# Regresyon

Hakan Mehmetcik

2024-12-19

## Regresyon Analizine Giriş

### Fonksiyonlar ve Modeller

Bir **fonksiyon**, belirli bir girdiyle belirli bir çıktıyı ilişkilendirir. Yani, her girdiye karşılık yalnızca bir çıktı elde edersiniz. Bu ilişkiyi daha somut hale getirmek için bir içecek otomatını örnek olarak düşünelim:

- Eğer "Kahve" tuşuna basarsanız, otomat her zaman kahve verir.
- Eğer "Çay" tuşuna basarsanız, otomat her zaman çay verir.
- Aynı tuşa bastığınızda farklı bir içecek çıkması mümkün değildir.

Bir fonksiyon genellikle f(x) olarak gösterilir, burada x girdidir. Örneğin, $(x) = x^2$ fonksiyonunda, f(x) fonksiyonu x değerini alır ve onun karesini çıktı olarak verir!

İstatistik açısından **fonksiyonlar, bir modelin nasıl çalıştığını anlamak için bir çerçeve sağlar.**

### Bir İstatiksel Modelleme Olarak Regresyon

**İstatistiksel modelleme**, değişkenler arasındaki ilişkileri belirlemenin bir yolunu sunar. Bu, incelediğimiz sistemi daha iyi anlamamıza yardımcı olur. Regresyon, istatistikteki en basit modelleme yöntemidir.

Regresyon analizi, tahmin ve hipotez testini tamamlayan temel bir çıkarımsal istatistik yöntemidir. Bu yöntem, değişkenler arasındaki ilişkileri incelemek üzere modeller geliştirmek için kullanılır.

## Regresyon Türleri

### 1. Doğrusal Regresyon (Linear Regression)

Bu, bağımlı değişken ile bir veya daha fazla bağımsız değişken arasındaki ilişkinin doğrusal olduğu varsayımıyla oluşturulan en basit regresyon türüdür.

Aşağıdaki gibi ifade edilir: $Y = \beta0 + \beta1X1 + \epsilon Y$

Burada:

- Y: Bağımlı değişken
- X1: Bağımsız değişken
- 0: Sabit terim (intercept)
- 1: Katsayı (eğim)
- \epsilon: Hata terimi

### 2. Çoklu Regresyon (Multiple Regression)

Doğrusal regresyonun genişletilmiş bir halidir ve bağımlı değişkeni daha iyi tahmin edebilmek için birden fazla bağımsız değişken kullanılır.

Aşağıdaki gibi ifade edilir:

$$Y = \beta0 + \beta1X1 + \beta2X2 + ... + \beta nXn + \epsilon$$

Burada:

- Y: Bağımlı değişken
- X1,X2,…,Xn: Bağımsız değişkenler
- 0, 1, 2,…, n: Katsayılar
- : Hata terimi

### 3. Lojistik Regresyon (Logistic Regression)

Bağımlı değişkenin kategorik olduğu durumlarda kullanılır. Genellikle ikili sonuçlar için bir olayın gerçekleşme olasılığını tahmin etmek amacıyla verileri lojistik bir eğriye uydurur.

### Çıkarımsal İstatistikte Regresyon

Çıkarımsal istatistikte regresyon analizi şu amaçlarla kullanılır:

1. **İlişkileri Tahmin Etme:** Değişkenler arasındaki ilişkinin gücünü ve anlamlılığını belirler.
2. **Sonuçları Tahmin Etme:** Regresyon modeli kullanılarak, bağımsız değişkenlerin belirli değerleri için bağımlı değişkenin değerleri tahmin edilir.
3. **Hipotezleri Test Etme:** Değişkenler arasındaki ilişkiler hakkında belirli hipotezleri test eder. Örneğin, bağımsız değişkenlerin etkisinin anlamlı olup olmadığını değerlendirir.

### Örnek 1:

Bu alıştırmada, bir grup üniversite öğrencisine ait iki değişken içeren ikili (bivariate) veri bulunmaktadır:

- Üniversite boyunca ders kitaplarına harcanan toplam miktar (dolar cinsinden),
- Genel not ortalaması (GPA).

Aşağıdaki doğrusal regresyon modeli, harcanan para miktarına (yüzlerce dolar cinsinden) göre GPA'yı tahmin etmek için kullanılmıştır:

$$\text{Tahmin Edilen GPA} = 2.84 + 0.04 \times \text{Ders Kitaplarına Harcanan Para (Dolar)}$$

### Soru 1:

Bir öğrenci üniversite boyunca ders kitaplarına toplam **970 dolar** harcamışsa, bu öğrencinin tahmin edilen GPA'sı nedir?

Harcanan para 0 dolar olduğunda (dolar cinsinden 0):

Tahmin Edilen GPA=2.84+0.04×0 Tahmin Edilen GPA=2.84

```
predictedgpa <- function(x) {
  2.84 + 0.04*x/100
}

predictedgpa(0)
```

```
[1] 2.84
```

**Soru 2:**

Bir öğrenci üniversite boyunca ders kitaplarına toplam **970 dolar** harcamışsa, bu öğrencinin tahmin edilen GPA'sı nedir?

```
predictedgpa(970)
```

```
[1] 3.228
```

**Soru 3:**

Bir öğrenci üniversitede ders kitaplarına **0 dolar** harcamış ve **3.71 GPA** ile mezun olmuşsa, bu öğrencinin **residüel değeri** nedir?

```
# Residüel hesaplama: Residüel = Gerçek Değer - Tahmin Edilen Değer

gercek_gpa <- 3.71  # Gerçek GPA değeri
tahmin_edilen_gpa <- 2.84  # Ders kitaplarına 0 dolar harcayan öğrencinin tahmin edilen GPA's

residuel <- gercek_gpa - tahmin_edilen_gpa  # Residüel hesaplama
cat("Residüel değer:", residuel)
```

```
Residüel değer: 0.87
```

> **Note**
>
> Residüeller, regresyon analizinin temel bir kavramıdır. Bir regresyon modelinde bağımlı değişkenin gözlemlenen değeri ile tahmin edilen değeri arasındaki farkı ölçerler. Residüeller, regresyon modelinin ne kadar iyi uyum sağladığını değerlendirmek ve modelin veriyi ne kadar iyi açıkladığını anlamak için kullanılır.
> Bir gözlem için **residüel** (e) şu şekilde hesaplanır:
>
> $$e = y_{\text{gözlemlenen}} - y_{\text{tahmin edilen}}$$

**Soru 4:**

Bir öğrenci ders kitaplarına **1,450 dolar** harcamış ve **2.91 GPA** ile mezun olmuşsa, bu öğrencinin residüel değeri nedir? Residüelin pozitif mi yoksa negatif mi olduğunu belirtin ve sonucu **virgülden sonra 2 basamak** olacak şekilde yuvarlayın.

```r
# Residüeli hespalayalım
residuel<- 2.91 - predictedgpa(1450)
cat("Residüel", residuel, "\n")
```

```
Residüel -0.51
```

```r
# Residüelin pozitif mi negatif mi olduğunu belirtelim
if (residuel > 0) {
  cat("Residüel pozitif, model gerçek GPA'yı düşük tahmin etmiş.\n")
} else if (residuel < 0) {
  cat("Residüel negatif, model gerçek GPA'yı yüksek tahmin etmiş.\n")
} else {
  cat("Residüel sıfır, model GPA'yı doğru tahmin etmiş.\n")
}
```

```
Residüel negatif, model gerçek GPA'yı yüksek tahmin etmiş.
```

**Soru 5:**

Bir birinci sınıf öğrencisinin **4 GPA** almak için ne kadar para harcaması gerektiğini hesaplayın.

```r
# Regresyon modeli katsayıları
beta_0 <- 2.84  # Sabit terim
beta_1 <- 0.04  # Katsayı
# Hedef GPA
hedef_gpa <- 4

# Harcanması gereken para (yüzlerce dolar cinsinden) için denklem çözümü
harcanan_para <- (hedef_gpa - beta_0) / beta_1

# Sonucu yazdır
cat(sprintf("4 GPA almak için harcanması gereken miktar: %.2f dolar\n", harcanan_para*100))
```

```
4 GPA almak için harcanması gereken miktar: 2900.00 dolar
```

> **Note**
>
> Bir birinci sınıf öğrencisi bu çalışmayı öğrendi ve 4.0 GPA almak için ders kitaplarına 2,900 dolar harcaması gerektiğini hesapladı. (Yukarıdaki denklemle bu hesabı az önce siz de doğruladınız). Bu öğrenci, GPA'sını artırmak için tüm ders kitaplarını ikinci el ve daha ucuz olanlar yerine sıfır olarak satın almaya karar verdi. Bu modeli istatistiksel olarak doğru bir şekilde mi kullanıyor?
> Cevap: Tabii ki hayır. Kesinlikle hayır!

### R'da Regresyon

**lm() fonksiyonu ile doğrusal modelleri simüle etme:** R'da doğrusal modeller genellikle **lm()** fonksiyonu kullanılarak oluşturulur. Bu fonksiyon, modelin toplam hata miktarını minimize eden en iyi uyum çizgisini (line of best fit) veriye uydurur.

```r
# Örnek Veri
set.seed(123)
study_hours <- runif(100, 0, 10)  # 0 ile 10 arasında rastgele çalışma saatleri üret
test_scores <- 50 + 5 * study_hours + rnorm(100, mean=0, sd=5)  # Test skorlarını üret

# Doğrusal Regresyon Modeli Oluştur
model <- lm(test_scores ~ study_hours)
summary(model)
```

```
Call:
lm(formula = test_scores ~ study_hours)

Residuals:
     Min       1Q   Median       3Q      Max
-11.1899  -3.0661  -0.0987   2.9817  11.0861

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  49.9552     0.9803   50.96   <2e-16 ***
study_hours   4.9551     0.1709   28.99   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.846 on 98 degrees of freedom
Multiple R-squared:  0.8956,    Adjusted R-squared:  0.8945
```

```
F-statistic: 840.6 on 1 and 98 DF,  p-value: < 2.2e-16
```

Bu kodda **lm(y ~ x)**, y değişkenini bağımlı değişken, x değişkenini ise bağımsız değişken olarak kullanarak doğrusal bir model oluşturur. **summary(model)** fonksiyonu ise modelin özetini verir; bu özet, katsayılar, residüeller ve diğer tanısal ölçümleri içerir.

### summary() Fonksiyonunun Sağladığı Bilgiler

1. **Call:**
   - **lm(formula = test_scores ~ study_hours):** Kullanılan modeli gösterir. Burada test_scores, study_hours'un bir fonksiyonu olarak tahmin edilmektedir.
2. **Residuals (Residüal/Artıklar):**
   - Artıklar, gözlemlenen değerler ile modelin tahmin ettiği değerler arasındaki farktır. Bu değerler (Min, 1Q (birinci çeyrek), Median, 3Q (üçüncü çeyrek) ve Max) artıkların dağılımını açıklar.
   - Daha küçük artıklar, modelin veriye daha iyi uyum sağladığını gösterir.
3. **Coefficients (Katsayılar):**
   - **(Intercept):** Çalışma saati 0 olduğunda beklenen test skorudur. Yaklaşık **49.9552**.
   - **study_hours:** Çalışma saatlerindeki her bir saatlik artış için test skorundaki artış tahmini. Yaklaşık **4.9551**.
   - **Std. Error:** Katsayı tahminlerinin doğruluğunu ölçer.
   - **t value:** Katsayının sıfırdan ne kadar uzak olduğunu gösterir (standart sapma cinsinden). Daha büyük mutlak değerler, daha anlamlı katsayılar olduğunu gösterir.
   - **Pr(>|t|):** Katsayının sıfırdan farklı olup olmadığını test eden p-değeri. **0.05'ten küçük** değerler katsayının anlamlı olduğunu gösterir. Her iki katsayı da anlamlıdır (<2e-16).
4. **Residual Standard Error:**
   - Yaklaşık **4.846**, yanıt değişkeninin gerçek regresyon doğrusundan ortalama sapmasını ölçer.
5. **Degrees of Freedom (Serbestlik Derecesi):**
   - Model **98 serbestlik derecesi** kullanmıştır (100 gözlem - 2 parametre).
6. **Multiple R-squared:**

- **0.8956**, bağımsız değişkenin bağımlı değişkendeki varyansın ne kadarını açıkladığını gösterir. Burada çalışma saatleri, test skorlarındaki değişkenliğin yaklaşık %89.56'sını açıklamaktadır.

7. **Adjusted R-squared:**

   - **0.8945**, R-kare değerinin bağımsız değişken sayısına göre düzeltilmiş hali. Bu değer genellikle R-kare'den daha düşüktür ve farklı sayıda bağımsız değişken içeren modelleri karşılaştırırken daha iyi bir ölçüdür.

8. **F-statistic:**

   - **840.6 (1 ve 98 DF ile):** Modelin genel uyumunu ölçer. Çok düşük p-değeri (<2.2e-16), modelin istatistiksel olarak anlamlı olduğunu gösterir.

**Genel Değerlendirme**

- Model veriye iyi uyum sağlamaktadır.
- **study_hours** değişkeni, test_scores'un önemli bir açıklayıcısıdır.
- Hem sabit terim (intercept) hem de eğim (slope) istatistiksel olarak anlamlıdır.
- Artıklar iyi dağılmış görünmekte ve model, test skorlarındaki değişkenliğin önemli bir kısmını açıklamaktadır.

### Örnek 2:

Havayolu gecikmeleri veri setinden bir soruyu ele alalım: Planlanan kalkış saati, beklenen uçuş gecikmesi üzerinde bir etkiye sahip mi? Çoğu kişi, sabah erken uçuşların daha az geciktiğini düşünür çünkü uçuş gecikmeleri gün boyunca kademeli olarak artma eğilimindedir. Bu teori veri ile destekleniyor mu?

**nycflights13** paketindeki **flights** veri çerçevesinde, kalkışın planlanan saatini belirten bir **hour** değişkeni bulunmaktadır. Bu veri setini kullanarak bu soruyu bir linear regresyon modeli kullanarak cevaplamaya çalışalım

```
library(nycflights13)
library(dplyr)
```

```
Attaching package: 'dplyr'
```

```
The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
SF <- flights |>
  filter(dest == "SFO", !is.na(arr_delay))

head(SF)
```

```
# A tibble: 6 x 19
   year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
  <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
1  2013     1     1      558            600        -2      923            937
2  2013     1     1      611            600        11      945            931
3  2013     1     1      655            700        -5     1037           1045
4  2013     1     1      729            730        -1     1049           1115
5  2013     1     1      734            737        -3     1047           1113
6  2013     1     1      745            745         0     1135           1125
# i 11 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
#   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
#   hour <dbl>, minute <dbl>, time_hour <dttm>
```

Çıktıda, dep_time olarak verilen değerleri saat olarak görmemiz gerekiyor. Bunun için bu veriyi saat formatına dönüştürüyoruz:

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v forcats   1.0.0     v readr     2.1.5
v ggplot2   3.5.1     v stringr   1.5.1
v lubridate 1.9.3     v tibble    3.2.1
v purrr     1.0.2     v tidyr     1.3.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```r
convert_time <- function(time_minutes) {
  # Calculate hours and minutes from total minutes
  hours <- time_minutes %/% 60
  minutes <- time_minutes %% 60

  # Create time strings, ensuring two-digit formatting
  time_str <- sprintf("%02d:%02d", hours, minutes)

  # Return formatted string
  time_str
}

# Times provided
times <- SF |>
  group_by(hour) |>
  count() |>
  pivot_wider(names_from = hour, values_from = n) |>
  data.frame()

# Apply the conversion function
formatted_times <- sapply(times, convert_time)

# Print formatted times
formatted_times
```

```
     X5      X6      X7      X8      X9     X10     X11     X12     X13     X14
"00:55" "11:03" "28:16" "16:27" "07:09" "29:04" "06:53" "08:24" "07:56" "08:48"
    X15     X16     X17     X18     X19     X20     X21
"15:46" "14:57" "24:51" "18:11" "12:11" "07:45" "00:57"
```

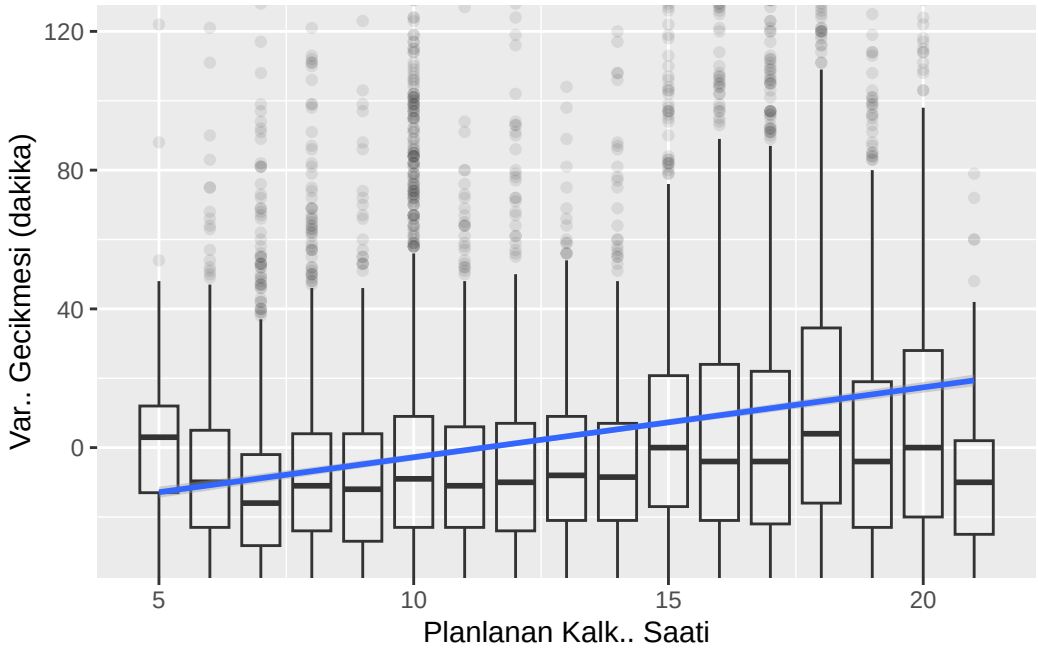
Şimdi daha açık bir şekilde birçok uçuşun sabah erken saatlerde ve öğleden sonra ile akşam arasında planlandığını görüyoruz. Hiçbir uçuş sabah 5'ten önce veya akşam 10'dan sonra planlanmamış. Şimdi varış gecikmesinin saate bağlı olarak nasıl değiştiğini iki şekilde inceleyeceğiz:

1. Standart kutu grafikleriyle (boxplot) gecikme dağılımını gösterme.
2. Gün boyunca ortalama varış gecikmesini izlemek için doğrusal model kullanma

```r
SF |>
  ggplot(aes(x = hour, y = arr_delay)) +
  geom_boxplot(alpha = 0.1, aes(group = hour)) +
```

```
  geom_smooth(method = "lm") +
  xlab("Planlanan Kalkış Saati") +
  ylab("Varış Gecikmesi (dakika)") +
  coord_cartesian(ylim = c(-30, 120))
```

```
`geom_smooth()` using formula = 'y ~ x'
```




Yukarıdaki grafik, varış gecikmesini planlanan kalkış saatine göre gösterir. Ortalama varış gecikmesi gün boyunca artış gösteriyor. Eğilim çizgisi, bir regresyon modeli kullanılarak oluşturulmuştur.

```
mod1 <- lm(arr_delay ~ hour, data = SF)
summary(mod1)
```

```
Call:
lm(formula = arr_delay ~ hour, data = SF)

Residuals:
   Min     1Q Median     3Q    Max
-97.32 -25.22  -9.17   9.83 993.66
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -22.93267    1.23275  -18.60   <2e-16 ***
hour          2.01487    0.09154   22.01   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 46.82 on 13171 degrees of freedom
Multiple R-squared:  0.03548,   Adjusted R-squared:  0.03541
F-statistic: 484.5 on 1 and 13171 DF,  p-value: < 2.2e-16
```

### Model Özeti

1. **Regresyon Denklemi:**

arr_delay = -22.93 + 2.01 x hour

```
-   **Intercept (-22.93):** Günün başında (saat 0) varış gecikmesinin ortalama olarak **-22.9

-   **hour Katsayısı (2.01):** Her saat başına, varış gecikmesinde **2.01 dakikalık bir artıs
```

2. **Artıklar (Residuals):**
   - **Min:** -97.32 — Model tahminlerinden en fazla 97.32 dakika daha erken varış yapılmış.
   - **Max:** 993.66 — Model tahminlerinden en fazla 993.66 dakika daha geç varış yapılmış.
   - **Median:** -9.17 — Çoğu tahmin, gerçek değerden yaklaşık 9.17 dakika sapmıştır.
   - Artıkların dağılımı, modelin hata payının (standart sapmanın) yüksek olduğunu gösterir.
3. **Katsayıların Anlamlılığı:**
   - **Intercept ve hour:** Her iki katsayının da p-değerleri $<$ **2e-16**, bu da katsayıların istatistiksel olarak anlamlı olduğunu ve tesadüfen oluşma ihtimalinin çok düşük olduğunu gösterir.
4. **Residual Standard Error (RSE):**

- **46.82:** Varış gecikmesinin tahmini değerlerden ortalama sapması 46.82 dakikadır. Bu, modelin hata payının oldukça yüksek olduğunu ve yalnızca saat bilgisinin uçuş gecikmelerini açıklamada yetersiz kaldığını gösterir.

5. **R-Kare ve Ayarlanmış R-Kare:**

   - **R-Kare (0.03548):** Model, uçuş gecikmelerindeki varyansın yalnızca **%3.5**'ini açıklayabilmektedir. Bu, saat değişkeninin gecikmeleri açıklamada çok zayıf bir değişken olduğunu gösterir.
   - **Ayarlanmış R-Kare (0.03541):** Bu değer, R-kare ile neredeyse aynı ve modelin genel açıklayıcılığının düşük olduğunu teyit eder.

6. **F-İstatistiği:**

   - **484.5 (p-değeri $<$ 2.2e-16):** Model genel olarak anlamlıdır; yani saat değişkeninin gecikmeler üzerinde anlamlı bir etkisi olduğu söylenebilir. Ancak, etkisinin gücü oldukça sınırlıdır.

**Yorum:**

- **Pozitif ilişki:** Planlanan kalkış saati arttıkça, varış gecikmesi de artma eğilimindedir. Her saat başına yaklaşık **2 dakika** gecikme eklenmektedir.
- **Modelin yetersizliği:** R-kare ve artık standart hatası (RSE) değerlerinden görüldüğü üzere, bu model uçuş gecikmelerindeki varyansın çok küçük bir kısmını açıklamaktadır. Sadece saat değişkeniyle uçuş gecikmelerini tam anlamıyla açıklamak mümkün değildir.
- **Ek değişkenlerin gerekliliği:** Gecikmeleri daha iyi modellemek için havayolu şirketi, mevsim, kalkış havalimanı gibi diğer değişkenlerin eklenmesi gereklidir.

Bu model, saat değişkeninin gecikmeler üzerindeki anlamlı etkisini göstermektedir ancak tek başına yeterli bir açıklama sağlayamamaktadır.

### Daha İyi Bir Model Oluşturabilir miyiz?

Uçuş gecikmelerini daha iyi açıklamak için başka hangi faktörler faydalı olabilir? Kalkış havalimanı, havayolu (taşıyıcı), yılın ayı ve haftanın günü gibi ek değişkenlere bakalım. Bazı veri düzenlemeleri yaparak haftanın gününü (**dow**) yıl, ay ve gün bilgisiyle çıkarabiliriz. Ayrıca, Haziran ve Temmuz aylarının uzun gecikmelerle bilindiğini özetleyen bir **season** değişkeni oluşturacağız. Bu değişkenler, varış gecikmesini (**arrival delay**) açıklamak için kullanılacak.

```r
library(lubridate)

SF <- SF |>
  mutate(
    day = as.Date(time_hour),
    dow = as.character(wday(day, label = TRUE)),
    season = ifelse(month %in% 6:7, "summer", "other month")
  )
```

### Yeni Model Oluşturma

Bu açıklayıcı değişkenleri içeren bir model oluşturabiliriz:

```r
mod2 <- lm(arr_delay ~ hour + origin + carrier + season + dow, data = SF)
summary(mod2)
```

```
Call:
lm(formula = arr_delay ~ hour + origin + carrier + season + dow,
    data = SF)

Residuals:
   Min     1Q Median     3Q    Max
-92.77 -24.73  -7.65  12.03 990.05

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -24.64131    2.17475 -11.331  < 2e-16 ***
hour           2.08154    0.08979  23.183  < 2e-16 ***
originJFK      4.11899    1.00484   4.099 4.17e-05 ***
carrierB6    -10.32394    1.88015  -5.491 4.07e-08 ***
carrierDL    -18.40939    1.61507 -11.399  < 2e-16 ***
carrierUA     -4.76173    1.48118  -3.215 0.001308 **
carrierVX     -5.06269    1.59833  -3.167 0.001541 **
seasonsummer  25.29210    1.03089  24.534  < 2e-16 ***
dowMon         1.74419    1.44774   1.205 0.228315
dowSat        -5.59915    1.54765  -3.618 0.000298 ***
dowSun         5.12336    1.47865   3.465 0.000532 ***
dowThu         3.16270    1.44868   2.183 0.029042 *
dowTue        -1.65371    1.44688  -1.143 0.253083
dowWed        -0.88422    1.44969  -0.610 0.541912
```

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

Residual standard error: 45.45 on 13159 degrees of freedom
Multiple R-squared:  0.09174,   Adjusted R-squared:  0.09084
F-statistic: 102.2 on 13 and 13159 DF,  p-value: < 2.2e-16
```

### Sonuçların Yorumu

Bu model, **planlanan kalkış saati**, **kalkış havalimanı**, **havayolu şirketi**, **mevsim** ve **haftanın günü** gibi değişkenlerin varış gecikmesi (**arr_delay**) üzerindeki etkisini inceleyen bir doğrusal regresyon modelidir.

### Model Özeti

#### 1. Regresyon Denklemi:

Model şu şekilde ifade edilebilir:

$$arr\_delay = -24.64 + 2.08 \cdot \text{hour} + 4.12 \cdot \text{originJFK} - 10.32 \cdot \text{carrierB6} + ...$$

- **Intercept (-24.64):** Kalkış saati 0 (gece yarısı) ve referans gruplar (EWR havalimanı, diğer aylar, Cuma günü) için ortalama varış gecikmesi **-24.64 dakika**. Negatif değer, bu koşullarda uçuşların genellikle erken varış yaptığını gösterebilir.

#### 2. Değişkenler Üzerindeki Etkiler:

- **Kalkış Saati (hour):** Her saatlik artış, varış gecikmesini **2.08 dakika artırır**. Bu, günün ilerleyen saatlerinde gecikmelerin arttığını doğrular.
- **Kalkış Havalimanı (originJFK):** JFK'den yapılan kalkışlar, Newark (EWR) referans grubuna kıyasla ortalama **4.12 dakika daha fazla gecikme** ile ilişkilidir.
- **Havayolu Şirketleri:**
  - **Delta (DL):** Ortalama olarak gecikmeler diğer taşıyıcılara kıyasla **18.41 dakika daha azdır**.
  - **JetBlue (B6):** Ortalama gecikme **10.32 dakika daha azdır**.
  - **United (UA):** Ortalama gecikme **4.76 dakika daha azdır**.
  - **Virgin America (VX):** Ortalama gecikme **5.06 dakika daha azdır**.

- **Mevsim (season):** Yaz aylarında (Haziran ve Temmuz) varış gecikmeleri ortalama **25.29 dakika daha uzundur**.
- **Haftanın Günleri (dow):**
  - **Cumartesi:** Ortalama gecikme **5.60 dakika daha azdır**.
  - **Pazar:** Ortalama gecikme **5.12 dakika daha fazladır**.
  - **Perşembe:** Ortalama gecikme **3.16 dakika daha fazladır**.
  - Diğer günlerin etkisi istatistiksel olarak anlamlı değildir.

### 3. Katsayıların Anlamlılığı (p-Değerleri):

- Değişkenlerin çoğu (örneğin, hour, originJFK, seasonsummer) $\mathbf{p < 0.001}$ ile anlamlıdır.
- **Haftanın bazı günleri (dowMon, dowTue, dowWed):** $p > 0.05$ olduğundan anlamlı değildir.

### 4. Artıklar (Residuals):

- **Artıkların Standart Hatası (RSE):** 45.45 dakika. Bu, modelin hata payının yüksek olduğunu ve varış gecikmesinin sadece bu açıklayıcı değişkenlerle tam olarak açıklanamayacağını gösterir.

### 5. R-Kare ve Ayarlanmış R-Kare:

- **R-Kare (0.09174):** Model, varış gecikmelerindeki toplam varyansın yalnızca **%9.2'sini** açıklayabilmektedir.
- **Ayarlanmış R-Kare (0.09084):** Eklenen açıklayıcı değişkenler hesaba katıldığında modelin açıklayıcılığı benzer düzeydedir.

### 6. F-İstatistiği:

- **F-İstatistiği (102.2, $\mathbf{p < 2.2e\text{-}16}$):** Modelin genel olarak anlamlı olduğunu ve açıklayıcı değişkenlerin gecikmeler üzerinde anlamlı bir etkisi olduğunu gösterir.

---

**Yorum:**

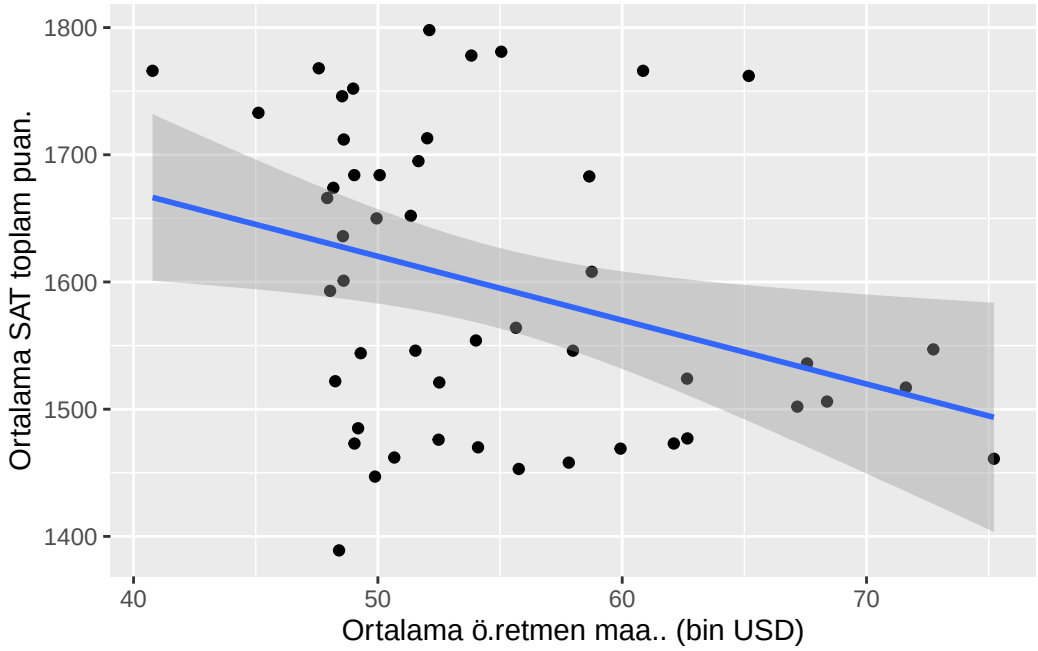
1. **Pozitif İlişkiler:**
   - Gün ilerledikçe uçuş gecikmeleri artar.
   - Yaz aylarında ve Pazar günleri gecikmeler daha uzundur.
   - JFK'den kalkış yapmak, daha uzun gecikmelerle ilişkilidir.
2. **Negatif İlişkiler:**
   - Delta ve JetBlue gibi bazı havayolu şirketleri, daha düşük ortalama gecikmelere sahiptir.
   - Cumartesi günleri gecikmeler daha azdır.
3. **Model Yetersizlikleri:**
   - **R-Kare değeri**, açıklayıcı değişkenlerin gecikmeleri açıklamada yetersiz olduğunu gösterir. Diğer faktörler (hava durumu, uçak tipi vb.) modele dahil edilmelidir.
4. **Genel Değerlendirme:**
   - Model bazı genel trendleri gösterse de, varış gecikmelerini tahmin etmede oldukça sınırlıdır. Daha fazla açıklayıcı değişken eklenmesi gereklidir.

**Örnek 3:**

2010 yılına ait gözlemsel veriler kullanarak, eyaletler düzeyinde öğretmen maaşlarının (ortalama öğretmen maaşı) ve SAT (Scholastic Aptitude Test) toplam puanlarının ilişkisini inceleyelim. SAT, üniversiteye giriş için kullanılan önemli bir sınavdır. Öğretmen maaşlarının yüksek olması, daha iyi sınav sonuçları ile ilişkilendirilebilir mi? Eğer öyleyse, sınav performansını artırmak için maaşlar düzenlenmeli midir? Aşağıdaki grafik, bu verilerin bir dağılım grafiğini ve bir doğrusal regresyon modelini göstermektedir:

```
library(mdsr)
SAT_2010 <- SAT_2010 |>
  mutate(Salary = salary / 1000)

SAT_plot <- ggplot(data = SAT_2010, aes(x = Salary, y = total)) +
  geom_point() +
  geom_smooth(method = "lm") +
  ylab("Ortalama SAT toplam puanı") +
  xlab("Ortalama öğretmen maaşı (bin USD)")
SAT_plot
```

```
`geom_smooth()` using formula = 'y ~ x'
```




### İlk Model: Sadece Maaş

Bir doğrusal regresyon modeli kullanarak, öğretmen maaşı ile SAT toplam puanları arasındaki ilişkiyi analiz edelim:

```r
SAT_mod1 <- lm(total ~ Salary, data = SAT_2010)
summary(SAT_mod1)
```

```
Call:
lm(formula = total ~ Salary, data = SAT_2010)

Residuals:
     Min       1Q   Median       3Q      Max
-239.136  -84.695   -8.943   84.418  218.027

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1871.104    113.141  16.538   <2e-16 ***
```

```
Salary        -5.019      2.048  -2.451   0.0179 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 111.2 on 48 degrees of freedom
Multiple R-squared:  0.1113,    Adjusted R-squared:  0.09273
F-statistic: 6.008 on 1 and 48 DF,  p-value: 0.01793
```

Bu model, öğretmen maaşları (Salary) ile SAT toplam puanları (total) arasındaki ilişkiyi analiz etmek için basit bir doğrusal regresyon kullanmaktadır. Modeli şu şekilde açıklayabiliriz:

$$\text{Predicted Total SAT Score} = 1871.104 - 5.019 \times \text{Salary}$$

Bu denklem, bağımsız değişken (Salary) kullanılarak bağımlı değişkenin (total) tahmin edilmesini sağlar.

**Coefficients (Katsayılar):**

- **Intercept (Sabit terim):** 1871.1041871.1041871.104
  - Maaşın 0 olduğu varsayıldığında (teorik olarak anlamlı olmasa da), SAT toplam puanının başlangıç tahmini.
  - Çok yüksek bir değere sahiptir çünkü maaşın sıfır olduğu bir durum pratikte mevcut değildir.
- **Salary:** $-5.019$
  - Maaş (bin USD cinsinden) her bir birim arttığında, SAT toplam puanında ortalama $-5.019$ birim azalma beklenir.
  - Negatif bir ilişki olduğunu gösterir; yani, daha yüksek öğretmen maaşları, daha düşük ortalama SAT puanları ile ilişkilidir.

**İstatistiksel Önem ve Anlam:**

- **Salary'nin p-değeri:** 0.01790
  - $p<0.05$, bu nedenle maaş katsayısı istatistiksel olarak anlamlıdır. Bu, maaş ile SAT toplam puanları arasında anlamlı bir ilişki olduğunu gösterir.
- **Intercept'in p-değeri:** $<2e-16$
  - Bu, sabit terimin de istatistiksel olarak anlamlı olduğunu gösterir.

**Artıklar (Residuals):**

- Artıklar, gözlemlenen değerler ile model tarafından tahmin edilen değerler arasındaki farklardır. Artıkların medyanı $-8.943$ ve bu, modelin tahminlerinde hafif bir negatif sapma olduğunu gösterir.
- Minimum artığın $-239.136$, maksimum artığın ise 218.027 olması, modelin bazı durumlarda büyük tahmin hataları yaptığını gösterir.

**Model Uygunluğu:**

- **Residual Standard Error (Artıkların Standart Hatası):** 111.2
  - Bu, modelin tahminlerinin ne kadar değişken olduğunu gösterir. Ortalama olarak, modelin tahminleri gerçek değerlerden yaklaşık 111.2 puan sapar.
- **R-squared (R-Kare):** 0.1113
  - Maaş değişkeni, SAT toplam puanlarındaki değişimin yalnızca %11.13'ünü açıklamaktadır. Bu, modelin açıklayıcı gücünün oldukça düşük olduğunu gösterir.
- **Adjusted R-squared (Düzeltilmiş R-Kare):** 0.09273
  - Açıklayıcı gücü, modeldeki değişken sayısına göre düzeltilmiş bir ölçüdür. Bu da düşük kalmaktadır.

**Genel Değerlendirme:**

Bu model, maaş ile SAT toplam puanları arasında anlamlı bir ilişki olduğunu gösterse de, ilişki negatif ve açıklayıcı gücü düşüktür. Bu, başka bir değişkenin (örneğin SAT sınavına katılım yüzdesi) modelde eksik olabileceğine işaret eder. Daha doğru bir analiz için ek değişkenler eklenmelidir.

**Örnek 4:**

Bu alıştırmada, İsveç Motor Sigortası veri seti kullanılacaktır. Veri seti, İsveç'teki Motor Sigortasında Risk Primi Analizi Komitesi tarafından hazırlanmıştır ve 1977 yılına ait üçüncü taraf motor sigortası taleplerini içermektedir. İlgi çekici sonuçlar, taleplerin sayısı (frekans) ve ödemelerin toplamı (şiddet) olup, İsveç kronu (SEK) cinsindendir. Sonuçlar, bir aracın sürüş mesafesine göre 5 kategori, 7 coğrafi bölge, sürücünün son talep deneyimine göre 7 kategori ve 9 araç türü bazında sınıflandırılmıştır.

```r
# Gerekli kütüphaneler
library(here)
```

```
here() starts at /Users/kobain/Desktop/IST2083
```

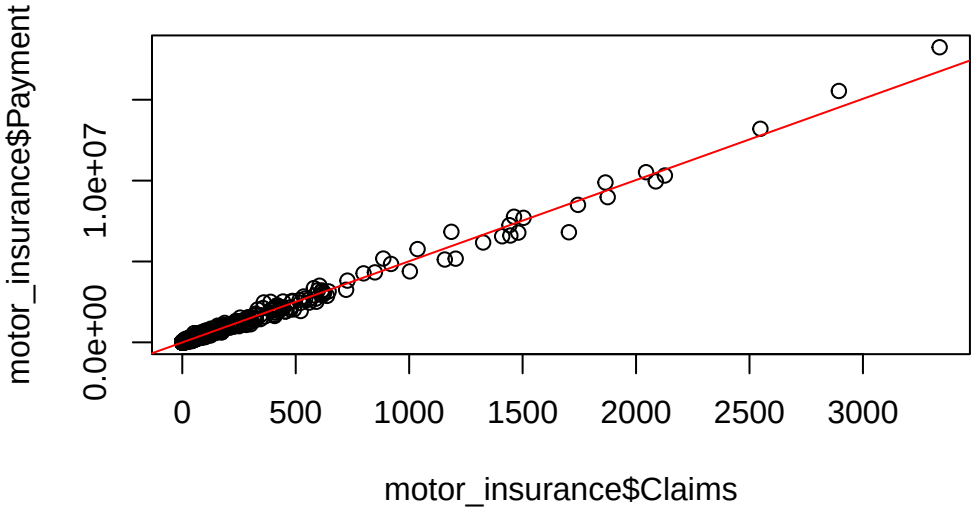
```r
library(fst)
library(tidyverse)

# Veriyi yükleyelim
motor_insurance <- read.csv(here("data", "SwedishMotorInsurance.csv"))

# Veriye göz atalım
str(motor_insurance)
```

```
'data.frame':   2182 obs. of  7 variables:
 $ Kilometres: int  1 1 1 1 1 1 1 1 1 1 ...
 $ Zone      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Bonus     : int  1 1 1 1 1 1 1 1 1 2 ...
 $ Make      : int  1 2 3 4 5 6 7 8 9 1 ...
 $ Insured   : num  455.1 69.2 72.9 1292.4 191 ...
 $ Claims    : int  108 19 13 124 40 57 23 14 1704 45 ...
 $ Payment   : int  392491 46221 15694 422201 119373 170913 56940 77487 6805992 214011 ...
```

```r
# View(motor_insurance)
```

Bu verideki talepler (Claims) ve ödemeler (Payment) için uygun bir model kuralım

```r
reg <- lm(Payment ~ Claims, data = motor_insurance)
plot(motor_insurance$Claims, motor_insurance$Payment)
abline(reg = reg, col = "red")
```




```
summary(reg)
```

```
Call:
lm(formula = Payment ~ Claims, data = motor_insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-1744858    -8545     2773    13386  1491369

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3362.29    2154.79   -1.56    0.119
Claims       5020.08      10.35  485.11   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 97480 on 2180 degrees of freedom
Multiple R-squared:  0.9908,    Adjusted R-squared:  0.9908
F-statistic: 2.353e+05 on 1 and 2180 DF,  p-value: < 2.2e-16
```

Regresyon denklemi şu şekildedir:

$$\text{Payment} = -3362.29 + 5020.08 \times \text{Claims}$$

- **Kesme Noktası (Intercept)**: -3362.29
  Bu, talepler (claims) sıfır olduğunda tahmin edilen ödemedir (payment). Bu negatif bir değer olduğu için gerçek dünyada anlamlı olmayabilir. Örneğin, sigorta şirketinin hiç talep almadığı bir durumda yine de belirli bir ödeme yükümlülüğü bulunmamalı.
- **Eğim (Slope)**: 5020.08
  Bu değer, her ek talebin ortalama 5020.08 SEK'lik bir ödemeyle ilişkilendirildiğini gösterir. Başka bir deyişle, bir talepteki artış, ödemelerde yaklaşık 5020 SEK'lik bir artışa yol açar.
- **Residuals (Artıklar)**:
  - Minimum: -1,744,858
  - Maksimum: 1,491,369
    Artıklar (residuals), gözlemlenen ve model tarafından tahmin edilen değerler arasındaki farktır. Büyük bir aralık olması, bazı gözlemlerde modelin iyi bir performans göstermediğini gösterebilir.
- **Std. Error (Standart Hata)**:
  - Kesme noktası için 2154.79, eğim için 10.35 olarak hesaplanmıştır. Standart hata, tahmin edilen katsayıların ne kadar kesin olduğunu gösterir. Daha küçük standart hata, tahminin daha güvenilir olduğunu ifade eder.
- **t Değeri ve p-Değeri**:
  - **Kesme Noktası (Intercept)**: t değeri -1.56 ve p-değeri 0.119. Bu, kesme noktasının anlamlı olmadığını gösterir ($p > 0.05$).
  - **Eğim (Claims)**: t değeri 485.11 ve p-değeri <2e-16. Bu, talep sayısının (claims) ödeme üzerindeki etkisinin son derece anlamlı olduğunu gösterir ($p < 0.001$).
  - **Residual Standard Error (RSE)**: 97,480
    Bu, ödemelerdeki (payment) gözlemlenen ve model tarafından tahmin edilen değerler arasındaki tipik sapmadır. Büyük bir değer, modelin bazı verilerde iyi performans göstermediğini gösterebilir.
  - **R-Squared ve Adjusted R-Squared**:
    - **R-Squared ($R^2$)**: 0.9908
      Bu, ödeme değişkenindeki varyansın %99.08'inin talepler (claims) ile açıklanabildiğini gösterir. Çok yüksek bir değer, modelin güçlü bir açıklayıcılığı olduğunu gösterir.

    * **Adjusted R-Squared**: 0.9908
      Bu değer, R-Squared'e çok yakın olduğu için modelin açıklayıcılığına eklenen değişken sayısından kaynaklanan bir yanlılık bulunmadığını gösterir.

  – **F-Statistic**: 2.353e+05 ve p-değeri <2.2e-16
    Bu, modelin genel olarak anlamlı olduğunu ve claims değişkeninin payment üzerindeki etkisinin istatistiksel olarak önemli olduğunu gösterir.

### Sonuç

1. **Talepler (Claims)** değişkeni, ödemeler (Payment) üzerinde güçlü ve anlamlı bir etkiye sahiptir.
2. **Kesme noktası (Intercept)** istatistiksel olarak anlamlı değildir, bu da talepler sıfır olduğunda modelin tahminlerinin gerçekçi olmayabileceğini gösterir.
3. Model, ödeme değişkenindeki varyansın büyük bir kısmını açıklayabiliyor (%99.08), bu da modelin genel olarak iyi bir performans sergilediğini gösterir.
4. Ancak, artıklar (residuals) arasında büyük bir aralık olduğu için, modelin bazı uç gözlemlerde kötü performans gösterebileceği göz önünde bulundurulmalıdır.

**Öneri**: Kesme noktasının negatif bir değer olduğu düşünüldüğünde, modelin bu kısmı daha dikkatlice incelenmeli veya bağlama uygun bir alternatif model düşünülebilir.