R ile Uygulamalı Ekonometri

Uraz Akgül

12/27/2022

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.

11/12/2022:

- **Regresyon Modellerinin Fonksiyon Yapıları** konusu eklendi.

27/12/2022:

- *rmfk2.xlsx* dosyasındaki, TÜİK'ten elde edilen mevsim ve takvim etkisinden arındırılmış harcama yöntemiyle GSYH, zincirlenmiş hacim endeksi verileri yıllık ortalama değerler ile değiştirilmiştir.
- **Nitel Açıklayıcı Değişkenli Regresyon Modelleri** konusu eklendi.

Veri Setleri

Kitap boyunca kullanılacak veri setlerine buradan ulaşabilirsiniz.

Kütüphaneler

```
library(readxl)
library(tidyverse)
library(ggthemes)
library(magrittr)
library(HistData)
library(zoo)
library(segmented)
```

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.

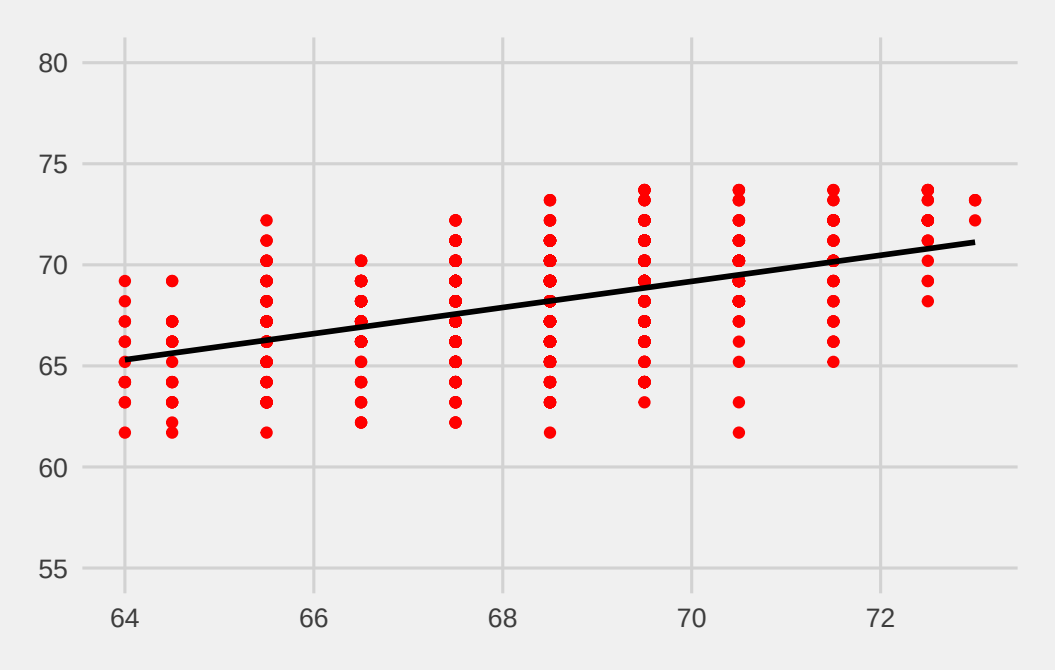
CROSS-TABULATION OF 928 ADULT CHILDREN BORN OF 205 MIDPARENTS, SORTED BY THEIR HEIGHT AND THEIR MIDPARENT'S HEIGHT

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.

3 Regresyon Modellerinin Fonksiyon Yapıları

Klasik Doğrusal Regresyon Modeli'nin varsayımlarından birisi şöyleydi:

- 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.

Bu başlıkta, parametrelere göre doğrusal olan ancak değişkenlere göre doğrusal olup olmamasının bir önemi olmayan modelleri ele alacağız.

3.1 Log-Log

Log-Log, hem bağımlı değişkenin hem de açıklayıcı değişkenlerin logaritmik yapıda olduğu bir fonksiyon kalıbı türüdür.

Üstel regresyon modelini inceleyelim.

$Y_i = \beta_1 X_i^{\beta_2} e^{\epsilon_i}$

Yukarıdaki model doğal logaritması alınarak şöyle yazılabilir:

$lnY_i = ln\beta_1 + \beta_2 lnX_i + \epsilon_i$

Son olarak aşağıdaki forma çevirebiliriz.

$lnY_i = \alpha + \beta_2 lnX_i + \epsilon_i$

Cobb-Douglas fonksiyonu ile devam edelim.

$Q_i = \beta_1 L_i^{\beta_2} K_i^{\beta_3}$

Q : Çıktı

L : Emek Girdisi

K : Sermaye

β_1 : Sabit

Yukarıdaki model parametreler açısından doğrusal değildir ki bu da doğrusal olmayan yöntemler ile ilerlenmesini gerektirir. Peki, (doğal) logaritmasını alırsak tam da varsayımda istediğimiz gibi parametreler açısından doğrusal olabilir mi?

$lnQ_i = ln\beta_1 + \beta_2 lnL_i + \beta_3 lnK_i$

$ln\beta_1$'e A dersek;

$lnQ_i = A + \beta_2 lnL_i + \beta_3 lnK_i$ olur. Yazılan eşitlik;

- Q, L ve K değişkenleri açısından doğrusal değildir
- A, β_2 ve β_3 parametrelerine göre doğrusaldır

Hata terimi ϵ_i de eklenirse doğrusal regresyon modelinin son hali elde edilir.

$lnQ_i = A + \beta_2 lnL_i + \beta_3 lnK_i + \epsilon_i$

3.1.1 Uygulama

ABD için Cobb-Douglas fonksiyonuna bakalım. İlgili verilere *rmfk1.xls* dosyası ile ulaşılabilir.

```
library(readxl)
library(tidyverse)
library(ggthemes)
library(magrittr)

df1 <- read_excel("./data/rmfk1.xls")

df1 %<>%
select(output,labor,capital)
```

Bağımlı değişken:

- **output:** ABD imalat sektörüne ait çıktı (katma değerle ölçülmüş, bin dolar)

Bağımsız değişkenler:

- **labor:** Emek girdisi (çalışma saati, bin saat)
- **capital:** Sermaye girdisi (sermaye harcaması, bin dolar)

Değişkenleri logaritmik yapıda yazalım.

```r
df1 %<>%
mutate(
    ln_output = log(output),
    ln_labor = log(labor),
    ln_capital = log(capital)
)
```

Model kurulabilir.

```r
loglog_model <- lm(ln_output ~ ln_labor + ln_capital, data = df1)
summary(loglog_model)
```

```
Call:
lm(formula = ln_output ~ ln_labor + ln_capital, data = df1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.45645 -0.12112 -0.05319  0.04518  1.21579

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.88760    0.39623   9.812 4.70e-13 ***
ln_labor     0.46833    0.09893   4.734 1.98e-05 ***
ln_capital   0.52128    0.09689   5.380 2.18e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2668 on 48 degrees of freedom
Multiple R-squared:  0.9642,    Adjusted R-squared:  0.9627
F-statistic: 645.9 on 2 and 48 DF,  p-value: < 2.2e-16
```

$Q_i = 48.79 L^{0.47} K^{0.52}$

Sabit değer 3.8876'nın ters logaritması yaklaşık olarak 48.79'dur.

```r
exp(as.numeric(coef(loglog_model)[1]))
```

```
[1] 48.79362
```

Yukarıdaki model çıktısının yorumları şöyledir:

- Çok küçük p değerlerine sahip regresyon parametreleri bireysel açıdan istatistiksel olarak anlamlıdır.
- F istatistiğine ait p değeri çok küçük olduğu için model istatistiksel olarak anlamlıdır.
- 0.96 gibi çok yüksek bir R^2 sevindirici olmak ile beraber bir soru işareti de yaratabilir.
- Log-Log modellerde eğim parametreleri esneklik olarak yorumlanabilir. Aynı zamanda ölçüm birimleri ortadan kalkar ve yüzde değişim altında değerlendirilirler.

Sermaye girdisi (ln_capital) sabit iken, emek girdisi (ln_labor) %1 artırılırsa imalat sektörüne ait çıktı ortalamada %0.47 artar.

Emek girdisi (ln_labor) sabit iken, sermaye girdisi (ln_capital) %1 artırılırsa imalat sektörüne ait çıktı ortalamada %0.52 artar.

3.2 Log-Lin

Bağımlı değişkenin logaritmik; diğer açıklayıcı değişkenlerin düzey ya da doğrusal yapıda olduğu bir fonksiyon kalıbı türüdür.

Bileşik faiz formülünü anımsayalım.

$Y_t = Y_0(1+r)^t$

r, Y'nin bileşik (zaman içindeki) büyüme hızıdır.

Eşitliğin logaritmasını alalım.

$lnY_t = lnY_0 + tln(1+r)$

lnY_0, β_1; $ln(1+r)$ ise β_2 olsun.

$lnY_t = \beta_1 + \beta_2 t + \epsilon_t$ olur.

Belli bir dönem (1998-2021 olsun) için reel GSYH'nin büyüme hızı ölçülmek isteniyorsa $RGDP_t = RGDP_{1998}(1+r)^t$ modeli kullanılabilir.

$RGDP$: Reel GSYH

r : Büyüme hızı

t : Kronolojik zaman

Her iki tarafın (doğal) logaritmasını alalım.

$lnRGDP_t = lnRGDP_{1998} + tln(1+r)$

β_1, $lnRGDP_{1998}$ ve β_2, $ln(1+r)$ olsun.

$lnRGDP_t = \beta_1 + \beta_2 t + \epsilon_t$ olur. Bu eşitlikte açıklayıcı değişken zamandır. Yani, t = 1,2,3,...,24.

3.2.1 Uygulama

Türkiye'nin reel GSYH büyüme hızına bakalım. Bu uygulamada, TÜİK'ten elde edilen mevsim ve takvim etkisinden arındırılmış harcama yöntemiyle GSYH, zincirlenmiş hacim endeksi kullanılacaktır. İlgili verilere *rmfk2.xlsx* dosyası ile ulaşılabilir.

Aşağıda açıklayıcı değişken olarak zaman değişkenini (t) ve RGDP'nin (doğal) logaritmasını ekledik.

```
df2 <- read_excel("./data/rmfk2.xlsx")

df2 %<>%
mutate(
    t = seq(1,nrow(.),1),
    ln_rgdp = log(rgdp)
)
```

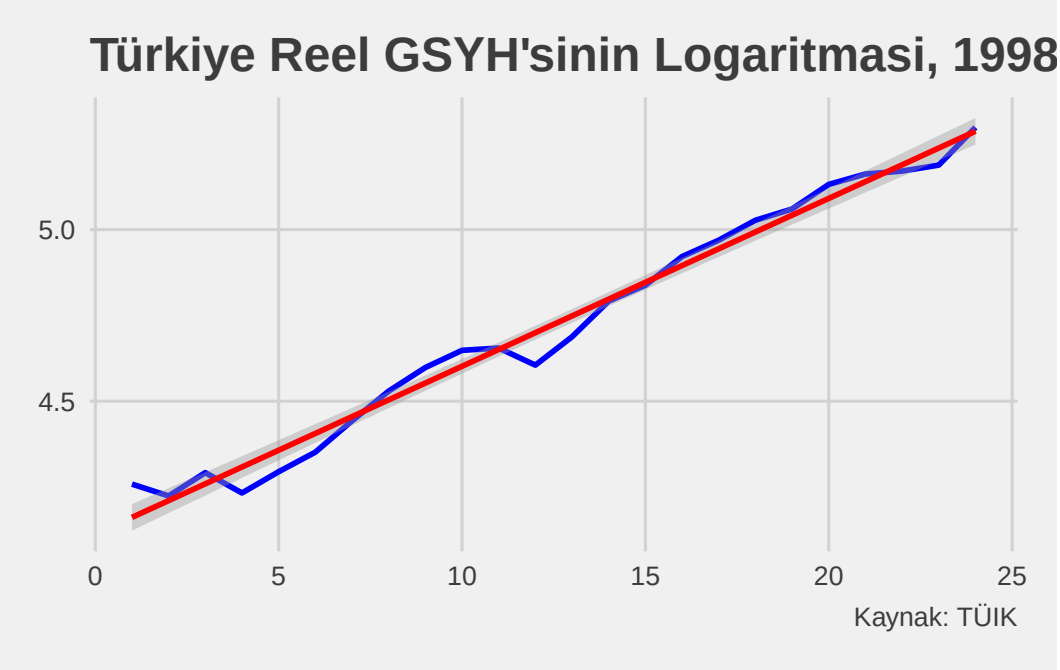
Bağımlı değişken:

- **ln_rgdp:** TÜİK'ten elde edilen mevsim ve takvim etkisinden arındırılmış harcama yöntemiyle GSYH, zincirlenmiş hacim endeksi

Bağımsız değişken:

- **t:** Zaman

```
ggplot(df2, aes(x = t, y = ln_rgdp)) +
geom_line(color = "blue", size = 1) +
geom_smooth(method = "lm", color = "red", size = 1) +
theme_fivethirtyeight() +
labs(
    title = "Türkiye Reel GSYH'sinin Logaritması, 1998-2021",
    caption = "Kaynak: TÜİK"
)
```



Modeli kuralım.

```
loglin_model <- lm(ln_rgdp ~ t, data = df2)
summary(loglin_model)
```

```
Call:
lm(formula = ln_rgdp ~ t, data = df2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.09438 -0.02607  0.01240  0.02780  0.09694

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.11259    0.01986  207.09   <2e-16 ***
t            0.04891    0.00139   35.19   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04713 on 22 degrees of freedom
Multiple R-squared:  0.9825,    Adjusted R-squared:  0.9818
```

```
F-statistic:  1239 on 1 and 22 DF,  p-value: < 2.2e-16
```

Yukarıdaki model çıktısının yorumları şöyledir:

- Çok küçük p değerlerine sahip regresyon parametreleri bireysel açıdan istatistiksel olarak anlamlıdır.
- F istatistiğine ait p değeri çok küçük olduğu için model istatistiksel olarak anlamlıdır.
- Büyüme hızını hesaplamak için β_2 100 ile çarpılır. Model, 1998-2021 döneminde Türkiye reel GSYH'sinin yıllık %4.9 oranında artmış olduğunu gösteriyor. 35.19 tahmin edilen t değeri istatistiksel açıdan oldukça anlamlıdır. Bu da aynı zamanda büyüme hızının anlamlı olduğunu gösterir.
- Kesme terimi 4.11'in ters logaritması 61.11'dir. Bu da aşağı yukarı 1998'e ait değerdir.

```
df2$rgdp[1]
```

```
[1] 70.7
```

β_2 parametresi bileşik büyüme hızı r'yi değil; anlık büyüme hızını verir. Eğer $r = exp(\beta_2) - 1$ yapılırsa bileşik büyüme hızına ulaşılır.

```
exp(as.numeric(coef(loglin_model)[2])) - 1
```

```
[1] 0.05012937
```

Yani, yaklaşık %5.1'dir. Bileşim olması sebebiyle anlık büyüme hızı %4.9'dan biraz daha fazla hesaplanmıştır.

3.2.2 Ek: Doğrusal Trend Modeli

Yukarıdaki büyüme modeli yerine doğrusal trend modeli ile tahmin etmek istediğimizi varsayalım.

$RGDP_t = \beta_1 + \beta_2 zaman + \epsilon_t$

Yukarıdaki eşitlikte, β_2 eğim parametresi birim dönemde RGDP'deki mutlak değişimi verir. β_2 pozitif ise RGDP'de (farklı bir değişken de olabilirdi) yükselen bir trend vardır. Tam tersi, negatif ise azalan bir trend anlamına gelecektir.

```r
dtrend_model <- lm(rgdp ~ t, data = df2)
summary(dtrend_model)
```

```
Call:
lm(formula = rgdp ~ t, data = df2)

Residuals:
     Min       1Q   Median       3Q      Max
-16.4597  -3.4598  -0.7431   3.3737  17.1847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  47.7931     3.3730   14.17 1.54e-12 ***
t             5.7222     0.2361   24.24  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.005 on 22 degrees of freedom
Multiple R-squared:  0.9639,    Adjusted R-squared:  0.9623
F-statistic: 587.6 on 1 and 22 DF,  p-value: < 2.2e-16
```

1998-2021 döneminde endeksin 5.7 arttığını görüyoruz. Aynı zamanda artı işaretli olması yükselen bir trendin olduğunu göstermektedir.

3.3 Lin-Log

Log-Lin ile açıklayıcı değişkendeki 1 birim değişime karşılık bağımlı değişkendeki yüzde büyüme ile ilgileniyorduk. Lin-Log'da ise aşağıdaki model tahmin edilir.

$Y_i = \beta_1 + \beta_2 lnX_i + \epsilon_i$

Yani, lin-log modellerde bağımlı değişken doğrusal yapıda iken en az bir açıklayıcı değişken logaritmik yapıdadır.

3.3.1 Uygulama

ABD için gıda harcamasına bakalım. İlgili verilere *rmfk3.xls* dosyası ile ulaşılabilir.

Aşağıda açıklayıcı değişkenin logaritmasını ekledik.

```r
df3 <- read_excel("./data/rmfk3.xls")

df3 %<>%
select(sfdho,expend) %>%
mutate(
    ln_expend = log(expend)
)
```

Bağımlı değişken:

- **sfdho:** Gıda harcamasının toplam harcamadaki payı

Bağımsız değişken:

- **expend:** Toplam hanehalkı harcaması

```r
linlog_model <- lm(sfdho ~ ln_expend, data = df3)
summary(linlog_model)
```

```
Call:
lm(formula = sfdho ~ ln_expend, data = df3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.18180 -0.04350 -0.00654  0.03373  0.48594

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.930387   0.036367   25.58   <2e-16 ***
ln_expend   -0.077737   0.003591  -21.65   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06875 on 867 degrees of freedom
Multiple R-squared:  0.3509,    Adjusted R-squared:  0.3501
F-statistic: 468.6 on 1 and 867 DF,  p-value: < 2.2e-16
```

Yukarıdaki model çıktısının yorumları şöyledir:

- Çok küçük p değerlerine sahip regresyon parametreleri bireysel açıdan istatistiksel olarak anlamlıdır.

- F istatistiğine ait p değeri çok küçük olduğu için model istatistiksel olarak anlamlıdır.
- Toplam harcama %1 arttığında gıda harcamasının toplam harcamadaki payı ortalamada 0.0008 birim düşecektir. Burada önemli bir not: Tahmin edilen eğim parametresi değeri 0.01 ile çarpılmalı ya da 100'e bölünmelidir. Yorum şöyle de yapılabilir: Toplam harcama %100 arttığında gıda harcamasının toplam harcamadaki payı ortalamada 0.08 birim azalır.

3.4 Ters Model

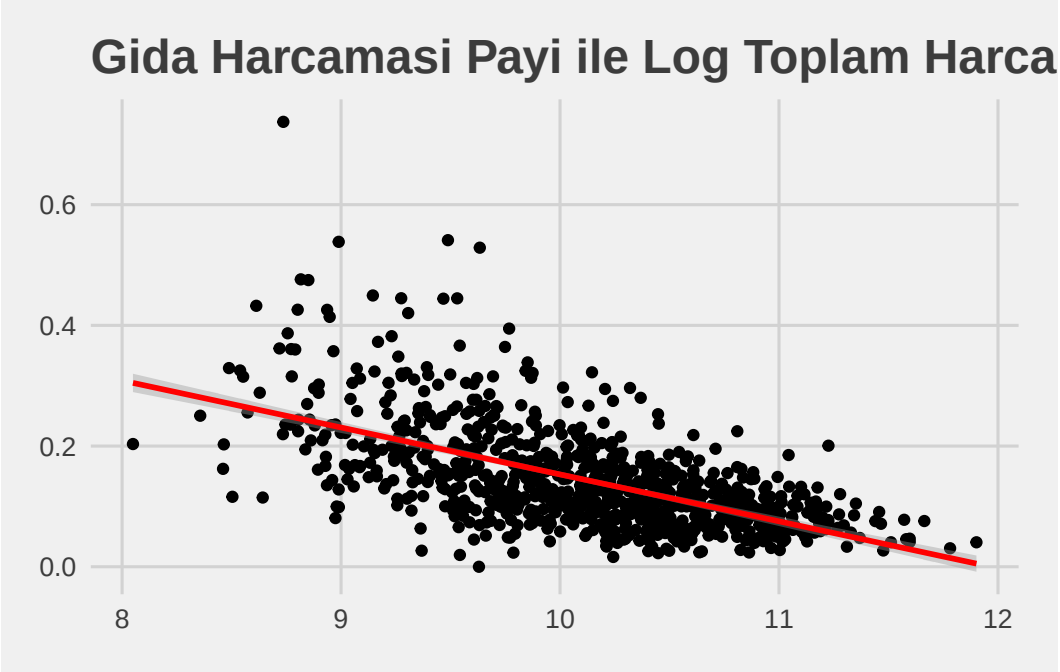
Yukarıdaki örneğe ait ilişki ters olabilir mi?

$Y_i = \beta_1 + \beta_2(\frac{1}{X_i}) + \epsilon_i$

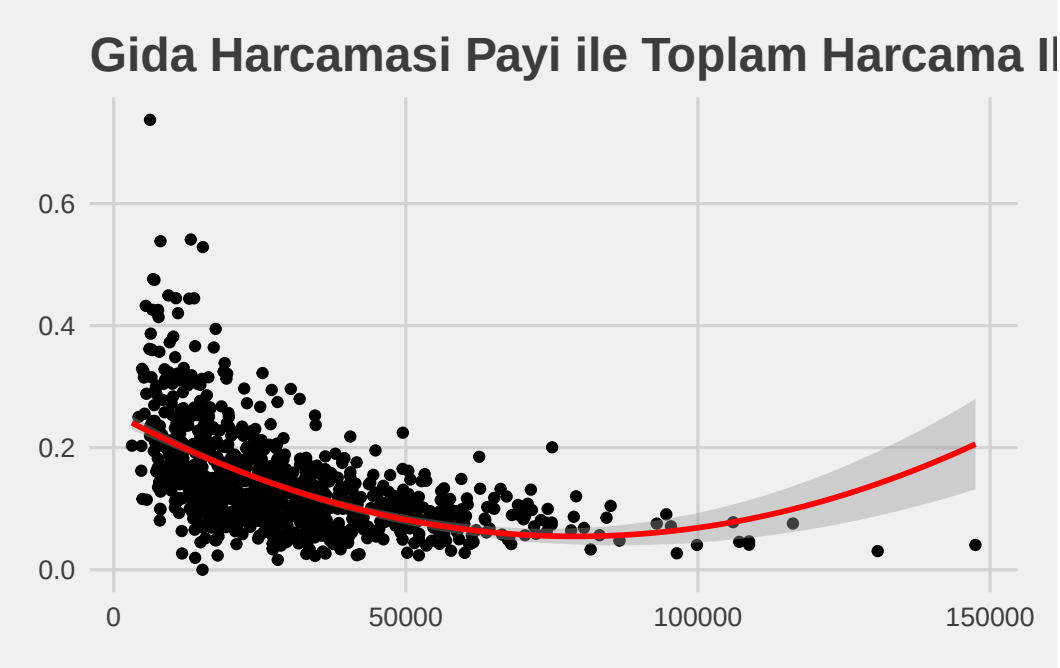
$SFDHO = \beta_1 + \beta_2(\frac{1}{Expend_i}) + \epsilon_i$

3.4.1 Uygulama

```
ggplot(df3, aes(x = ln_expend, y = sfdho)) +
geom_point() +
geom_smooth(method = "lm", color = "red") +
theme_fivethirtyeight() +
labs(
    title = "Gıda Harcaması Payı ile Log Toplam Harcama İlişkisi"
)
```




```
ggplot(df3, aes(x = expend, y = sfdho)) +
geom_point() +
geom_smooth(method = "lm", formula = y ~ poly(x, 2, raw = TRUE), color = "red") +
theme_fivethirtyeight() +
labs(
    title = "Gıda Harcaması Payı ile Toplam Harcama İlişkisi"
)
```




```
df3 %<>%
mutate(
    ters_expend = 1 / expend
)
```

Modeli kuralım.

```
ters_model <- lm(sfdho ~ ters_expend, data = df3)
summary(ters_model)
```

```
Call:
lm(formula = sfdho ~ ters_expend, data = df3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.29889 -0.04205 -0.01120  0.03229  0.44606

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.726e-02  4.012e-03   19.26   <2e-16 ***
ters_expend 1.331e+03  6.396e+01   20.82   <2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06968 on 867 degrees of freedom
Multiple R-squared:  0.3332,    Adjusted R-squared:  0.3325
F-statistic: 433.3 on 1 and 867 DF,  p-value: < 2.2e-16
```

Yukarıdaki model çıktısının yorumları şöyledir:

- Çok küçük p değerlerine sahip regresyon parametreleri bireysel açıdan istatistiksel olarak anlamlıdır.
- F istatistiğine ait p değeri çok küçük olduğu için model istatistiksel olarak anlamlıdır.
- Toplam harcama sonsuza gittiğinde gıda harcaması payı yaklaşık olarak %8'e yerleşecektir. Çünkü açıklayıcı değişken X sonsuza giderken $\beta_2(1/X_i)$ sıfıra yaklaşır. Y ise limit değer olan β_1'e yaklaşır.
- β_2 eğim parametresi pozitiftir. Yani, gıda harcaması payının toplam harcamaya göre değişim hızı her noktada negatif olacaktır. Eğer β_2 negatif olsaydı bu defa her noktada pozitif olacaktı.

3.5 Polinom Regresyon

3.5.1 Uygulama

Türkiye'nin reel GSYH'si için aşağıdaki modele bakalım.

$RGDP_t = \beta_1 + \beta_2 zaman + \beta_3 zaman^2 + \epsilon_t$

Yukarıdaki eşitlik zaman değişkenine göre karesel fonksiyon ya da ikinci derece polinomdur. Eğer bu modele $zaman^3$ eklenseydi üçüncü derece bir polinom denklemi olacaktı. Kısaca, açıklayıcı değişkenin en büyük kuvveti polinomun derecesine eşittir.

```
df2 %<>%
mutate(
    t2 = t^2
)
```

Modeli kuralım.

```
polinom2_model <- lm(rgdp ~ t + t2, data = df2)
summary(polinom2_model)
```

```
Call:
lm(formula = rgdp ~ t + t2, data = df2)

Residuals:
   Min     1Q Median     3Q    Max
-9.538 -3.735  1.451  3.059  5.853

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 63.52376    3.20715  19.807 4.55e-15 ***
t            2.09207    0.59110   3.539  0.00194 **
t2           0.14521    0.02295   6.326 2.84e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.807 on 21 degrees of freedom
Multiple R-squared:  0.9876,    Adjusted R-squared:  0.9864
F-statistic: 834.9 on 2 and 21 DF,  p-value: < 2.2e-16
```

$\frac{dRGDP}{dt} = 2.09 + 2(0.15)zaman = 2.09 + 0.3zaman$

Yukarıda, RGDP'nin değişim hızı bu değişim hızının ölçüldüğü zamana bağlıdır. Zamana göre ikinci türev alınırsa, 0.3 değeri elde edilir. Bu da değişim hızının zaman içinde sabit olan değişim hızıdır. İkinci türev pozitif ise RGDP artan bir oranda artmaktadır.

Az önce tahmin edilen model yerine aşağıdaki model tahmin edilmek istensin.

$lnRGDP_t = \beta_1 + \beta_2 t + \beta_3 t^2 + \epsilon_t$

Modeli kuralım.

```
polinom2_model_l <- lm(ln_rgdp ~ t + t2, data = df2)
summary(polinom2_model_l)
```

```
Call:
lm(formula = ln_rgdp ~ t + t2, data = df2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.085964 -0.032355  0.007408  0.030899  0.082041
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.1317305  0.0317325 130.205  < 2e-16 ***
t           0.0444972  0.0058485   7.608 1.82e-07 ***
t2          0.0001766  0.0002271   0.778    0.445
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04756 on 21 degrees of freedom
Multiple R-squared:  0.983, Adjusted R-squared:  0.9814
F-statistic: 608.5 on 2 and 21 DF,  p-value: < 2.2e-16
```

$\frac{dlnRGDP}{dt} = \beta_2 + 2\beta_3 t$

$\frac{1}{RGDP}\frac{dRGDP}{dt} = \beta_2 + 2\beta_3 t$

RGDP'nin büyüme hızı $= \beta_2 + 2\beta_3 t = 0.044 + 0.0002t$

RGDP'nin büyüme hızı birim zaman başına 0.0002 oranında artmaktadır.

4 Nitel Açıklayıcı Değişkenli Regresyon Modelleri

4.1 Ücret Modeli

Saatlik ücreti (dolar bazında) belirleyen faktörleri örneklemde araştırmak üzere, Mart 1995'te görüşme yapılan 1289 kişilik (anakütleden alınan örneklem) bir yatay-kesite bakmış ve bir regresyon modeli kurmuştuk. Burada kullandığımız kukla değişkenler aşağıdaki gibi gösterilebilir.

$Wage_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 D_{4i} + \beta_5 Educ_i + \beta_6 Exper_i + \epsilon_i$

Yukarıdaki modelde D'ler kukla (dummy) değişkenlerdir.

D_{2i} : Kadın 1, Erkek 0

D_{3i} : Beyaz Değil 1, Beyaz 0

D_{4i} : Sendikalı 1, Sendikasız 0

Kukla değişkenler söz konusu olduğu zaman aşağıdaki değerlendirmeler dikkate alınmalıdır.

- Eğer modele kesme terimi dahil edilmiş ise kukla değişken sayısının 1 eksiği kadar kukla değişken belirlenir. Yani, m kategori varsa (m-1) kukla değişken belirlenir.

Yukarıdaki değerlendirme dikkate alınmazsa kukla değişken tuzağı denilen duruma düşülür ki bu da tam doğrusal bağlantı durumudur. Örneğin, 3 kategori için 3 kukla değişken ve bir de kesme terimi olsun. Üç kukla değişkenin toplamı 1 olacaktır. Bu değer ise 1 olan ortak kesme terimine eşit olacaktır. Buna tam doğrusallık denir.

- Eğer tüm kategoriler kukla değişken olarak dahil edilmek isteniyorsa bir önceki maddede bahsedilen kukla değişken tuzağından kurtulmak için kesme terimi modelden çıkarılmalıdır.
- Bir kategori 0 değerini alıyorsa (kadın 1; erkek 0 gibi) referans, gösterge ya da karşılaştırma kategorisi ismini alır.
- Kukla değişkenlerin (1 ve 0 olarak belirlendiğinde) logaritması alınmamalıdır.

Kukla değişkenleri yorumlamadan önce model çıktısını hatırlayalım.

```r
library(readxl)
library(tidyverse)
library(ggthemes)
library(magrittr)
library(zoo)
#library(segmented)

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

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

model <- lm(formula = wage ~ female + nonwhite + union + education + exper, 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
```

Diğer tüm değişkenler sabit tutulduğunda;

- female: Bir kadın işçinin ortalama saatlik ücreti bir erkek işçinin ortalama saatlik ücretine göre $3.07 daha düşüktür. Referans kategori erkektir.
- nonwhite: Beyaz olmayan bir işçinin ortalama saatlik ücreti beyaz olan bir işçinin ortalama saatlik ücretine göre $1.56 daha düşüktür. Referans kategori beyaz.
- union: Sendikalı bir işçinin ortalama saatlik ücreti sendikasız bir işçinin ortalama saatlik ücretine göre $1.09 daha düşüktür. Referans kategori sendikasız.

Tüm p'ler neredeyse sıfır olduğu için kukla parametrelerinin tamamı istatistiksel olarak anlamlıdır.

-7.18 olan ortak kesme değeri, 0 değeri alan kategorileri temsil eder. Beyaz, sendikasız ve erkek işçi için beklenen saatlik ücrettir denilebilir.

4.2 Yapısal Değişimdeki Rolü

1959 - 2007 dönemi için ABD'de brüt özel yatırımlar (GPI) ile brüt özel tasarruflar (GPS) arasındaki ilişkiye bakalım. İlgili verilere *kd1.xls* dosyasından ulaşılabilir.

Yatırım fonksiyonu:

$GPI_t = \beta_1 + \beta_2 GPS_t + \epsilon_t, \beta_2 > 0$

Herhangi bir yapısal kırılma olup olmadığı değerlendirmesine girmeden model aşağıdaki gibi kurulsun.

```
df1 <- read_excel("./data/kd1.xls") %>%
select(obs,gpi,gps)

model_yapisal_normal <- lm(gpi ~ gps, data = df1)
summary(model_yapisal_normal)
```

```
Call:
lm(formula = gpi ~ gps, data = df1)

Residuals:
    Min      1Q  Median      3Q     Max
-275.06  -65.89   29.12   55.31  336.84

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -78.72105   27.48474  -2.864  0.00623 **
```

```
gps           1.10740    0.02908  38.081  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 114.9 on 47 degrees of freedom
Multiple R-squared:  0.9686,    Adjusted R-squared:  0.9679
F-statistic:  1450 on 1 and 47 DF,  p-value: < 2.2e-16
```

GPS $1 arttığında ortalama GPI $1.1 artmaktadır.

Peki, yapısal kırılma var mıdır?

1981'den itibaren olan gözlemler için 1 değerini alan kategori Recession81 olsun. Bu durumda 1981 öncesi de 0 değerini alacaktır.

$GPI_t = \beta_1 + \beta_2 GPS_t + \beta_3 Recession81_t + \epsilon_t$

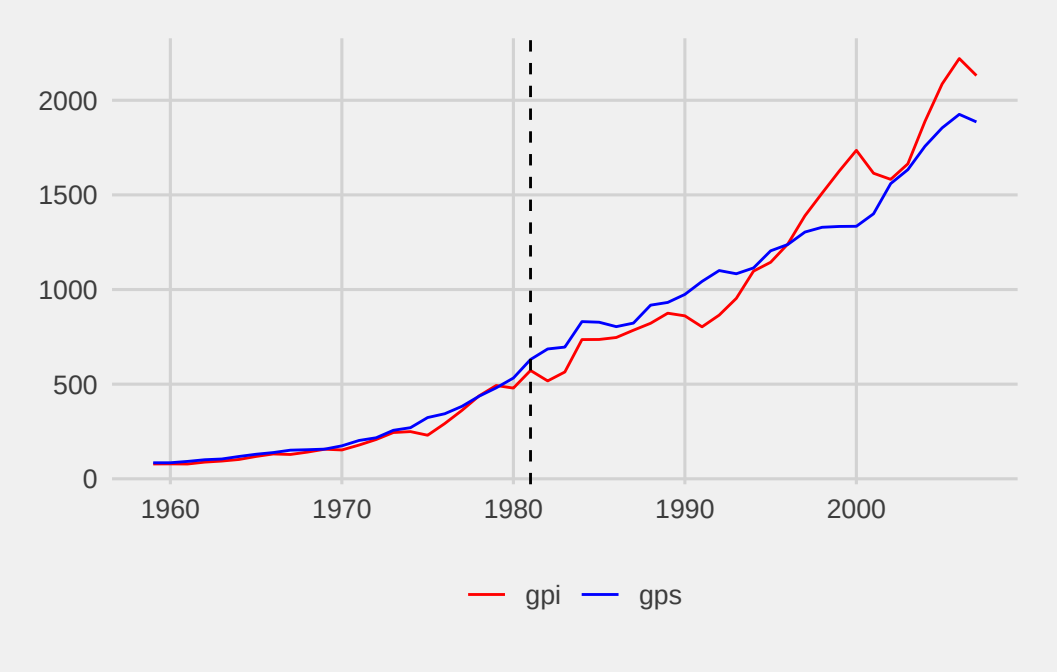
```r
df1 %<>%
mutate(
    recession81 = ifelse(
        obs >= 1981, 1, 0
    )
)

model_yapisal_81 <- lm(gpi ~ gps + recession81, data = df1)
summary(model_yapisal_81)
```

```
Call:
lm(formula = gpi ~ gps + recession81, data = df1)

Residuals:
    Min      1Q  Median      3Q     Max
-228.98  -42.82    0.73   37.82  340.56

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -83.48603   23.15913  -3.605 0.000765 ***
gps            1.28867    0.04707  27.380  < 2e-16 ***
recession81 -240.78785   53.39663  -4.509 4.46e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 96.69 on 46 degrees of freedom
Multiple R-squared:  0.9782,    Adjusted R-squared:  0.9773
F-statistic:  1034 on 2 and 46 DF,  p-value: < 2.2e-16
```

-240.78 olan kukla parametresi istatistiksel olarak anlamlıdır. Resesyon öncesi dönem ise (-83.48 - 240.78) = -324.26'dır.

```
df1 %>%
select(-recession81) %>%
pivot_longer(!obs, names_to = "vars", values_to = "vals") %>%
ggplot(aes(x = obs, y = vals, color = vars)) +
geom_line() +
geom_vline(xintercept = 1981, linetype = "dashed") +
theme_fivethirtyeight() +
theme(legend.title = element_blank()) +
scale_color_manual(values = c("red","blue"))
```



4.3 Mevsimsel Verilerdeki Rolü

Mevsime duyarlı olan moda giyim satışlarına bakalım. İlgili verilere *kd2.xls* dosyasından ulaşılabilir.

Model aşağıdaki gibi olsun.

$Sales_t = \beta_1 + \beta_2 D_{2t} + \beta_3 D_{3t} + \beta_4 D_{4t} + \epsilon_t$

D_2 : İkinci çeyrek için 1

D_3 : Üçüncü çeyrek için 1

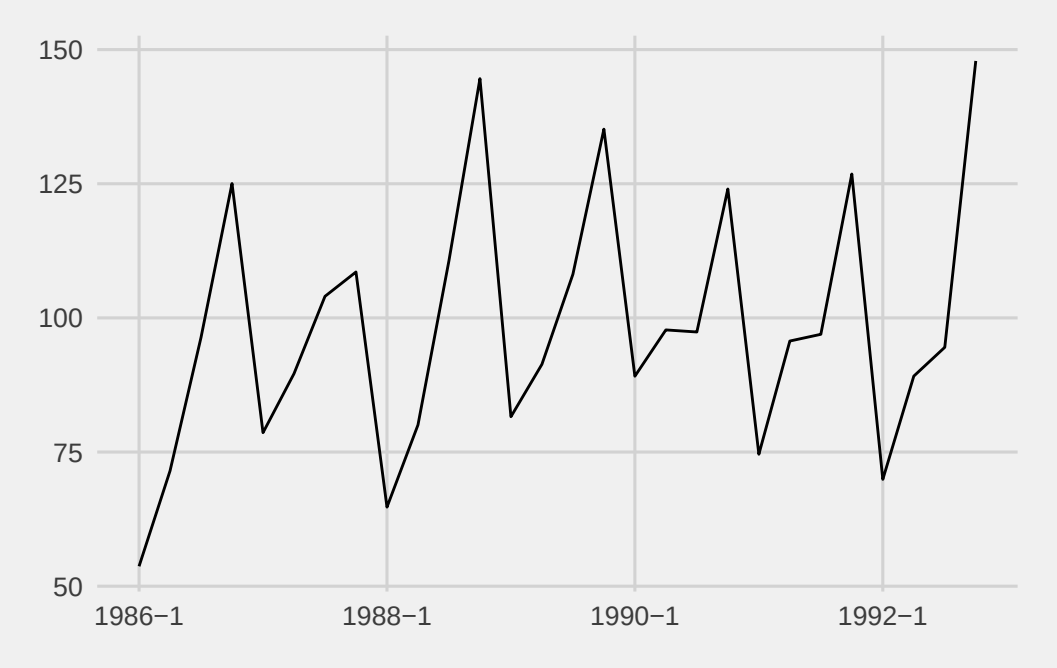
D_4 : Dördüncü çeyrek için 1

$Sales$: Bin fit-karelik perakende alanı başına reel satış

β_1 : Birinci çeyrek ortalama satış değeri

İlk çeyrek referans çeyrektir.

```
df2 <- read_excel("./data/kd2.xls") %>%
select(obs,sales) %>%
mutate(
    obs = as.yearqtr(gsub(":","-",obs))
)

ggplot(df2, aes(x = obs, y = sales)) +
geom_line() +
theme_fivethirtyeight()
```



```r
df2 %<>%
mutate(
    quarter = substr(obs,6,7),
    "q1" = ifelse(quarter == "Q1", 1, 0),
    "q2" = ifelse(quarter == "Q2", 1, 0),
    "q3" = ifelse(quarter == "Q3", 1, 0),
    "q4" = ifelse(quarter == "Q4", 1, 0)
)

model_mevsim <- lm(sales ~ q2 + q3 + q4, data = df2)
summary(model_mevsim)
```

```
Call:
lm(formula = sales ~ q2 + q3 + q4, data = df2)

Residuals:
    Min      1Q  Median      3Q     Max
-21.740  -5.511   1.340   7.193  17.587

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   73.183      3.977  18.399 1.18e-15 ***
q2            14.692      5.625   2.612   0.0153 *
q3            27.965      5.625   4.971 4.47e-05 ***
q4            57.115      5.625  10.154 3.65e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.52 on 24 degrees of freedom
Multiple R-squared:  0.8235,    Adjusted R-squared:  0.8014
F-statistic: 37.32 on 3 and 24 DF,  p-value: 3.372e-09
```

- D_2 : İkinci çeyrekteki ortalama satış değeri referans çeyrekteki ortalama satıştan 14.692 birim daha yüksektir. İkinci çeyrek ortalama satış değeri = 73.183 + 14.692 = 87.875'tir.
- D_3 : Üçüncü çeyrekteki ortalama satış değeri referans çeyrekteki ortalama satıştan 27.965 birim daha yüksektir. Üçüncü çeyrek ortalama satış değeri = 73.183 + 27.965 = 101.148'tir.
- D_4 : Dördüncü çeyrekteki ortalama satış değeri referans çeyrekteki ortalama satıştan 57.115 birim daha yüksektir. Dördüncü çeyrek ortalama satış değeri = 73.183 + 57.115 = 130.298'tir.

p değerlerine göre kukla parametreleri istatistiksel olarak anlamlıdır.

Mevsimsel etkiden arındırmak için aşağıdaki yol izlenebilir:

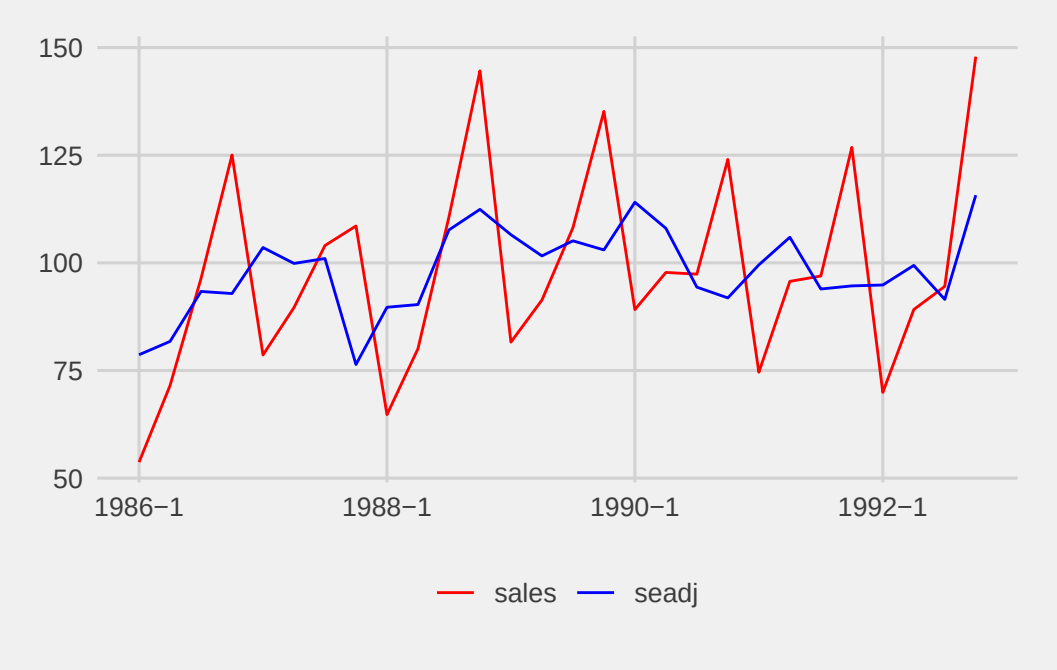
i. Tahmin edilen modelden satış hacmi tahmini bulunur.

ii. Gerçek satış hacminden tahmin edilen satış değeri çıkartılır ve kalıntılar elde edilir.

iii. Örneklem ortalama satış değeri elde edilen kalıntılara eklenir.

```
df2 %<>%
mutate(
    f_sales = model_mevsim$fitted.values,
    resid = sales - f_sales,
    seadj = mean(sales) + resid
)
```

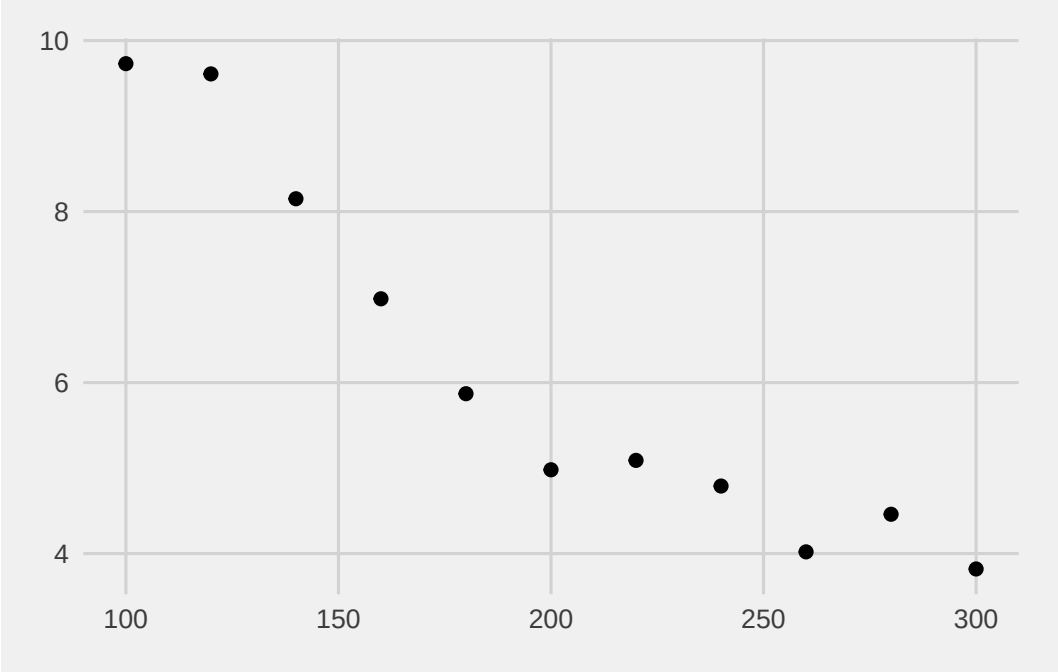
```
df2 %>%
select(obs,sales,seadj) %>%
pivot_longer(!obs, names_to = "vars", values_to = "vals") %>%
ggplot(aes(x = obs, y = vals, color = vars)) +
geom_line() +
theme_fivethirtyeight() +
theme(legend.title = element_blank()) +
scale_color_manual(values = c("red","blue"))
```



4.4 Parçalı Doğrusal Regresyon

Hipotetik olarak bir üreticinin birim başına ortalama maliyeti (ac) ile üretim partisi büyüklüğüne (lot_size) bakalım. İlgili verilere *kd3.xlsx* dosyasından ulaşılabilir.

```
df3 <- readxl::read_excel("./data/kd3.xlsx")

ggplot(df3, aes(x = lot_size, y = ac)) +
geom_point(size = 2) +
theme_fivethirtyeight() +
labs(
    x = "Parti Büyüklüğü",
    y = "Ortalama Maliyet"
)
```



Yukarıdaki görselde parti büyüklüğü 200 öncesi ve sonrası için farklı doğruların çalıştırılabileceği düşünülebilir. Yani, 200 eşik değerdir.

```
model_parcali <- lm(ac ~ lot_size, data = df3)
segmented_model_parcali <- segmented::segmented(
    model_parcali, seg.Z = ~lot_size, psi = 200
)
summary(segmented_model_parcali)
```

```
    ***Regression Model with Segmented Relationship(s)***

Call:
segmented.lm(obj = model_parcali, seg.Z = ~lot_size, psi = 200)

Estimated Break-Point(s):
                  Est. St.Err
psi1.lot_size 195.766 10.524

Meaningful coefficients of the linear terms:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.313000   0.753708  20.317 1.75e-07 ***
```

```
lot_size    -0.051750   0.005277  -9.807 2.43e-05 ***
U1.lot_size  0.039664   0.006615   5.996       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3337 on 7 degrees of freedom
Multiple R-Squared: 0.9833,  Adjusted R-squared: 0.9762

Boot restarting based on 6 samples. Last fit:
Convergence attained in 2 iterations (rel. change 2.359e-12)
```

Kullanılan fonksiyon x = 195.766'da bir kırılma tespit etti.

Eğer x $\leq$ 195.766 ise;

y = 15.313 - 0.052 * lot_size

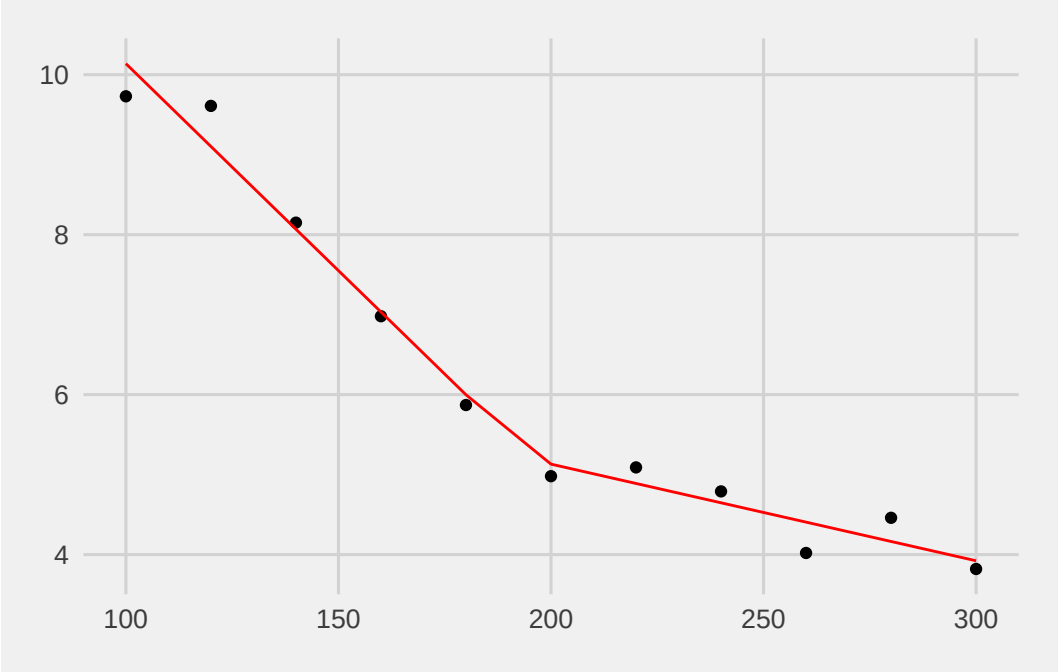
Eğer x $>$ 195.766 ise;

(15.313 - 0.052 * 195.766) + (-0.052 + 0.0396) * (lot_size-195.766)

5.133 - 0.0124 * (lot_size-195.766)

195.766 birimlik parti büyüklüğüne kadar parti büyüklüğündeki birim artış başına birim maliyet 5 kuruş azalıyor. 195.766 birimlik parti büyüklüğünden sonra ise yaklaşık 1 kuruş azalıyor.

```r
df3 %<>%
mutate(
    pr = segmented_model_parcali$fitted.values
)
ggplot(df3) +
geom_point(aes(x = lot_size, y = ac)) +
geom_line(aes(x = lot_size, y = pr), color = "red") +
theme_fivethirtyeight() +
labs(
    x = "Parti Büyüklüğü",
    y = "Ortalama Maliyet"
)
```


10
8
6
4
100
150
200
250
300

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