.

$\beta_2, ..., \beta_k$: Kısmi eğim parametreleri ya da kısmi regresyon parametreleri.

$\beta_1, \beta_2, ..., \beta_k$: Regresyon parametreleri.

i : i-nci gözlem.

DRM'nin gösterimini aşağıdaki gibi sadeleştirmek mümkündür.

$Y_i = \beta X + \epsilon_i$

Yukarıdaki eşitliklere anakütle modeli denir. Bu model, deterministik bileşen βX ile rassal bileşen ϵ_i'nin birleşimidir. βX için X değerleri verildiğinde Y_i'nin koşullu ortalaması $E(Y_i|X)$ diyebiliriz.

Regresyon analizinde öncelikli amacımız X değişkenlerinin değerlerindeki değişimlere Y'nin verdiği ortalama tepkiyi ölçmektir. Bu noktada eğim parametrelerinden bahsedebiliriz. Eğim parametresi, diğer açıklayıcı değişkenlerin değerleri sabit tutulduğunda bir açıklayıcı değişken değerindeki bir birim değişim karşılığında Y'nin ortalama değerindeki değişimi ölçer. Yeri gelmişken kısmi eğim parametreleri ya da kısmi regresyon parametrelerini açıklayalım. Buradaki kısmilik şuradan gelir: Bir açıklayıcı değişkendeki bir birimlik değişimin Y'nin ortalaması üzerindeki doğrudan ya da net etkisi, diğer açıklayıcı değişkenlerin Y'nin ortalaması üzerinde olabilecek etkisinden arındırılarak ölçülür. Bu nedenle kısmi kavramı kullanılır.

1.2 Sıradan En Küçük Kareler (OLS)

Bu bölümün sonunda bir ücret regresyonu kuracağız. Buna geçmeden önce yukarıda yazdığımız aşağıdaki eşitliğin nasıl tahmin edileceğine bakalım.

$Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i$

Regresyon parametrelerini tahmin etmede sıradan en küçük kareler (OLS, Ordinary Least Squares) ciddi bir kullanımı olan bir yöntemdir.

ϵ_i dediğimiz hata terimi, gerçek Y değerleri ile regresyon modelinden elde edilen Y değerleri arasındaki farktır.

$\epsilon_i = Y_i - (\beta_1 + \beta_2 X_{2i} + ... + \beta_k X_{ki})$

$\epsilon_i = Y_i - \beta X$

OLS, bu hata teriminin kareler toplamını minimize eder.

$\sum \epsilon_i^2 = \sum (\beta_1 - \beta_2 X_{2i} - ... - \beta_k X_{ki})^2$

Hata kareler toplamını mümkün olduğunca küçük yapan β parametresi değerleri bulunmalıdır. Tahmin edilen β parametrelerini θ ile gösterelim.

$\hat{Y}_i = \theta_1 + \theta_2 X_{2i} + ... + \theta_k X_{ki} + e_i$

Yukarıdaki eşitliği aşağıdaki gibi sadeleştirebiliriz.

$Y_i = \hat{Y}_i + e_i = \theta X + e_i$

Henüz gördüğümüz $\hat{Y}_i$, βX'in tahmincisidir. β parametrelerinin tahmincileri θ parametreleri; ϵ_i hata teriminin tahmincisi ise e_i'dir.

1.3 Klasik DRM Varsayımları

Klasik DRM'nin varsayımları model kurma sürecinin önemli bir parçası olacaktır. İlgili varsayımlar aşağıdaki gibidir.

- Regresyon modeli parametreler açısından doğrusaldır. Y ve X değişkenlerine göre ise doğrusallık aranmaz.

Parametreler açısından doğrusallık: Parametrelerin kuvveti alınmamış (β_2^2 gibi), parametreler diğer parametrelere bölünmemiş (β_2/β_3 gibi) veya dönüştürülmemiştir ($ln\beta_4$ gibi).

Değişkenler açısından: Koşul aranmaz. Örneğin, X değişkeninin doğal logaritması (lnX_2 gibi), tersi ($1/X_3$ gibi) veya kuvveti (X_2^3 gibi) alınmış olabilir.

- $cov(\epsilon_i, X) = 0$. Değerlerinin tekrarlanmış örneklemlerde sabit olmasına bağlı olarak, açıklayıcı değişkenlerin sabit olduğu veya stokastik olmadığı varsayılır. Bu varsayıma sabit X değerleri ya da hata teriminden bağımsız X değerleri diyebiliriz. Yani, her X değişkeni ile ϵ_i arasındaki ortak varyans sıfırdır.
- $E(\epsilon_i|X) = 0$. X değişkenlerinin değerleri verildiğinde hata teriminin beklenen ya da ortalama değeri sıfırdır. Bu durumda yazdığımız $Y_i = \beta X + \epsilon_i$ eşitliğini $E(Y_i|X) = \beta X + E(\epsilon_i|X) = \beta X$ şeklinde yazabiliriz.
- $var(\epsilon_i|X) = \sigma^2$. X değerleri verildiğinde her bir ϵ_i'nin varyansı sabittir (sabit varyans).
- $cov(\epsilon_i, \epsilon_j|X) = 0, i \neq j$. İki hata terimi arasında korelasyon (otokorelasyon) yoktur.
- X değişkenleri arasında tam doğrusal ilişki ya da çoklu doğrusal bağlantı yoktur.
- Regresyon modeli doğru tanımlanmış olup herhangi bir tanımlama yanlılığı ya da hatası yoktur.
- Gözlem sayısı tahmin edilecek anakütle parametrelerinden fazla olmalıdır. Daha sade bir anlatım ile gözlem sayısı açıklayıcı değişken sayısından büyük olmalıdır diyebiliriz.
- $\epsilon_i \sim N(0, \sigma^2)$. Hata terimi sıfır ortalamalı ve σ^2 (sabit) varyanslı normal dağılıma sahiptir.

Ayrıca OLS tahmincileri BLUE'dur. BLUE (Best Linear Unbiased Estimator), en iyi doğrusal yansız tahminci anlamına gelmektedir. Akılda tutmak için Doğrusal En iyi Sapmasız Tahmin Edici olan DESTE de kullanılabilir.

- Doğrusal tahminci: Tahminciler Y bağımlı değişkeninin doğrusal fonksiyonlarıdır.
- Yansız tahminci: Yöntemin tekrarlanan uygulamalarında tahminciler ortalama olarak gerçek değerlerine eşittir.
- En küçük varyansa sahip tahminci / etkin tahminci: Doğrusal yansız tahminciler sınıfı içinde OLS tahmincileri minimum varyansa sahiptir.

1.4 OLS Tahmincilerinin Standart Hataları

θ OLS tahmincileri, değerleri örnekleme bağlı olarak değiştiği için rassal değişkendir. Bu noktada değişkenliği ölçmemiz gerekmektedir. İstatistikte de bu değişkenlik varyans (σ^2) ve standart sapma (σ) ile ölçülür. Bir tahmincinin standart sapması regresyon bağlamında standart hatadır. Standart hata, tahmin edicinin örneklem dağılımının standart sapmasıdır. DRM'de u_i hata terimi varyansına (σ^2) ait tahmin şöyledir:

$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k}$

Burada, n örneklem büyüklüğü ilen k tahmin edilen regresyon parametre sayısıdır. $\sqrt{\hat{\sigma}^2}$ ya da $\hat{\sigma}$, regresyonun standart hatası ya da kök ortalama karedir.

1.5 Hipotez Testleri

1.5.1 t Testi

Anakütle regresyon parametresi için $\beta_k = 0$ hipotezini test edelim.

$H_0 : \beta_k = 0$

$H_1 : \beta_k \neq 0$

Bu hipotezin testi için ise şunu yazalım:

$t = \frac{\theta_k}{se(\theta_k)}$

Burada t testini kullanyoruz. $se(\theta_k)$, θ_k'nın standart hatasıdır. t değerinin (n-k) serbestlik derecesi vardır. Hesaplanan t değeri olasılığı düşük çıkarsa (%5 veya daha az gibi) $\beta_k = 0$ sıfır hipotezi reddedilir. Sıfır hipotezinin reddi ise t değerinin istatistiksel olarak anlamlı olduğu anlamına gelir. Daha geniş bir ifade ile diğer açıklayıcı değişkenler sabit iken incelenen değişkenin bağımlı değişken üzerinde istatistiksel olarak anlamlı bir etkisinin olduğu söylenebilir. Diyelim ki sıfır hipotezini kabul ettik. Burada edilen kabul, örneklem verilerine göre bu hipotezi reddedecek nedeni henüz bulamadığımızdandır. Yani, bu kabul net doğru anlamına gelmez. Özetle, kabul ederiz yerine reddedemeyiz demek yerinde olacaktır.

1.5.2 F Testi

t testi ile bireysel anlamlılığa bakıyorduk. Bütün eğim parametrelerinin aynı anda anlamlı olup olmadığına bakmak için ise F testini kullanacağız. Buna regresyonun genel anlamlılığı da diyebiliriz.

$H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0$

H_1 : En az bir $\beta_i \neq 0$

F istatistiği şöyledir:

$F = \frac{ESS/sd}{RSS/sd} = \frac{Ortalama\ ESS}{Ortalama\ RSS}$

$ESS = \sum(\hat{Y}_i - \bar{Y})^2$

$RSS = \sum(Y_i - \hat{Y}_i)$

$TSS = \sum(Y_i - \bar{Y})^2$

ESS, Y bağımlı değişkenindeki değişkenliğin model tarafından açıklanan kısmı iken, RSS, Y bağımlı değişkenindeki değişkenliğin model tarafından açıklanmayan kısmıdır. Y bağımlı değişkenindeki toplam değişkenlik ise TSS olup ESS ile RSS'in toplamıdır. sd, serbestlik derecesidir ve payın serbestlik derecesi k-1 iken, paydanın serbestlik derecesi n-k'dır. Yukarıdaki eşitliği tekrar yazalım.

$F = \frac{ESS}{TSS} = \frac{\sum(\hat{Y}_i - \bar{Y})/(k-1)}{\sum(Y_i - \hat{Y}_i)/(n-k)}$

Eğer hesaplayacağımız F değeri α seviyesindeki kritik F değerinden büyük ise sıfır hipotezi reddedilebilir ki bu da en az bir açıklayıcı değişkenin istatistiksel olarak anlamlı olduğu anlamına gelir.

1.6 Güven Aralığı

Bir tek nokta tahminine güvenmek yerine belli bir olasılıkla (örneğin %95) gerçek parametreyi içerecek şekilde bir aralık oluşturabiliriz. Herhangi bir anakütle parametresi β_k için $(1-\alpha)$ güven aralığı şöyledir:

$Pr[\theta_k \pm t_{\alpha/2} se(\theta_k)] = (1-\alpha)$

$[\theta_k - t_{\alpha/2} se(\theta_k)]$: Alt sınır

$[\theta_k + t_{\alpha/2} se(\theta_k)]$: Üst sınır

Güven aralığının genişliği tahmin edicinin güvenilirliği olan standart hatası ile orantılıdır. Dikkat etmemiz gereken nokta bu güven aralığı gerçek β_k'nın verilen alt ve üst sınırlar arasında yer alma olasılığının $(1-\alpha)$ olduğunu söylemez. Aksine gerçek β_k değerini sabit bir sayı kabul ederiz ve bu olasılık da ya 1'dir ya da 0. Asıl söylediği, her 100 aralığın 95'inde (güven katsayısının %95 olduğunu varsayalım) gerçek β_k'yı içerdiğidir.

1.7 Uyum İyiliği

Tahmin edilen regresyon doğrusunun uyum iyiliğinin ölçüsü R^2'dir ve şöyle hesaplanır:

$R^2 = \frac{ESS}{TSS}$ ya da $R^2 = 1 - \frac{RSS}{TSS}$

R^2 değeri 0 ile 1 arasında yer alır ($0 \leq R^2 \leq 1$) ve 1'e yaklaştıkça uyum iyileşir. Uyumun iyileşmesi açıklayıcılık gücünün arttığı anlamına gelir. Bu noktada modele açıklayıcı değişken ekledikçe R^2 değerinin artacağı bilinmelidir. Bu durumda düzeltilmiş R^2 ya da $\bar{R^2}$ kullanılabilir ve şöyle hesaplanır:

$\bar{R^2} = 1 - (1 - R^2)\frac{n-1}{n-k}$

Düzeltilmiş ile serbestlik derecesi düzeltmesi kastedilmektedir. İki nokta:

- $k > 1$ ise $\bar{R^2} < R^2$'dir.
- R^2 daima pozitif iken $\bar{R^2}$ negatif olabilir.

Yüksek $\bar{R^2}$ bulma yarışına bir not düşmek gerekir. Bizim asıl ilgimiz, açıklayıcı değişkenlerin bağımlı değişken ile olan mantıksal ilişkilerine ve onların istatistiksel anlamlılıklarına odaklı olmalıdır. Yüksek bir $\bar{R^2}$ bulamasak da bu modelin kötü olduğu anlamına gelmez.

1.8 Uygulama

Saatlik ücreti (dolar bazında) belirleyen faktörleri örneklemde araştırmak üzere, Mart 1995'te görüşme yapılan 1289 kişilik (anakütleden alınan örneklem) bir yatay-kesite bakalım. İlgili verilere *drm.xls* dosyası ile ulaşılabilir.

Bağımlı değişken:

- **wage:** Saatlik ücret ($)

Bağımsız değişkenler:

- **female:** Kadın ise 1; değilse 0
- **nonwhite:** Beyaz olmayan işçi ise 1; değilse 0
- **union:** Sendikalı bir işte ise 1; değilse 0
- **education:** Yıl bazlı eğitim
- **exper:** Yıl bazlı iş deneyimi. Yaş – eğitim süresi – 6 okula başlama yaşı

```
library(readxl)
library(tidyverse)
```

```r
library(magrittr)
```

```r
df <- read_excel("./data/drm.xls")

df %<>%
  select(wage, female, nonwhite, union, education, exper)
```

DRM'yi kurabiliriz.

```r
model <- lm(formula = wage ~ female + nonwhite + union + education + exper, data = df)
#ya da model <- lm(formula = wage ~., data = df)
summary(model)
```

```
Call:
lm(formula = wage ~ female + nonwhite + union + education + exper,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-20.781  -3.760  -1.044   2.418  50.414

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.18334    1.01579  -7.072 2.51e-12 ***
female      -3.07488    0.36462  -8.433  < 2e-16 ***
nonwhite    -1.56531    0.50919  -3.074  0.00216 **
union        1.09598    0.50608   2.166  0.03052 *
education    1.37030    0.06590  20.792  < 2e-16 ***
exper        0.16661    0.01605  10.382  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.508 on 1283 degrees of freedom
Multiple R-squared:  0.3233,    Adjusted R-squared:  0.3207
F-statistic: 122.6 on 5 and 1283 DF,  p-value: < 2.2e-16
```

- Regresyondaki açıklayıcı değişkenleri sıfıra eşitlediğimizde ortalama wage -$7.18 olur. Tabi bunun iktisadi açıdan bir anlamı yoktur. Ancak buna rağmen kesme terimini bırakmak faydalı olabilir.

Diğer değişkenler sabit tutulduğunda;

- Kadınların ortalama wage'i erkeklerin ortalama wage'inden $3.07 daha düşüktür (female).
- Beyaz olmayan bir işçinin ortalama wage'i beyaz bir işçinin ortalama wage'inden $1.56 daha düşüktür (nonwhite).
- Sendikalı bir işte çalışanın ortalama wage'i sendikalı bir işte çalışmayanın ortalama wage'inden $1.09 daha fazladır (union).
- Her ilave eğitim yılı için ortalama wage $1.37 artmaktadır (education).
- Her ilave deneyim için ortalama wage $0.16 artmaktadır (exper).

Diğer yorumlara bakalım.

- Bu model yardımı ile bir kişinin alacağı ücreti kesin olarak söyleyemeyiz. Sadece bu kişinin niteliklerine göre ne kazanabileceğini öngörebiliriz.
- p değeri küçüldükçe sıfır hipotezi aleyhindeki kanıtlar daha da güçlenir. Örneğin, yaklaşık 1.37 olan education parametresine ait değerin yaklaşık 20.79 olan bir t değeri hesaplandı. p değeri neredeyse sıfırdır (2e-16). Bu durumda education parametresi istatistiksel olarak oldukça anlamlıdır. Yani, education değişkeni wage değişkeninin önemli bir belirleyicisidir. %5 gibi bir p değeri aldığımızda tüm parametrelerin istatistiksel olarak anlamlı olduğunu görüyoruz. Yani tüm değişkenler wage'in önemli bir belirleyicisidir.
- R^2 değeri yaklaşık olarak 0.32'dir. Wage değişkenindeki değişkenliğin yaklaşık %32'si beş açıklayıcı değişken tarafından açıklanmaktadır.
- F değerine ait p değeri neredeyse sıfır (2.2e-16) olduğu için en az bir değişkenin wage değişkeni üzerinde anlamlı bir etkisi vardır. F değerini manuel hesaplayalım. Manuel hesaplarken F değerinin R^2 ile olan ilişkisini göreceğiz.

(k-1): Kesme terimi dışarıda tutulduğunda açıklayıcı değişken sayısı (5)

n: Gözlem sayısı 1289 ile kesme terimi dahil tahmin edilen parametre sayısı 6'nın farkı (1283)

$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} = \frac{0.3233/5}{(1-0.3233)/(1283)} = 122.6$

- Gerçek education parametresinin en iyi tek tahmini 1.37'dir ama bunu örneğin %95 güven aralığında da yorumlayabiliriz. Aşağıdan da görüleceği üzere, diğer şeyler sabit iken ek 1 yıllık eğitimin wage üzerindeki etkisinin minimum $1.24 ve maksimum $1.49 olduğu konusunda %95 güvendeyiz. Hatırlayalım: gerçek education parametresinin $1.32 olduğunu varsayalım. Bu durumda $1.32 bu aralıkta ya yer alır ya da yer almaz. Olasılık ya 1'dir ya da 0'dır. Aralık, her 100 aralığın 95'inde (güven katsayısının %95 olduğunu varsayalım) gerçek education parametresini içerir. %95 güven katsayısı ile hareket ettiğimizde %5'inde hatalı oluruz.

```r
confint(model, level = 0.95)
```

```
                2.5 %     97.5 %
(Intercept) -9.1761258 -5.1905507
female      -3.7901849 -2.3595660
nonwhite    -2.5642450 -0.5663817
union        0.1031443  2.0888072
education    1.2410091  1.4995928
exper        0.1351242  0.1980889
```

Aşağıdaki değerler ile beklenen ortalama ücreti bulalım.

- female: 1
- nonwhite: 1
- union: 0
- education: 12
- exper: 20

Kadın, beyaz olmayan, sendikası olmayan, 12 yıllık eğitime sahip, 20 yıllık iş deneyimi olan bir işçinin beklenen ortalama ücretine bakıyoruz.

```r
df2 <- data.frame(
    female = 1,
    nonwhite = 1,
    union = 0,
    education = 12,
    exper = 20
)

wage_pred <- predict(
    model, newdata = df2
)

paste0("$",round(wage_pred,digits=2))
```

```
[1] "$7.95"
```

2 Francis Galton ve Regresyon Terimi

Regresyon terimi ilk kez Francis Galton tarafından kullanılmıştır.

Saymak ve ölçmek Galton'ın hobisiymiş. Bırakın hobiyi saplantı da denilebilir.

Yapabildiğin her yerde say (Wherever you can, count). Kendisi hakkında yazılan çok şey var. Gerçekten de yapabildiği her yerde saymış.

Sokakta yürürken karşılaştığı kızları çekicilik derecelerine göre sınıflandırmış, kız alımlıysa sol cebinde taşıdığı kartı, sıradansa sağ cebindekini işaretlermiş. Böyle böyle İngiltere'nin güzellik haritasını çıkarmış ve Londralı kızlar en yüksek puanı; Aberdeenli kızlar ise son sırayı almış.

Kurduğu Galton Antropometrik (antropolojik ölçüm) Laboratuvarı'nda parmak izleri dahil insan vücuduyla ilgili mümkün olan her ölçümü yapmış, bu ölçümlerin yelpazesini ve karakterini izleyerek kaydını tutmuş. Parmak izleri Galton'ı büyülüyormuş. Nedeni ise vücudun diğer kısımlarından farklı olarak parmak izlerinin şeklinin kişi yaşlansa da hiçbir zaman değişmemesi. Galton bu konuda 200 sayfalık bir kitap yayınlamış ve bu çalışması kısa zamanda polisin parmak izini yaygın biçimde kullanmasına öncülük etmiş.

Galton, Britanya Bilimi İlerletme Birliği Başkanlığı'na seçilmesi sebebiyle bir konuşma yapar ve bu konuşma sırasında gerçekleştirdiği bir deneyde ortalamaya dönüşü destekleyen yeni kanıtlar bulduğunu açıklar. Bu deney için ise kendisine veri sağlayacak kişilere nakit ödeme yapacağını ilan eder ve insanlarla ilgili muazzam miktarda veri toplar: 205 ebeveynden doğmuş 928 yetişkin çocuk.

CROSS-TABULATION OF 928 ADULT CHILDREN BORN OF 205 MIDPARENTS, SORTED BY THEIR HEIGHT AND THEIR MIDPARENT'S HEIGHT

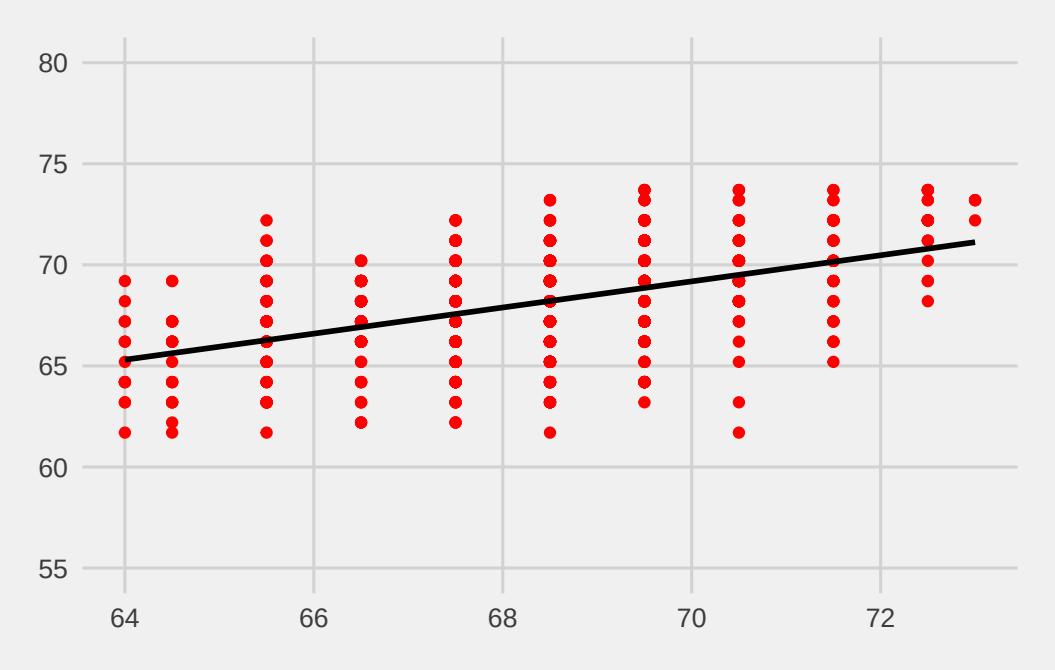
Height of Mid-parents (inches)	*Height of the Adult Child*														*Total No. of Adult Children*	*Total No. of Mid-parents*	*Medians*
	<61.7	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	>75.7			
>73.0	—	—	—	—	—	—	—	—	—	—	—	1	3	—	4	5	—
72.5	—	—	—	—	—	—	—	1	2	1	2	7	2	4	19	6	72.2
71.5	—	—	—	—	1	3	4	3	5	10	4	9	2	2	43	11	69.9
70.5	1	—	1	—	1	1	3	12	18	14	7	4	3	3	68	22	69.5
69.5	—	—	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68.9
68.5	1	—	7	11	16	25	31	34	48	21	18	4	3	—	219	49	68.2
67.5	—	3	5	14	15	36	38	28	38	19	11	4	—	—	211	33	67.6
66.5	—	3	3	5	2	17	17	14	13	4	—	—	—	—	78	20	67.2
65.5	1	—	9	5	7	11	11	7	7	5	2	1	—	—	66	12	66.7
64.5	1	1	4	4	1	5	5	—	2	—	—	—	—	—	23	5	65.8
<64.0	1	—	2	4	1	2	2	1	1	—	—	—	—	—	14	1	—
Totals	5	7	21	59	48	117	138	120	167	99	64	41	17	14	928	205	—
Medians	—	—	66.3	67.8	67.9	67.7	67.9	68.3	68.5	69.0	70.0	—	—	—	—	—	—

(From Francis Galton, 1886, "Regression Toward Mediocrity in Hereditary Stature," Journal of the Anthropological Institute, Vol. 15, pp. 246–263.)

R ile verilere ulaşabilmek mümkündür.

```
library(HistData)
library(tidyverse)
library(ggthemes)

Galton %>%
ggplot(aes(x=parent, y=child)) +
geom_point(color="red") +
geom_smooth(method = "lm", color="black", se=FALSE) +
scale_y_continuous(limits = c(55,80)) +
theme_minimal() +
theme_fivethirtyeight()
```



Gözlemleri inceleyebilmek için kadınlar ve erkekler arasındaki boy farklarıyla ilgili bir düzeltme yapar. Bunun için bütün annelerin boylarını 1.08 ile çarpar, ardından da anne ve babaların boylarını toplayıp ikiye böler. Elde ettiği birime de *orta ebeveyn boyu* adını verir. Tabi bu arada uzunlar uzunlarla, kısalar kısalarla evleniyor gibi bir eğilim var mı diye de hesaplamalar yapar ama böyle bir eğilimin bulunmadığını varsayacağı noktaya yeterince yakındır.

Tabloda... Sayıların sol alt köşeden sağ üst köşeye çarpraz bir yapı segilediğini görüyoruz. Yani, uzun boylu ebeveynlerin uzun boylu, kısa boylu ebeveynlerin de kısa boylu çocukları olduğunu gösteriyor: Kalıtım. Büyük sayıların ise tablonun ortasında toplandığı görülebilir. Bu ise her boy grubunun çocuklar arasında normal dağıldığını, aynı şekilde aynı ebeveynlerle ilgili her

boy grubundan her çocuk dizisinin de, normal bir dağılım gösterdiğini söyler.

Matematiksel analizin hükümranlığına ve muhteşem ruhuna hiç bu kadar derin bir bağlılık ve saygı duymamıştım (I never felt such a glow of loyalty and respect towards the sovereignty and magnificent sway of mathematical analysis) der Galton.

O bizi günlük yaşama, insanların soluk aldığı, terlediği, cinsel ilişkide bulunduğu ve geleceğinden endişelendiği dünyaya götürür. Artık daha önceki matematikçilerin teorilerini doğrulama aracı olarak kullandıkları kumar masalarından da, yıldızlardan da uzaklaşmış bulunuyoruz. Galton, teorileri bulduğu şekilde ele almış ve onları neyin önemli kıldığını keşfetmeye çalışmıştır.

Yararlandığım Kaynaklar

Örneklerle Ekonometri, D. Gujarati (Çeviren: N. Bolatoğlu)

Temel Ekonometri, D.N. Gujarati & D.C. Porter (Çeviren: Ü. Şenesen, G.G. Şenesen)

Ekonometri Kılavuzu, P. Kennedy (Çeviren: M. Sarımeşeli, Ş. Açıkgöz)

Ekonometrik Okuryazarlık, S. Güriş

Tanrılara Karşı-Riskin Olağanüstü Tarihi, P.L. Bernstein

Francis Galton: The man who drew up the 'ugly map' of Britain