Regresyon_Devam

Hakan Mehmetcik

2024-12-26

Regresyon II

Hatırlayalım: Regresyon Analizi Nedir?

Regresyon analizi, bir **bağımlı değişken** ile bir veya daha fazla **bağımsız değişken** arasındaki ilişkiyi incelemek ve bu ilişkiyi bir matematiksel modelle açıklamak için kullanılan bir istatistiksel yöntemdir. Temel amacı, bağımlı değişkenin değerini tahmin etmek veya bağımsız değişkenlerin bağımlı değişken üzerindeki etkisini değerlendirmektir.

Doğrusal (Linear) Regresyon Nedir?

Doğrusal regresyon, bağımlı değişken ile bağımsız değişken(ler) arasındaki ilişkinin doğrusal olduğunu varsayar. Matematiksel olarak model:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

şeklinde ifade edilir. Burada:

- Y: Bağımlı değişken
- X: Bağımsız değişken
- 0: Kesme noktası (Intercept, X=0 olduğunda Y'nin değeri)
- 1: Eğim katsayısı (Bağımsız değişkende bir birimlik değişiklik olduğunda Y'deki değişim)
- : Hata terimi (modelin açıklayamadığı rastgele değişkenlik)

> **i** R'da Doğrusal Regresyon
>
> R'da doğrusal regresyon yapmak için `lm()` fonksiyonu kullanılır.

Modeli tanımlama

model <- lm(Y ~ X, data = dataset)

Model özetini görüntüleme

summary(model)

`summary()` Fonksiyonunun Ürettiği Sonuçlar

1. **Call:**
   - Modelin formülü ve kullanılan veri seti hakkında bilgi verir.
2. **Residuals:**
   - Modeldeki tahmin hatalarının (residuals) özet istatistiklerini (min, 1. çeyrek, medyan, 3. çeyrek, max) gösterir. Hataların küçük ve dengeli olması modelin iyi bir uyum sağladığını gösterir.
3. **Coefficients (Katsayılar):**
   - **Estimate:** Her bir değişken için tahmini katsayıyı gösterir.
     - **Intercept (** 0): Bağımlı değişkenin X=0 olduğundaki tahmini değeri.
     - **Slope (** 1): Bağımsız değişkende bir birimlik artış olduğunda bağımlı değişkendeki tahmini değişim.
   - **Std. Error:** Tahminlerin standart hatası.
   - **t value:** Katsayıların 0'dan farklı olup olmadığını test eden t-istatistiği.
   - **Pr(>|t|):** Katsayının anlamlı olup olmadığını belirten p-değeri (<0.05 ise anlamlıdır).
4. **Residual Standard Error (Hata Terimlerinin Standart Hatası):**
   - Modelin hata terimlerinin standart sapmasını ifade eder. Daha düşük bir değer daha iyi bir uyum anlamına gelir.
5. **R-squared (R^2):**
   - Bağımlı değişkenin toplam varyansının ne kadarının model tarafından açıklandığını gösterir. 1'e yakın bir değer modelin iyi bir uyum sağladığını ifade eder.
6. **Adjusted R-squared:**

- R^2'nin düzeltilmiş hali. Özellikle çoklu bağımsız değişkenler olduğunda daha güvenilir bir ölçüttür.

7. **F-statistic:**
    - Modelin genel anlamlılığını test eder. P-değeri <0.05 ise model genel olarak anlamlıdır.

Bu hafta regresyon üzerine farklı örnekler yapacağız. Ancak her bir örnekte, veri setine yönelik farklı soruların cevaplarını da araştıracağız. Bu yöntemle, şimdiye kadar öğrendiğimiz çeşitli teknik ve analizleri regresyon analiziyle birleştirmiş olacağız. Ayrıca, bu haftanın bir diğer konu başlığı, kategorik değişkenlerin regresyon analizinde nasıl kullanılacağı olacak.

Örnek 1:

Bu alıştırma için, dünya rekoru kırılan atletizm (örneğin 100 metre koşu) ve saha (örneğin gülle atma) etkinliklerinin dünya rekor sürelerini ve mesafelerini içeren bir veri setimiz var. Bu veri seti, rekoru kıran kişi, milliyeti, rekorun kırıldığı yer ve kırıldığı yıl gibi bilgileri içeriyor. Unutmayın ki tüm dünya rekorları Olimpiyatlarda kırılmamıştır; birçok rekor bölgesel veya ulusal yarışmalarda gerçekleşmiştir.

```
# gerekli paketler
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- 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 beco
```

```
library(here)
```

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

```r
# veri setini R'a ekle
wr <- read_csv(here("data", "worldrecord.csv"))
```

```
New names:
Rows: 285 Columns: 8
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(5): Event, Type, Athlete, Nationality, Location dbl (3): ...1, Record, Year
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
* `` -> `...1`
```

Veriyi İnceleme

Soru 1:

Veri setinde kaç farklı etkinlik türü (“Erkekler 100m”, “Kadınlar gülle atma” vb.) bulunmaktadır?

```r
# Kaç farklı etkinlik olduğunu görmek için
wr %>%
  count(Event) %>%
  arrange(desc(n))
```

```
# A tibble: 10 x 2
   Event               n
   <chr>           <int>
 1 Mens Polevault     55
 2 Womens Shotput     41
 3 Mens Shotput       39
 4 Mens Mile          32
 5 Womens 800m        29
 6 Mens TripleJump    25
 7 Mens 800m          24
 8 Mens 100m          17
 9 Womens Mile        13
10 Womens 100m        10
```

Soru 2:

Usain Bolt erkekler 100 metre koşusunda ilk dünya rekorunu hangi yıl kırdı?

```
wr %>%
  filter(Athlete == "Usain Bolt") %>%
  select(Event, Year, Record)
```

```
# A tibble: 3 x 3
  Event      Year Record
  <chr>     <dbl>  <dbl>
1 Mens 100m  2008   9.72
2 Mens 100m  2008   9.69
3 Mens 100m  2009   9.58
```

Soru 3:

Kadınlar 1 mil dünya rekorunu 260 saniyenin altında bir sürede ilk kez kıran kadın kimdi?

```
wr %>%
  filter(Event == "Womens Mile", Record < 260) %>%
  arrange(Year) %>%
  slice(1)
```

```
# A tibble: 1 x 8
   ...1 Event       Type  Record Athlete     Nationality         Location  Year
  <dbl> <chr>       <chr>  <dbl> <chr>       <chr>               <chr>    <dbl>
1   146 Womens Mile time    258. Mary Slaney "\xa0United States" Paris     1982
```

Soru 4:

Rekoru belirleyen süre veya mesafeyi hangi değişken gösteriyor? Bu değişkenin türü nedir?

```
wr %>%
  glimpse()
```

```
Rows: 285
Columns: 8
$ ...1        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,~
$ Event       <chr> "Mens 100m", "Mens 100m", "Mens 100m", "Mens 100m", "Mens ~
$ Type        <chr> "time", "time", "time", "time", "time", "time", "time", "t~
```

```
$ Record      <dbl> 10.06, 10.03, 10.02, 9.95, 9.93, 9.92, 9.90, 9.86, 9.85, 9~
$ Athlete     <chr> "Bob Hayes", "Jim Hines", "Charles Greene", "Jim Hines", "~
$ Nationality <chr> "United States", "United States", "United States", "United~
$ Location    <chr> "Tokyo, Japan", "Sacramento, USA", "Mexico City, Mexico", ~
$ Year        <dbl> 1964, 1968, 1968, 1968, 1983, 1988, 1991, 1991, 1994, 1996~
```

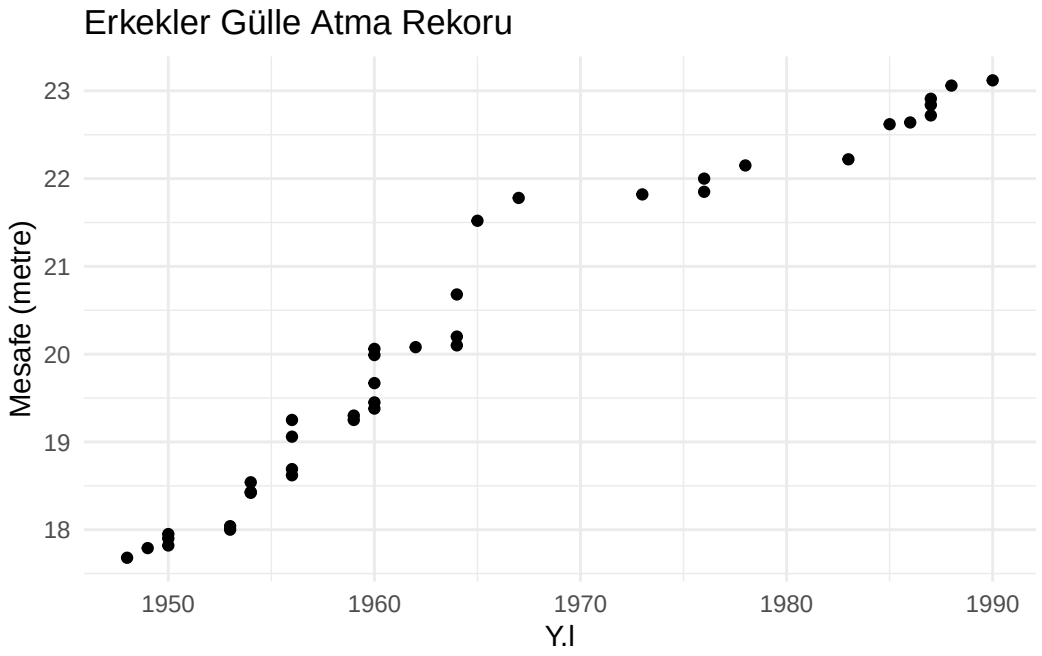
Soru 5:

Kadınlar ve erkekler için gülle atma rekoru mesafesi ile yıl arasındaki dağılımı gösteren bir scatterplot (dağılım grafiği) oluşturalım. Neden?

Cevap: Scatterplot, bu iki sayısal değişken arasında doğrusal bir ilişki olup olmadığını görmemizi sağlar.

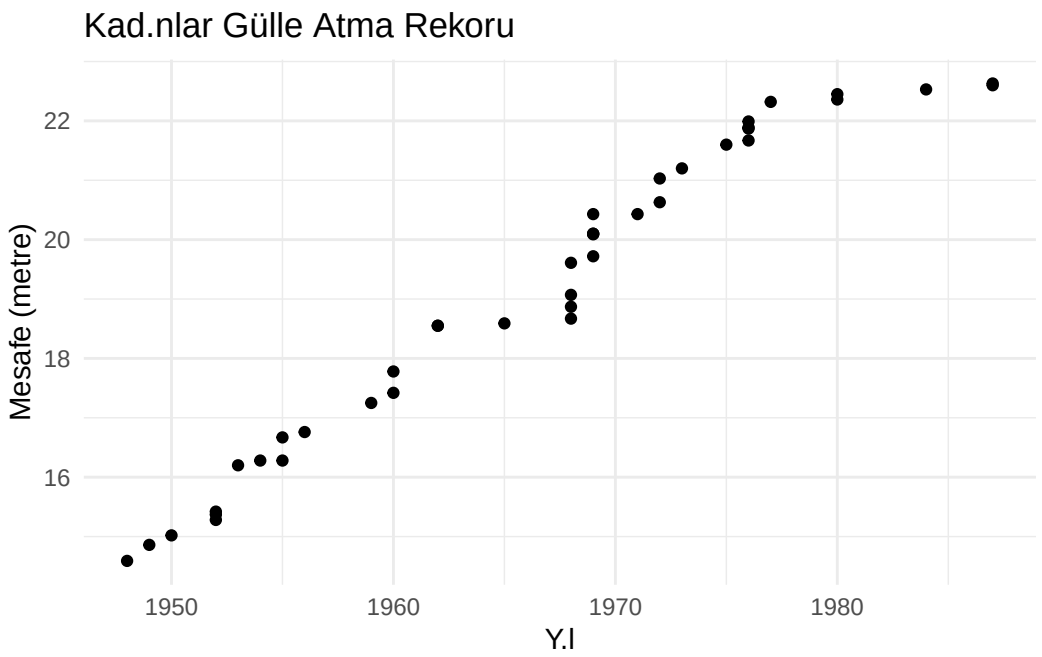
Erkeler için gülle atma rekoru ile yıl arasındaki ilişkiyi gösteren scatterplot

```
mens_shotput <- wr %>%
  filter(Event == "Mens Shotput")

mens_shotput %>%
  ggplot(aes(x = Year, y = Record)) +
  geom_point() +
  labs(title = "Erkekler Gülle Atma Rekoru", x = "Yıl", y = "Mesafe (metre)") +
  theme_minimal()
```



Kadınlar için gülle atma rekoru ile yıl arasındaki ilişkiyi gösteren scatterplo

```
womens_shotput <- wr %>%
  filter(Event == "Womens Shotput")

womens_shotput %>%
  ggplot(aes(x = Year, y = Record)) +
  geom_point() +
  labs(title = "Kadınlar Gülle Atma Rekoru", x = "Yıl", y = "Mesafe (metre)") +
  theme_minimal()
```



Scatterplot Analizi: Erkekler ve kadınlar için rekor mesafesi ile yıl arasında pozitif doğrusal bir ilişki görülmektedir.

Soru 7:

Bu gülle atma dünya rekoru verilerine bir doğrusal model (linear model) oluşturduğumuzda neyi belirleyebiliriz? Erkekler için dünya rekoru mesafesini tahmin eden model denklemi nedir?

```
lm_mdl_women <- lm(Record~Year, data = mens_shotput)
summary(lm_mdl_women)
```

```
Call:
lm(formula = Record ~ Year, data = mens_shotput)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5718 -0.2851 -0.1659  0.2096  1.3389

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.433e+02  1.084e+01  -22.45   <2e-16 ***
Year         1.341e-01  5.515e-03   24.32   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4499 on 37 degrees of freedom
Multiple R-squared:  0.9411,    Adjusted R-squared:  0.9395
F-statistic: 591.2 on 1 and 37 DF,  p-value: < 2.2e-16
```

Intercept Yorumlaması: Yıl sıfırdan başlamadığı için intercept (kesişme noktası) anlamlı değildir.

```r
mens_shotput <- mens_shotput %>%
  mutate(Year = Year - min(Year))  # Yılı sıfırdan başlatma

lm_mdl_men <- lm(Record ~ Year, data = mens_shotput)
summary(lm_mdl_men)
```

```
Call:
lm(formula = Record ~ Year, data = mens_shotput)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5718 -0.2851 -0.1659  0.2096  1.3389

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.901305   0.118691  150.82   <2e-16 ***
Year         0.134107   0.005515   24.32   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.4499 on 37 degrees of freedom
Multiple R-squared:  0.9411,    Adjusted R-squared:  0.9395
F-statistic: 591.2 on 1 and 37 DF,  p-value: < 2.2e-16
```

Kadınlar için dünya rekoru mesafesini tahmin eden model denklemi nedir?

```r
womens_shotput <- womens_shotput %>%
  mutate(Year = Year - min(Year))  # Yılı sıfırdan başlatma

lm_mdl_women <- lm(Record ~ Year, data = womens_shotput)
summary(lm_mdl_women)
```

```
Call:
lm(formula = Record ~ Year, data = womens_shotput)

Residuals:
     Min       1Q   Median       3Q      Max
-1.34968 -0.24704  0.09981  0.35615  0.70689

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.837040   0.157671   94.10   <2e-16 ***
Year         0.233657   0.007409   31.54   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5061 on 39 degrees of freedom
Multiple R-squared:  0.9623,    Adjusted R-squared:  0.9613
F-statistic: 994.6 on 1 and 39 DF,  p-value: < 2.2e-16
```

Sonuç

Erkekler ve kadınlar gülle atma dünya rekorları için dağılım grafikleri, bu etkinliklerin yıllar boyunca güçlü ve pozitif bir doğrusal ilişki gösterdiğini ortaya koymaktadır. Erkeklerde rekor mesafesi yılda ortalama 0.13 metre, kadınlarda ise 0.23 metre artmaktadır. Ancak, başlangıç yılına bağlı olan intercept (kesişme noktası), sorunun bağlamına göre doğrudan yorumlanamaz. Bu nedenle, yılı yeniden düzenleyerek modeller oluşturulmuş ve verilerle iyi bir uyum gösterilmiştir. Fakat çarpıcı bir biçimde, modelimize göre kadınlar için yıllık değişim oranı erkeklere göre daha büyüktür ve bu sonuç istatistiksel olarak anlamlıdır.

Örnek 2:

Araştırma Sorusu: *"Erkekler ve kadınlar için 1 mil dünya rekoru süreleri yıllar içinde nasıl değişti?"*

Soru 1: Veri modeline geçmeden önce ilk ne yapılmalı?

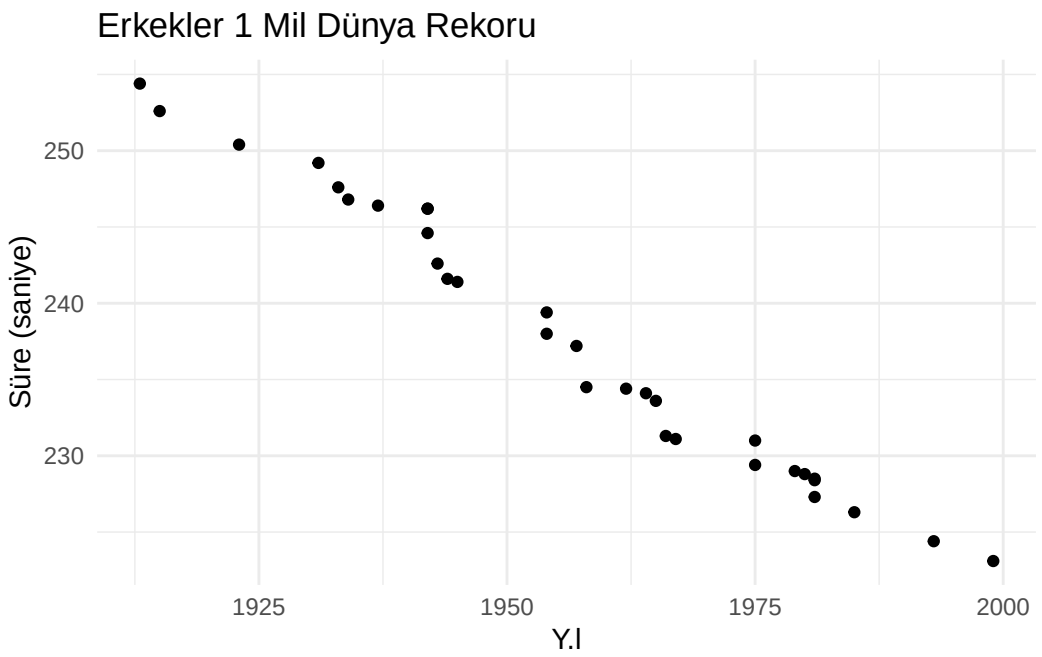
Cevap: İlgi duyulan iki değişkenin scatterplot'unu (dağılım grafiği) oluşturun.

Soru 2: Bir doğrusal modelde, bağımlı değişkendeki varyansın ne kadarı bağımsız değişkenle açıklanabilir?

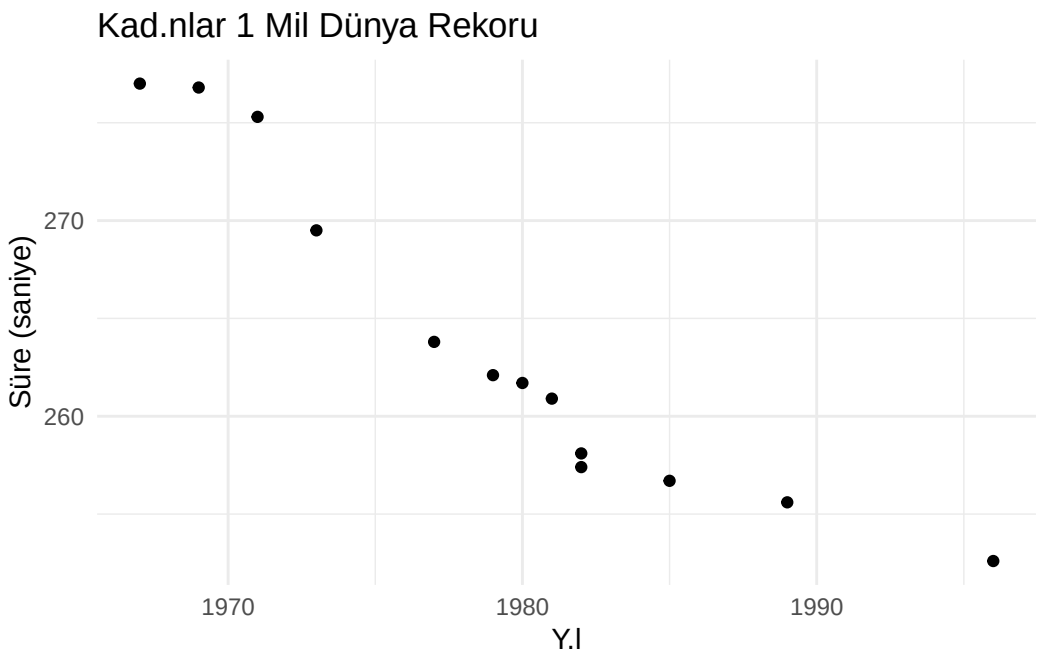
Cevap: R-kare değeri bu bilgiyi verir.

Soru 3: Hangi scatterplot daha güçlü bir doğrusal ilişki gösteriyor?

```
library(tidyverse)

# Erkekler ve kadınlar için veri alt kümeleri
mens_mile <- wr %>% filter(Event == "Mens Mile")
womens_mile <- wr %>% filter(Event == "Womens Mile")

# Erkekler için scatterplot
mens_mile %>%
  ggplot(aes(x = Year, y = Record)) +
  geom_point() +
  labs(title = "Erkekler 1 Mil Dünya Rekoru", x = "Yıl", y = "Süre (saniye)") +
  theme_minimal()
```




```
# Kadınlar için scatterplot
womens_mile %>%
  ggplot(aes(x = Year, y = Record)) +
  geom_point() +
  labs(title = "Kadınlar 1 Mil Dünya Rekoru", x = "Yıl", y = "Süre (saniye)") +
  theme_minimal()
```



Soru 4: Erkekler için dünya rekoru süresi her yıl ortalama kaç saniye azalıyor?

```
men_model <- lm(Record ~ Year, data = mens_mile)
summary(men_model)
```

```
Call:
lm(formula = Record ~ Year, data = mens_mile)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6033 -0.7175 -0.1860  0.7278  2.8533

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 1007.47075   21.35031   47.19   <2e-16 ***
Year          -0.39347    0.01091  -36.07   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.371 on 30 degrees of freedom
Multiple R-squared:  0.9775,    Adjusted R-squared:  0.9767
```

```
F-statistic:  1301 on 1 and 30 DF,  p-value: < 2.2e-16
```

Soru 5: Kadınlar için dünya rekoru süresi her yıl ortalama kaç saniye azalıyor?

```r
women_model <- lm(Record ~ Year, data = womens_mile)
summary(women_model)
```

```
Call:
lm(formula = Record ~ Year, data = womens_mile)

Residuals:
   Min     1Q Median     3Q    Max
-3.635 -1.853 -1.107  1.376  5.186

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2189.2834   198.2509  11.043 2.72e-07 ***
Year          -0.9729     0.1002  -9.713 9.88e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.824 on 11 degrees of freedom
Multiple R-squared:  0.8956,    Adjusted R-squared:  0.8861
F-statistic: 94.35 on 1 and 11 DF,  p-value: 9.88e-07
```

Soru 6: Erkeklerin rekor süresinin 1 saniye azalması için kaç yıl geçmesi gerekir?

```r
1 / abs(coef(men_model)["Year"])
```

```
    Year
2.541472
```

Soru 7: Kadınların rekor süresinin 1 saniye azalması için kaç yıl geçmesi gerekir?

```r
1 / abs(coef(women_model)["Year"])
```

```
    Year
1.027876
```

Soru 8: Erkekler ve kadınlar için dünya rekoru sürelerinin varyansının ne kadarı yıl ile açıklanabilir?

Cevap:

- Erkekler: Adjusted R-squared: 0.9767
- Kadınlar: Adjusted R-squared: 0.8861

Çıkarım:

Erkekler ve kadınlar için 1 mil dünya rekoru etkinliklerine ait scatterplot'lara (dağılım grafikleri) bakıldığında, her iki etkinliğin de zamanla güçlü ve negatif bir ilişki izlediği görülmektedir. Her iki grup için de doğrusal (lineer) ilişki varsayımı sağlanmış görünmektedir.

- Erkekler için 1 mil dünya rekoru süresi yılda ortalama 0.393 saniye azalırken, kadınlar için bu süre yılda ortalama 0.976 saniye azalmaktadır.
- **Kesişme noktası (intercept)** tahmini, yıl sıfıra eşit olduğunda rekor süresinin değerini ifade ettiğinden, bu bağlamda yorumlanamaz.

Her iki lineer model de verilere iyi bir uyum sağlamaktadır:

- Erkekler için **R-kare değeri** 0.976'dır, yani performanstaki değişimin %97.7'si yıl değişkeniyle açıklanabilmektedir.
- Kadınlar için **R-kare değeri** 0.886'dır, yani performanstaki değişimin %88.6'sı yıl değişkeniyle açıklanabilmektedir.

Bu analiz, erkekler ve kadınlar için dünya rekoru sürelerinin son birkaç on yılda doğrusal olarak azaldığını ortaya koymaktadır.

Örnek 3:

Dünya Rekoru Verileriyle Çalışmaya Devam Ediyoruz: Şimdi, 1970'ten itibaren erkekler sırıkla atlama dünya rekorları için en iyi uyum sağlayan doğrusal modeli bulmak istiyoruz.

Adım 1: İlk olarak, 1970 ve sonrası yıllarda erkekler sırıkla atlama etkinliğindeki dünya rekorlarını içeren bir veri çerçevesi oluşturalım.

```
menspole <- wr %>%
  filter(Event == "Mens Polevault" & Year >= 1970)
```

Soru 1: Erkekler sırıkla atlamadaki güncel dünya rekoru yüksekliği (metre cinsinden) nedir?

```r
max_record <- menspole %>%
  summarise(Max_Record = max(Record))
max_record
```

```
# A tibble: 1 x 1
  Max_Record
       <dbl>
1       6.14
```

Soru 2: Sırıkla atlama rekoru ilk kez 6 metreyi hangi yıl geçti?

```r
menspole %>%
  filter(Record > 6)
```

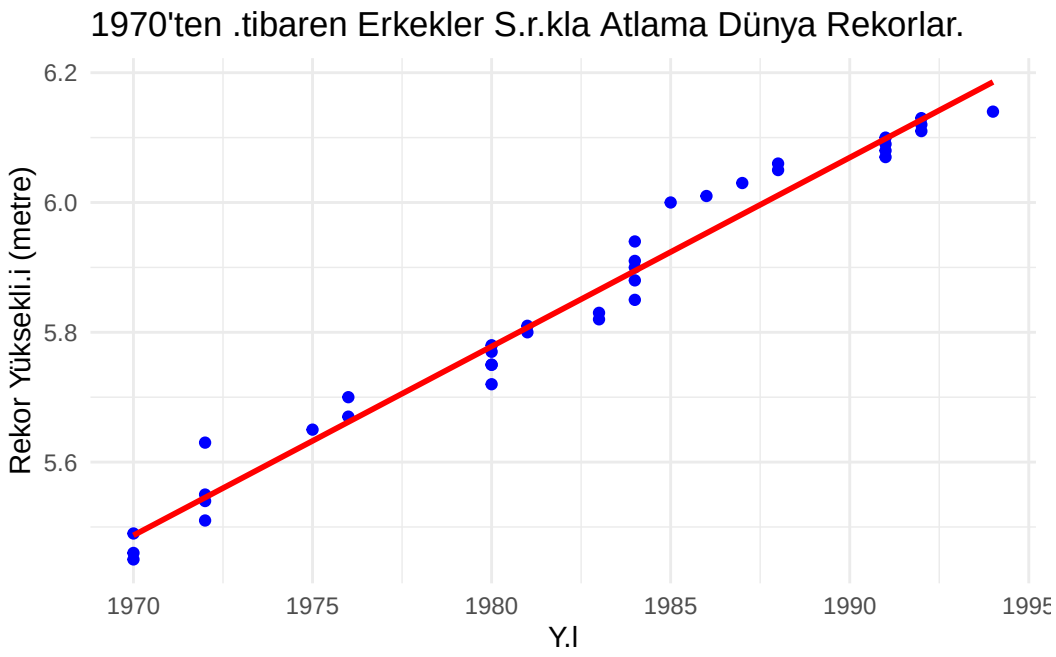
```
# A tibble: 12 x 8
    ...1 Event          Type     Record Athlete      Nationality  Location  Year
   <dbl> <chr>          <chr>     <dbl> <chr>        <chr>        <chr>    <dbl>
 1   194 Mens Polevault distance   6.01 Sergey Bubka "\xa0Soviet~ "Moscow~  1986
 2   195 Mens Polevault distance   6.03 Sergey Bubka "\xa0Soviet~ "Prague~  1987
 3   196 Mens Polevault distance   6.05 Sergey Bubka "\xa0Soviet~ "Bratis~  1988
 4   197 Mens Polevault distance   6.06 Sergey Bubka "\xa0Soviet~ "Nice, ~  1988
 5   198 Mens Polevault distance   6.07 Sergey Bubka "\xa0Soviet~ "Shizuo~  1991
 6   199 Mens Polevault distance   6.08 Sergey Bubka "\xa0Soviet~ "Moscow~  1991
 7   200 Mens Polevault distance   6.09 Sergey Bubka "\xa0Soviet~ "Formia~  1991
 8   201 Mens Polevault distance   6.1  Sergey Bubka "\xa0Soviet~ "Malm\x~  1991
 9   202 Mens Polevault distance   6.11 Sergey Bubka "\xa0Ukrain~ "Dijon,~  1992
10   203 Mens Polevault distance   6.12 Sergey Bubka "\xa0Ukrain~ "Padua,~  1992
11   204 Mens Polevault distance   6.13 Sergey Bubka "\xa0Ukrain~ "Tokyo,~  1992
12   205 Mens Polevault distance   6.14 Sergey Bubka "\xa0Ukrain~ "Sestri~  1994
```

Soru 3: 1970'ten itibaren erkekler sırıkla atlama rekorlarını yıl fonksiyonu olarak gösteren bir scatterplot oluşturun ve verilere doğrusal bir model uydurun.

```r
# Lineer modeli oluşturma
men_mdl_pole <- lm(Record ~ Year, data = menspole)

# Scatterplot ve modelin çizimi
menspole %>%
  ggplot(aes(x = Year, y = Record)) +
  geom_point(color = "blue") +
```

```
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(
    title = "1970'ten İtibaren Erkekler Sırıkla Atlama Dünya Rekorları",
    x = "Yıl",
    y = "Rekor Yüksekliği (metre)"
  ) +
  theme_minimal()
```

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



Soru 4: Rekorun değişimini en iyi şekilde ne açıklar?

Cevap: Dağılım grafiğinde açıkça görüldüğü üzere, sırıkla atlama rekor yüksekliği zamanla sürekli artmıştır. Bunun için rekoru yıl (zaman) bağlamında değerlendirmek (bir model oluşturmak) mantıklı olacaktır.

Soru 5: 1970'ten itibaren erkekler sırıkla atlama dünya rekorunun değişimini tanımlayan doğrusal model için katsayı tahminlerini rapor edin.

```
summary(men_mdl_pole)
```

```
Call:
lm(formula = Record ~ Year, data = menspole)

Residuals:
      Min        1Q    Median        3Q       Max
-0.058171 -0.028171 -0.005313  0.015400  0.084687

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.185e+01  1.625e+00  -31.91   <2e-16 ***
Year         2.911e-02  8.199e-04   35.50   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0356 on 35 degrees of freedom
Multiple R-squared:  0.973, Adjusted R-squared:  0.9722
F-statistic:  1260 on 1 and 35 DF,  p-value: < 2.2e-16
```

Intercept'in negatif olması, yıl değişkeninin başlangıç noktasının modele uygun olmadığını gösterir. Bu nedenle, yılı transpoze ederek modeli tekrar oluşturabiliriz. Yıl değişkenini 1970'ten başlayacak şekilde yeniden düzenleyelim ve ardından doğrusal modeli tekrar çalıştıralım.

```r
# Yılı transpoze et
menspole <- menspole %>%
  mutate(Year_Transposed = Year - 1970)

# Yeni model oluştur
men_mdl_pole_transposed <- lm(Record ~ Year_Transposed, data = menspole)

# Modelin özetini incele
summary(men_mdl_pole_transposed)
```

```
Call:
lm(formula = Record ~ Year_Transposed, data = menspole)

Residuals:
      Min        1Q    Median        3Q       Max
-0.058171 -0.028171 -0.005313  0.015400  0.084687

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)     5.4870986  0.0115819   473.8   <2e-16 ***
Year_Transposed 0.0291072  0.0008199    35.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0356 on 35 degrees of freedom
Multiple R-squared:  0.973, Adjusted R-squared:  0.9722
F-statistic:  1260 on 1 and 35 DF,  p-value: < 2.2e-16
```

Modelin Yorumu:

1. Kesme Noktası (Intercept):

- **Değer:** 5.4871
- Bu değer, **1970 yılında erkekler sırıkla atlama dünya rekorunun** yüksekliğini temsil eder ve anlamlı bir sonuçtur.
- 1970 yılında dünya rekoru yaklaşık **5.49 metre** olarak tahmin edilmiştir.

2. Eğim (Slope):

- **Değer:** 0.0291
- Bu katsayı, **1970'ten sonra geçen her yıl için sırıkla atlama dünya rekorunun ortalama 0.029 metre (yaklaşık 2.91 cm) arttığını** ifade eder.
- Bu artış oranı, rekorların zamanla düzenli bir şekilde geliştiğini göstermektedir.

3. Model Uyumunun Gücü (R-squared):

- **Multiple R-squared:** 0.973
- **Adjusted R-squared:** 0.9722
- Model, rekor yüksekliğindeki varyansın %97.3'ünü açıklamaktadır. Bu da modelin oldukça güçlü bir uyum sağladığını gösterir.

4. F-Testi ve p-Değeri:

- **F-statistic:** 1260
- **p-value:** $<$ 2.2e-16
- Modelin geneli istatistiksel olarak anlamlıdır. Bu, "Yıl" değişkeninin rekor yüksekliğini açıklamada önemli bir faktör olduğunu ifade eder.

Kategorik Açıklayıcı Değişkenler

Kategorik değişkenler, regresyon analizinde özel bir dikkat gerektirir. Bu değişkenler, sürekli ya da iki kategorili değişkenler gibi doğrudan regresyon denklemine eklenemez. Bunun yerine, regresyon modeline dahil edilebilmeleri için farklı bir kodlama sistemi kullanılarak bir dizi değişkene dönüştürülmeleri gerekir. Kodlama sistemleri farklı olabilir, ancak seçilen kodlama sistemi değişkenin genel etkisini değiştirmez.

Örnek 4:

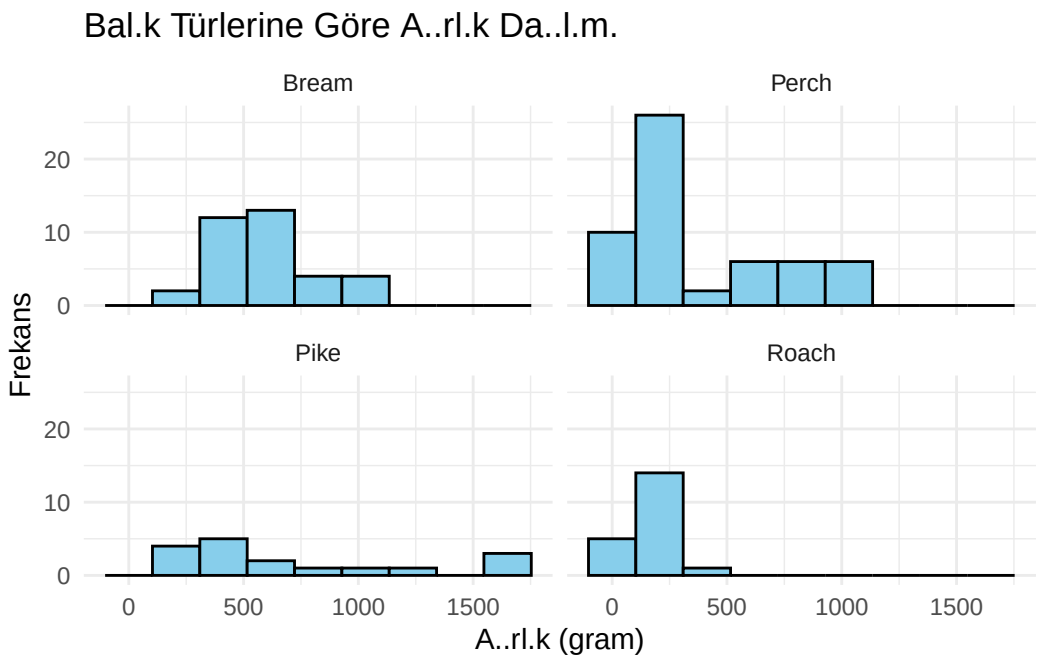
Balık veri seti, yedi yaygın balık türünü ve bu balıkların ağırlıkları gibi bilgilerle bir tahmin modeli geliştirmek için kullanılabilir. Amaç, balık türlerine göre ağırlık tahmini yapmaktır.

```
fish <- read.csv(here("data", "Fish.csv"))
species <- c("Bream", "Perch", "Pike", "Roach")
fish <- fish[fish$Species %in% species,]
head(fish)
```

```
  Species Weight Length1 Length2 Length3  Height  Width
1   Bream    242    23.2    25.4    30.0 11.5200 4.0200
2   Bream    290    24.0    26.3    31.2 12.4800 4.3056
3   Bream    340    23.9    26.5    31.1 12.3778 4.6961
4   Bream    363    26.3    29.0    33.5 12.7300 4.4555
5   Bream    430    26.5    29.0    34.0 12.4440 5.1340
6   Bream    450    26.8    29.7    34.7 13.6024 4.9274
```

Soru 1: Dağılım Grafiği Oluşturma

```
ggplot(data = fish, aes(x = Weight)) +
  geom_histogram(bins = 9, fill = "skyblue", color = "black") +
  facet_wrap(~Species) +
  labs(title = "Balık Türlerine Göre Ağırlık Dağılımı",
       x = "Ağırlık (gram)", y = "Frekans") +
  theme_minimal()
```



Soru 2: Türlere Göre Ortalama Ağırlık

```
fish |>
  group_by(Species) |>
  summarise(mean_weight = mean(Weight, na.rm = TRUE)) |>
  arrange(desc(mean_weight))
```

```
# A tibble: 4 x 2
  Species mean_weight
  <chr>         <dbl>
1 Pike           719.
2 Bream          618.
3 Perch          382.
4 Roach          152.
```

Soru 3: Türlere Göre Ağırlık Değişimini Açıklayan Modelin Katsayıları

```
model_with_intercept <- lm(Weight ~ Species, data = fish)
summary(model_with_intercept)
```

```
Call:
lm(formula = Weight ~ Species, data = fish)

Residuals:
    Min      1Q  Median      3Q     Max
-518.71 -239.74  -78.14  133.31  931.29

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    617.83      53.00  11.657  < 2e-16 ***
SpeciesPerch  -235.59      67.56  -3.487 0.000676 ***
SpeciesPike    100.88      92.69   1.088 0.278579
SpeciesRoach  -465.78      87.89  -5.300 5.13e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 313.6 on 124 degrees of freedom
Multiple R-squared:  0.2581,    Adjusted R-squared:  0.2402
F-statistic: 14.38 on 3 and 124 DF,  p-value: 4.231e-08
```

Soru 4: Modelin Problemi

Bu modelde, intercept değeri **Bream** türü için raporlanmıştır ve diğer türler için negatif ağırlık değerleri üretmektedir. Bu durum anlamlı değildir. Çözüm bir sıfır kesişimli model üretmektir:

```r
model_no_intercept <- lm(Weight ~ Species + 0, data = fish)
summary(model_no_intercept)
```

```
Call:
lm(formula = Weight ~ Species + 0, data = fish)

Residuals:
    Min      1Q  Median      3Q     Max
-518.71 -239.74  -78.14  133.31  931.29

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
SpeciesBream   617.83      53.00  11.657  < 2e-16 ***
SpeciesPerch   382.24      41.90   9.123 1.66e-15 ***
```

```
SpeciesPike    718.71      76.05   9.451 2.70e-16 ***
SpeciesRoach   152.05      70.11   2.169    0.032 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 313.6 on 124 degrees of freedom
Multiple R-squared:  0.7163,    Adjusted R-squared:  0.7072
F-statistic: 78.28 on 4 and 124 DF,  p-value: < 2.2e-16
```

Her bir balık türü için katsayılar (coefficients) tahmini ağırlıkları verir. Bu modelde kesme noktası (intercept) sıfırdır, bu nedenle her türün ağırlığı bağımsız olarak raporlanmıştır:

- **SpeciesBream (617.83 gram)**: Levrek türü için tahmini ağırlık. Bu katsayı, yüksek bir t-değerine (11.657) ve çok düşük bir p-değerine (< 2e-16) sahip, bu da tahminin oldukça anlamlı olduğunu gösteriyor.
- **SpeciesPerch (382.24 gram)**: Sazan türü için tahmini ağırlık. Anlamlı bir tahmin (p-değeri 1.66e-15).
- **SpeciesPike (718.71 gram)**: Turna balığı için tahmini ağırlık. Bu da oldukça anlamlı bir tahmindir (p-değeri 2.70e-16).
- **SpeciesRoach (152.05 gram)**: Kızılkanat balığı için tahmini ağırlık. Anlamlılık seviyesi daha düşüktür (p-değeri 0.032), ancak %5 anlamlılık düzeyinde istatistiksel olarak anlamlıdır.

Model, türlere göre ağırlık tahmin etmekte oldukça başarılıdır (R-Square: %71.63 varyans açıklanıyor).

Örnek 5:

Regresyon, açıklayıcı değişkenlerin bilinen değerlerinden bir yanıt değişkeninin değerlerini tahmin etmenizi sağlar. Hangi değişkeni yanıt değişkeni olarak kullanacağınız, yanıtlamak istediğiniz soruya bağlıdır, ancak birçok veri setinde tahmin edilmesi ilginç olacak değişkenler için açık bir seçim bulunur. Bu alıştırmalar boyunca, Tayvan emlak veri setini keşfedeceksiniz. Bu veri seti, Yeni Taipei Şehri, Tayvan'ın Sindian Bölgesi'nden toplanan bir gayrimenkul değerleme tarihçesine dayanmaktadır. Bu veri setinde, gayrimenkul değerleme bir regresyon problemidir.

```
library(fst)
taiwan_real_estate <- read.fst(here("data", "taiwan_real_estate.fst"))
str(taiwan_real_estate)
```

```
'data.frame':   414 obs. of  4 variables:
 $ dist_to_mrt_m  : num  84.9 306.6 562 562 390.6 ...
 $ n_convenience  : num  10 9 5 5 5 3 7 6 1 3 ...
 $ house_age_years: Ord.factor w/ 3 levels "0 to 15"<"15 to 30"<..: 3 2 1 1 1 1 3 2 3 2 ...
 $ price_twd_msq  : num  11.5 12.8 14.3 16.6 13 ...
```

Soru 1:

Veri setini görüntüleyerek hangi değişkenin iyi bir yanıt değişkeni olacağını belirleyin.

```
View(taiwan_real_estate)
```

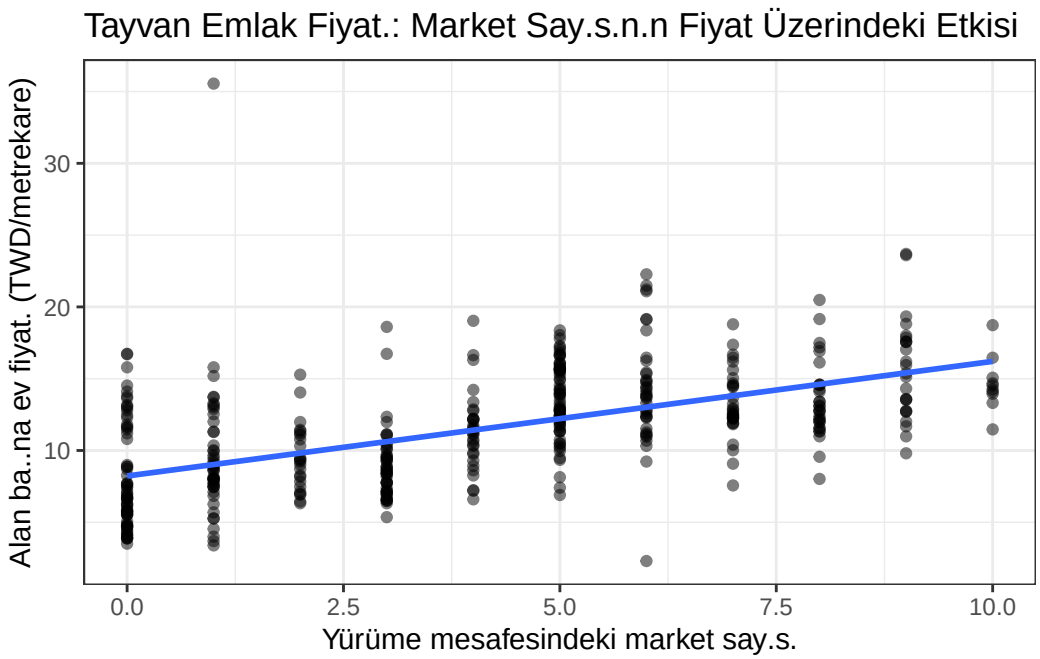
Cevap: Fiyat tahmini, yaygın bir iş görevidir, bu nedenle **house price** (ev fiyatı) iyi bir yanıt değişkenidir. Bu veri setinde, `price_twd_msq` değişkeni yanıt değişkenlerinden biri olabilir.

Soru 2:

Bir istatistiksel model çalıştırmadan önce, veri setini görselleştirmek genellikle iyi bir fikirdir. Burada, ev fiyatı (alan başına fiyat) ile yakınlardaki market (convenience store) sayısı arasındaki ilişkiyi inceleyeceğiz.

```
ggplot(data = taiwan_real_estate, aes(x = n_convenience, y = price_twd_msq)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  scale_x_continuous("Yürüme mesafesindeki market sayısı") +
  scale_y_continuous("Alan başına ev fiyatı (TWD/metrekare)") +
  ggtitle("Tayvan Emlak Fiyatı: Market Sayısının Fiyat Üzerindeki Etkisi") +
  theme_bw()
```

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

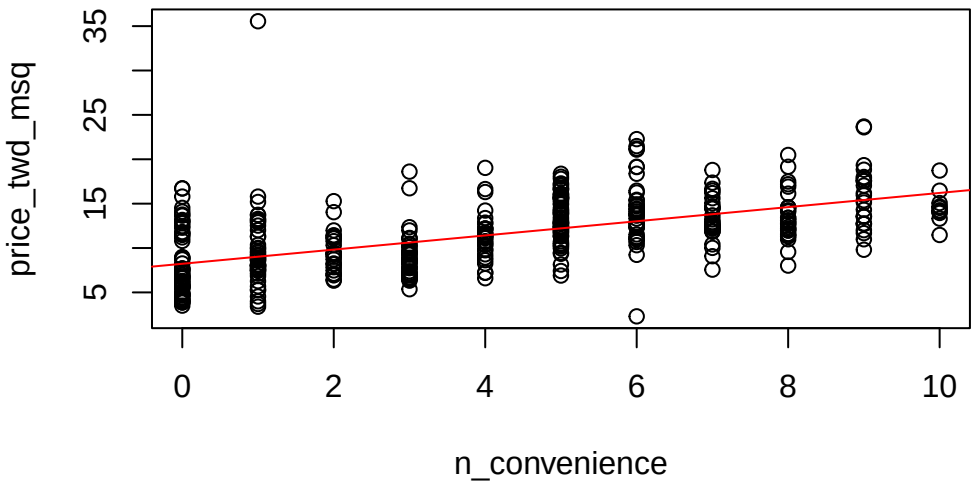


Soru 3:

Doğrusal regresyon modelleri her zaman verilere bir doğru çizgisi uydurur. Doğru çizgileri iki özellikle tanımlanır: kesişim noktası (intercept) ve eğim (slope).

Bu alıştırmada, `price_twd_msq` yanıt değişkeni, `n_convenience` ise açıklayıcı değişken olarak kullanılarak bir doğrusal regresyon modeli çalıştıracağız.

```
taiwan_model <- lm(price_twd_msq ~ n_convenience, data = taiwan_real_estate)
plot(price_twd_msq ~ n_convenience, data = taiwan_real_estate)
abline(reg = taiwan_model, col = "red")
```




```
summary(taiwan_model)
```

```
Call:
lm(formula = price_twd_msq ~ n_convenience, data = taiwan_real_estate)

Residuals:
     Min       1Q   Median       3Q      Max
-10.7132  -2.2213  -0.5409   1.8105  26.5299

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.22424    0.28500   28.86   <2e-16 ***
n_convenience  0.79808    0.05653   14.12   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.384 on 412 degrees of freedom
Multiple R-squared:  0.326, Adjusted R-squared:  0.3244
F-statistic: 199.3 on 1 and 412 DF,  p-value: < 2.2e-16
```

Soru 4:

Modelin `(Intercept)` katsayısı `8.2242` olarak hesaplanmıştır. Bu ne anlama gelir?

Cevap: Ortalama olarak, yakında market bulunmayan bir evin metrekare fiyatı **8.2242 TWD**'dir.

Soru 5:

Modelin `n_convenience` katsayısı `0.7981` olarak hesaplanmıştır. Bu ne anlama gelir?

Cevap: Yakındaki market sayısını bir artırdığınızda, ev fiyatında beklenen artış **0.7981 TWD/metrekare** olur.

Örnek 6:

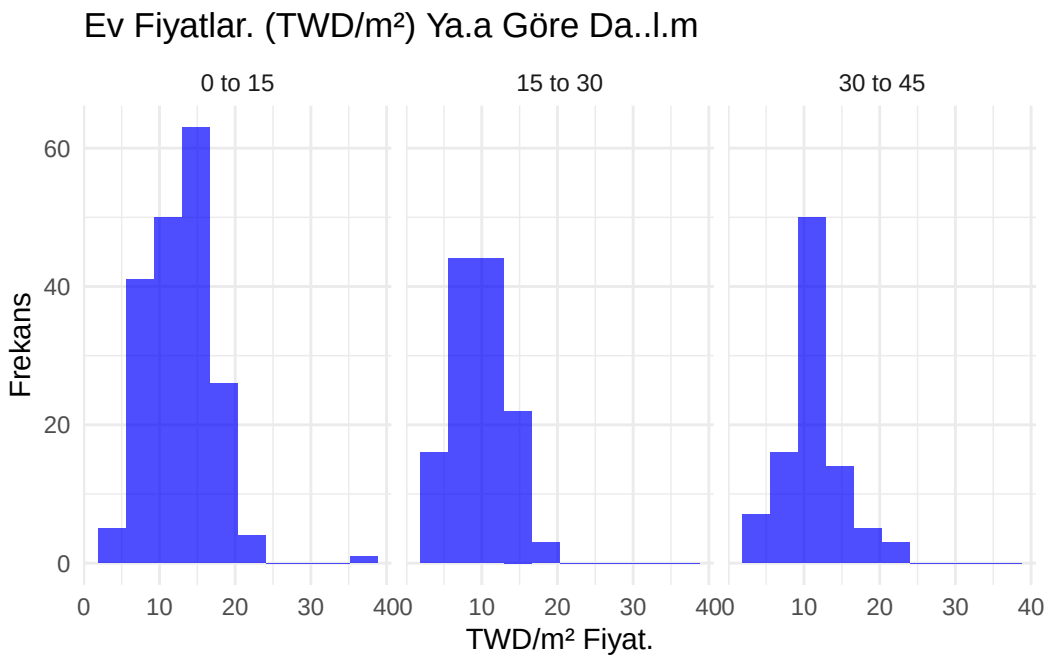
Açıklayıcı değişken kategorik olduğunda, verileri görselleştirmek için daha önce kullanılan dağılım grafiği uygun değildir. Bunun yerine, her kategori için bir çubuk grafiği veya histogram oluşturmak iyi bir seçenektir.

Tayvan emlak veri setinde, her evin yaşını gösteren kategorik bir değişken bulunmaktadır. Yaşlar şu şekilde 3 gruba ayrılmıştır: **0-15 yıl**, **15-30 yıl**, ve **30-45 yıl**.

Soru 1:

Tayvan emlak veri setini kullanarak `price_twd_msq` için bir histogram çizin.

```
# ggplot ile histogram ve kategorik gruplandırma
ggplot(data = taiwan_real_estate, aes(x = price_twd_msq)) +
  geom_histogram(bins = 10, fill = "blue", alpha = 0.7) +
  facet_wrap(~house_age_years) +
  labs(
    title = "Ev Fiyatları (TWD/m²) Yaşa Göre Dağılım",
    x = "TWD/m² Fiyatı",
    y = "Frekans"
  ) +
  theme_minimal()
```



Soru 2:

Ev yaşına (`house_age_years`) göre verileri gruplandırın.

```
taiwan_real_estate |>
  count(house_age_years)
```

```
  house_age_years   n
1         0 to 15 190
2        15 to 30 129
3        30 to 45  95
```

Soru 3:

Her grup için `price_twd_msq` ortalamasını hesaplayın.

```
taiwan_real_estate |>
  group_by(house_age_years) |>
  summarise(mean_price = mean(price_twd_msq, na.rm = TRUE))
```

```
# A tibble: 3 x 2
  house_age_years mean_price
  <ord>                <dbl>
1 0 to 15              12.6
2 15 to 30              9.88
3 30 to 45             11.4
```

Soru 4:

`price_twd_msq`'i yanıt değişkeni, `house_age_years`'i açıklayıcı değişken olarak kullanarak bir doğrusal regresyon modeli çalıştırın. Modeli `mdl_price_vs_age` olarak atayın.

```r
mdl_price_vs_age <- lm(price_twd_msq ~ house_age_years + 0, data = taiwan_real_estate)
summary(mdl_price_vs_age)
```

```
Call:
lm(formula = price_twd_msq ~ house_age_years + 0, data = taiwan_real_estate)

Residuals:
     Min       1Q   Median       3Q      Max
-10.3379  -2.9119   0.2218   2.5544  22.9147

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
house_age_years0 to 15   12.6375     0.2866   44.10   <2e-16 ***
house_age_years15 to 30   9.8767     0.3478   28.40   <2e-16 ***
house_age_years30 to 45  11.3933     0.4053   28.11   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.95 on 411 degrees of freedom
Multiple R-squared:  0.896, Adjusted R-squared:  0.8953
F-statistic:  1180 on 3 and 411 DF,  p-value: < 2.2e-16
```

Sonuç:

- **0-15 yaş** arası evler ortalama olarak **12.635 TWD/m²** maliyete sahiptir.
- **15-30 yaş** arası evler daha ucuz olup ortalama **9.987 TWD/m²** maliyete sahiptir.

- **30 yıldan daha eski evler**, orta yaşlı evlerden daha pahalıdır. Bu durum dikkatle incelenmelidir.

Açıklama: Daha eski evlerin daha büyük olması, fiyatlarının daha yüksek olmasının bir nedeni olabilir. Bu nedenle, ev büyüklüğü Tayvan emlak piyasasında fiyat tahminlerinde daha iyi bir belirleyici olabilir.

Regresyon Analizinde Dikkat Edilmesi Gereken Noktalar

1. **Negatif Kesme Noktası (Intercept):**
    - Kesme noktası (0) negatif olduğunda, bu değerin fiziksel anlamı olup olmadığını değerlendirmek önemlidir. Örneğin, negatif bir fiyat fiziksel olarak anlamlı değildir ve model bu durumda yeniden değerlendirilmelidir.
2. **Kategorik Bağımsız Değişkenler:**
    - Kategorik değişkenler regresyon modeline doğrudan dahil edilemez. Bunun yerine, kategorik değişkenler **dummy değişkenlere** dönüştürülmelidir. Örneğin, "evet/hayır" gibi iki kategorili bir değişken bir 0 ve 1 değeriyle kodlanabilir.
3. **Lineerlik Varsayımı:**
    - Doğrusal regresyon, bağımlı ve bağımsız değişkenler arasındaki ilişkinin doğrusal olduğunu varsayar. Bu varsayımı doğrulamak için grafiksel analiz (dağılım grafikleri) yapılmalıdır.
4. **Ölçüm Birimleri:**
    - Bağımsız değişkenlerin birimleri modelde büyük farklar yaratabilir. Bu nedenle, bazı durumlarda değişkenleri standardize etmek faydalıdır. (**z-skorları hatırlayın!**)
5. **Hata Terimleri:**
    - Hata terimlerinin normal dağılım ve sabit varyans varsayımlarını sağladığından emin olunmalıdır.