LİNEER REGRESYON VE SCİKİT-LEARN

Dr. Öğr. Üyesi Caner Erden, cerden@subu.edu.tr

@canererden /c/CanerErdenn

İçerik

Regresyon Analizi

Lineer Regresyon: Bağımlı bir değişken(Y) ile bağımsız bir değişken (X) serisinin arasındaki ilişkinin fonksiyonel biçimi.

Lineer regresyon analizi: «veriyi lineer bir çizgiye uydurma çalışması» veya «lineer modelleme»

Y: Çıktı, bağımlı, sonuç, etkilenen değişken

X: Girdi, bağımsız, neden(faktör), etkileyen değişken

Regresyon sorusu: Y'nin X'e bağlı koşullarını nasıl tahmin edebiliriz sorusudur.

Basit Lineer Regresyon

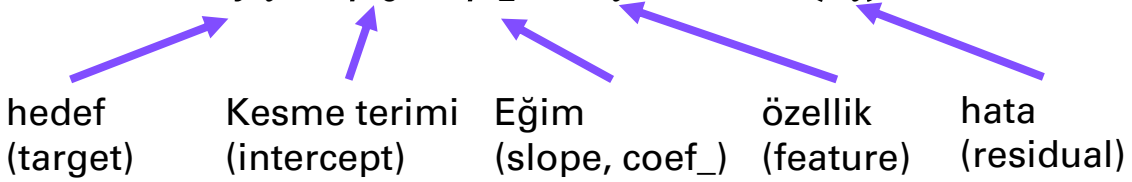
En fazla kullanılan ve en basit regresyon çeşididir.

Bağımlı ve bağımsız değişken arasında düz bir çizgi çizerek aradaki bağıntıyı ortaya çıkarır.

Sadece sürekli verilerde çalışır. Kategorik verilerde çalışmaz.

Basit Lineer Regresyon Denklemi

$$y_i = \beta_0 + \beta_1 \times x_i + hata(e_i)$$



Katsayıların Tahmini

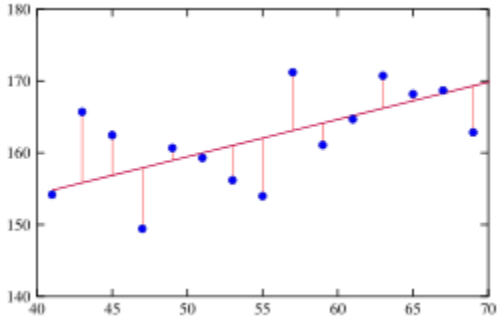
Regresyon ile β_0 ve β_1 parametrelerinin(katsayıların) tahmini yapılır.

Tahmini yapılan denklem:

$$\hat{y} = \beta_0 + \beta_1 \times x$$

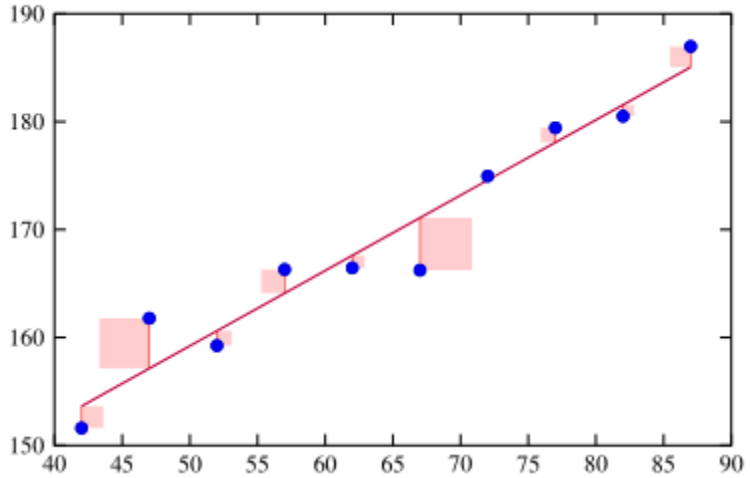
Şapkalı değerler gerçek değerleri değil, tahmini yapılan değerleri gösterir. $\hat{y}$ y'nin tahmini olarak tanımlanır.

Esas amaç hata değerinin minimize edilmesidir.



En Küçük Kareler Metodu

$\beta_0\beta_1$ değerleri gösterilen karelerin alanlarını en aza indirgemek için ayarlanır. Grafikte gösterilen kırmızı alanların en az olduğu zaman regresyon çizgisi ortaya çıkarılmış olur.



En Küçük Kareler Metodu Hesaplama

$$\sum e^2 = \sum \left(Y_i - \widehat{Y}_i\right)^2 = \sum (Y_i - b_0 - b_1 X_i)^2 = min$$

Yukarıdaki denklemin minimum noktası olabilmesi için birinci türevinin sıfır olması gerekir. Buna göre b_0 ve b_1 için ayrı ayrı türev alıp sıfıra eşitlediğimizde aşağıdaki eşitliklere ulaşırız.

$\sum Y_i = b_0(n) + b_1 \sum X_i \qquad \sum Y_i X_i = b_0 \sum x_i + b_1 \sum X_i^2$

b_0 ve b_1 çekildiğinde;

$b_0 = \frac{\sum X^2 \sum Y - \sum X \sum XY}{n\sum X^2 - (\sum X)^2} \qquad b_1 = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$ olur ve $\widehat{Y}_i = b_0 + b_1 X_i$ denkleminde yerine yazarak istenilen değerler için tahmin gerçekleştirilebilir.

Örnek Çalışma

Yıl	Satışlar	Reklam Giderleri
2001	223	7
2002	215	11
2003	233	15
2004	264	22
2005	305	26
2006	316	28
2007	320	31

Bağımlı Değişken(Y): Satışlar
Bağımsız Değişken(X): Reklam Giderleri

Çözüm

$b_0 = \frac{\sum X^2 \sum Y - \sum X \sum XY}{n\sum X^2 - (\sum X)^2}$ $b_1 = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$ değerleri için aşağıdaki tablodaki hesaplamalar yapılır.

Y_i	X_i	X_i^2	$X_i Y_i$
223	7	49	1561
215	11	121	2365
233	15	225	3495
264	22	484	5808
305	26	676	7930
316	28	784	8848
320	31	961	9920
1876	**140**	**3300**	**39927**

$$b_0 = \frac{3300(1876) - 140(39927)}{7(3300) - (140)^2} = 171{,}72$$

$$b_1 = \frac{7(39927) - 140(1876)}{7(3300) - (140)^2} = 4{,}841$$

$$Y_i = 171{,}72 + 4{,}814X_i$$

Sapmalarla Tahmin

En küçük kareler yöntemine alternatif olarak $x_i = X_i - \bar{X}$ ve $y_i = Y_i - \bar{Y}$ dönüşümleri ile de hesap yapabiliriz. Bunun için birinci denklemi yeniden yazalım.

$\sum Y_i = b_0(n) + b_1 \sum X_i$ her iki tarafı da n ile bölersek $\bar{Y} = b_0 + b_1 \bar{X}$ olur. Regresyon denklemi ile alt alta yazarsak,

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

$$\bar{Y} = b_0 + b_1 \bar{X}$$

$$y_i = b_1 x_i + e_i$$

Bu modelin b_1 e göre türevi alındığında $b_1 = \frac{\sum xy}{\sum x^2}$ eşitliği bulunur. Bulunan b_1 değeri $\bar{Y} = b_0 + b_1 \bar{X}$ denkleminde yerine koyularak b_0 değerine ulaşırız.

Örnek Çalışma

$$b_1 = \frac{\sum xy}{\sum x^2} = \frac{2407}{500} = 4{,}8$$

$$\bar{Y} = b_0 + b_1 \bar{X}$$

$$= 268 - 4{,}8(20) = 171{,}7$$

Y_i	X_i	x_i	y_i	$x_i y_i$	x_i^2
223	7	-13	-45	585	169
215	11	-9	-53	477	81
233	15	-5	-35	175	25
264	22	2	-4	-8	4
305	26	6	37	222	36
316	28	8	48	384	64
320	31	11	52	572	121
1876	**140**	**0**	**0**	**2407**	**500**

Regresyonun Standart Hatası

$Y_i = \beta_0 + \beta_1 X_i + e_i$ ilişkisindeki e_i değerlerinin ortalaması 0'dır $\bar{e} = 0$ sapmaların karesi $\sum(e_i - \bar{e})^2 = \sum e_i^2$ olur. Buradan hataların varyansı, $\sigma^2 = \frac{\sum e_i^2}{n-k}$ olarak bulunur. K regresyondaki katsayı sayısıdır. Standart hata ne kadar küçük olursa noktaların doğru etrafında kümelenmesi o ölçüde yakın olur.

Determinasyon Katsayısı

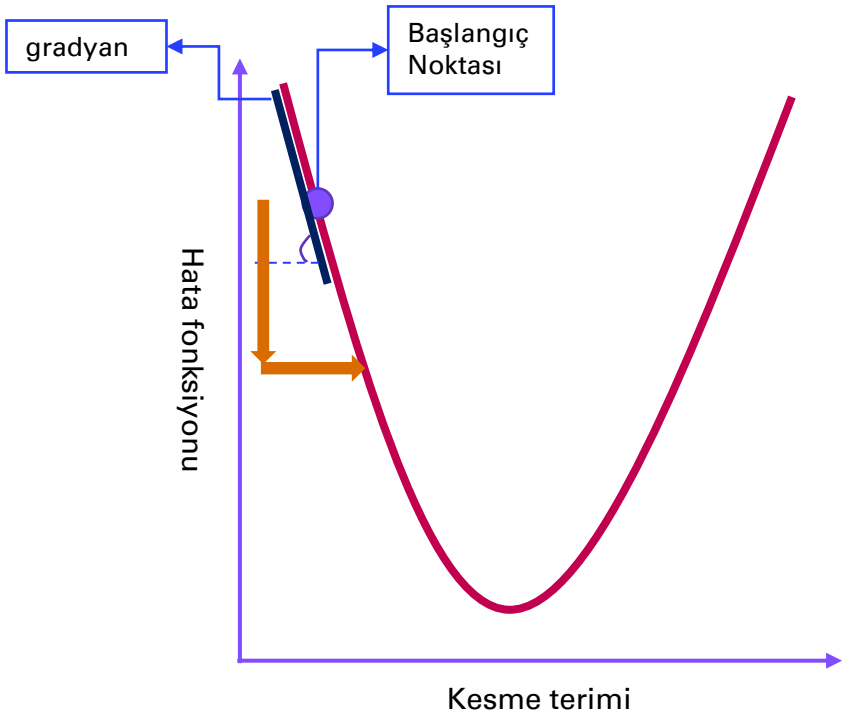
Tahmin edilen bir regresyonun genel başarısı yüzdelik bir derece olarak determinasyon katsayısı ile ölçülür. R^2 ile gösterilen determinasyon katsayısı, basit regresyon için bağımlı ve bağımsız değişkenler arasındaki basit korelasyon katsayısıdır.

$R^2 = \frac{(\sum xy)^2}{\sum x^2 \sum y^2}$ olur. $b_1 = \frac{\sum xy}{\sum x^2}$ olduğu için

$R^2 = b_1 \frac{\sum xy}{\sum y^2}$ ile determinasyon katsayısı hesaplanır. 0 ile 1 arasında değerler alır.

Gradient Descent

Makine öğrenmesinde toplam hataların minimize edilmesi amacıyla kullanılır.



1- Rastgele bir başlangıç noktası belirle.

2- Noktanın kayıp fonksiyonunu hesapla. $RMSE = \sqrt{\frac{\sum(Y_i - \hat{Y}_i)^2}{n}}$

$$\beta_i^k = \beta_i^{k-1} - \Delta\beta_i$$

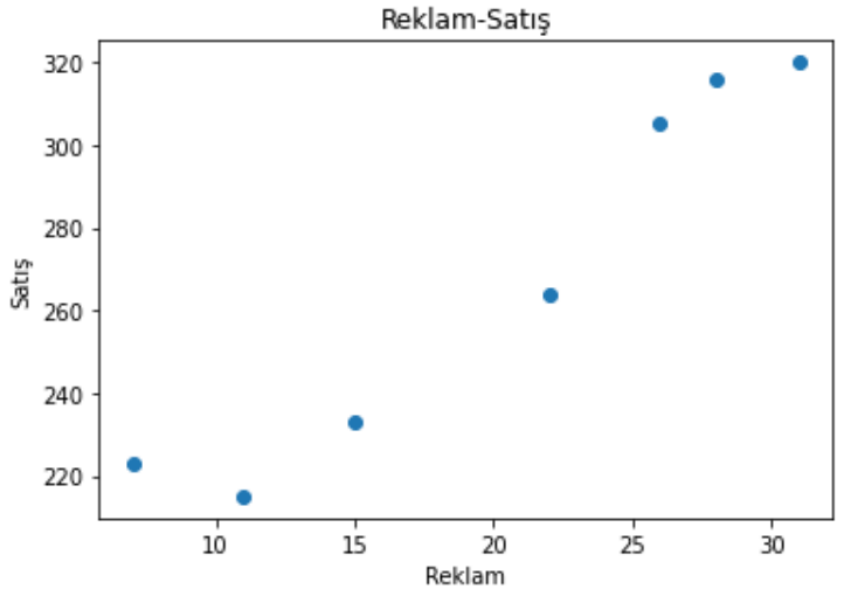
3- Belirlenen noktanın türevini alarak eğimini hesapla.

4- Adım boyutunu belirle.

5- 2-4 arasını tekrar et.

Satışlar	Reklam Giderleri
223	7
215	11
233	15
264	22
305	26
316	28
320	31

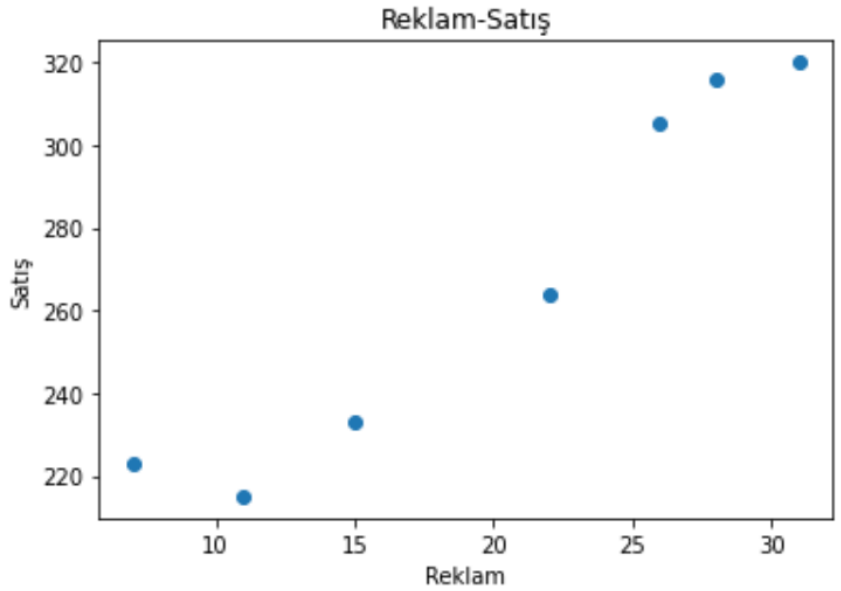
$$Satışlar = \beta_0 + \beta_1 \times Reklam$$



Satışlar	Reklam Giderleri
223	7
215	11
233	15
264	22
305	26
316	28
320	31

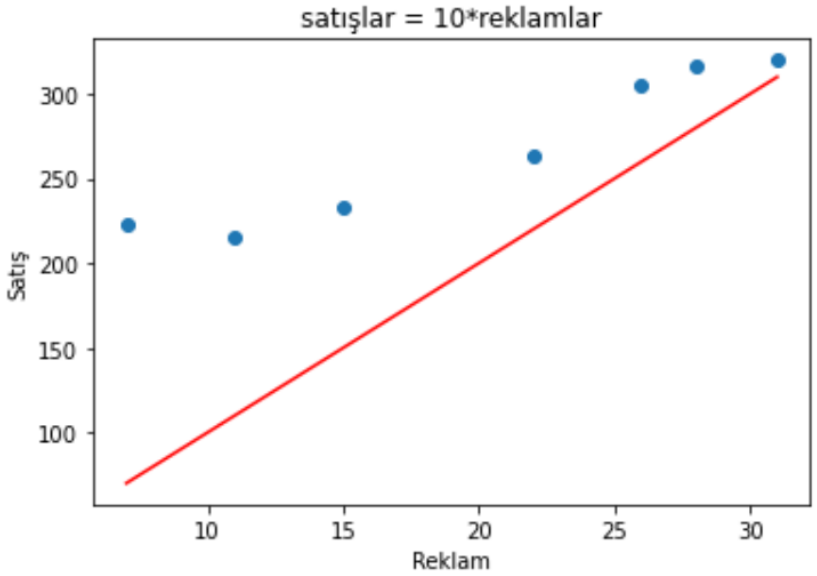
$$Satışlar = \mathbf{0}\beta_0 + \beta_1\mathbf{10} \times Reklam$$

Rassal bir değer ile başlanır.



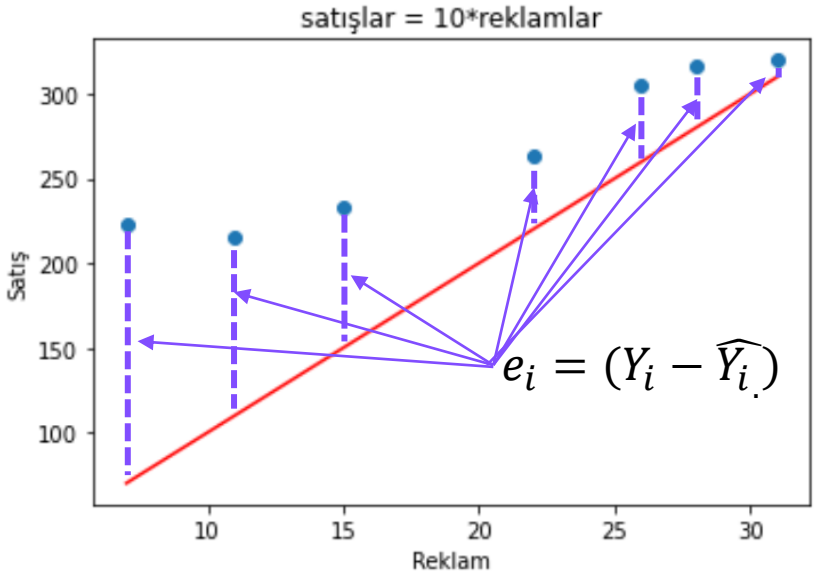
Satışlar	Reklam Giderleri
223	7
215	11
233	15
264	22
305	26
316	28
320	31

$$Satışlar = \mathbf{0} + \mathbf{10} \times Reklam$$



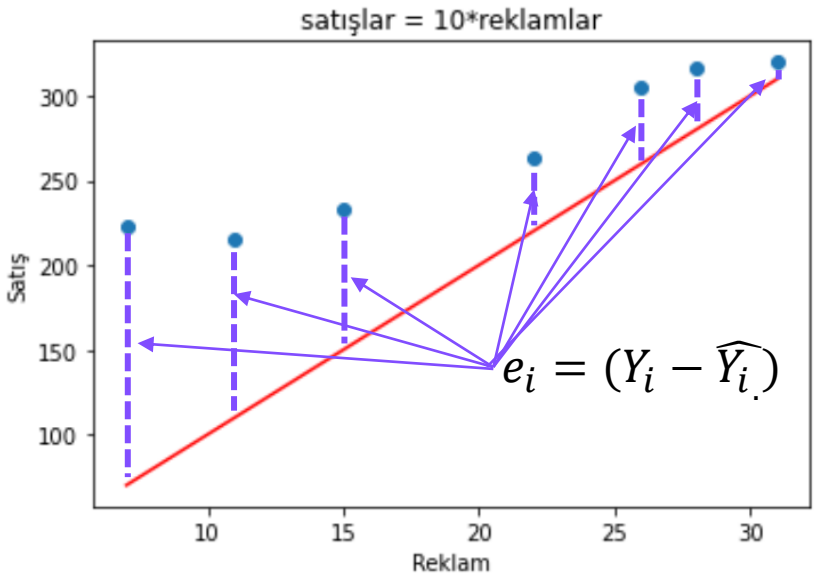
Satışlar	Reklam Giderleri
223	7
215	11
233	15
264	22
305	26
316	28
320	31

$$Satışlar = \mathbf{0} + \mathbf{10} \times Reklam$$



```
In [29]:
1 satislar = np.array([223, 215, 233, 264, 305, 316, 320])
2 reklam = np.array([7, 11, 15, 22, 26, 28, 31])

In [30]:
1 y_prediction = 10*reklam
2 (y_prediction-satislar)

Out[30]: array([-153, -105,  -83,  -44,  -45,  -36,  -10])

In [31]:
1 np.mean((y_prediction-satislar)**2)

Out[31]: 6668.571428571428

In [32]:
1 RMSE = np.sqrt(np.mean((y_prediction-satislar)**2))
2 RMSE

Out[32]: 81.66132149660223
```

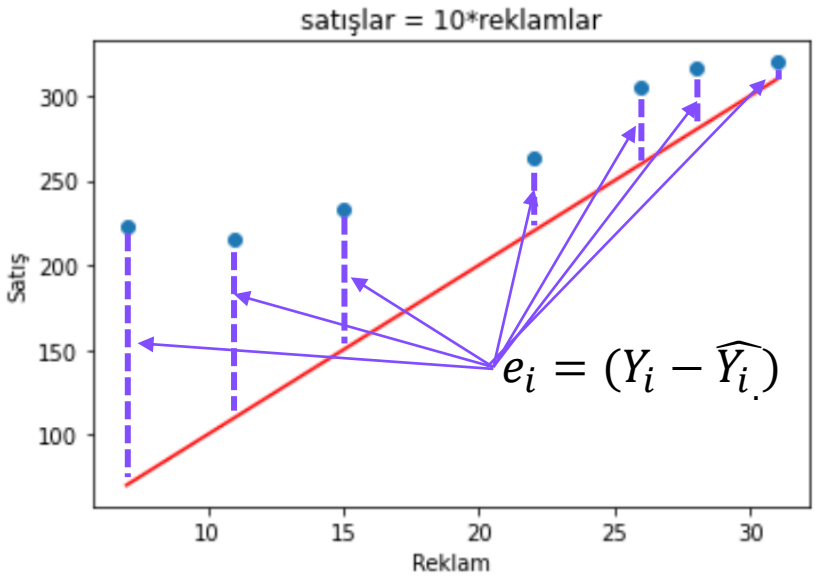
$$Satışlar = \mathbf{0} + \mathbf{10} \times Reklam$$



```
In [29]:
1 satislar = np.array([223, 215, 233, 264, 305, 316, 320])
2 reklam = np.array([7, 11, 15, 22, 26, 28, 31])

In [30]:
1 y_prediction = 10*reklam
2 (y_prediction-satislar)

Out[30]: array([-153, -105,  -83,  -44,  -45,  -36,  -10])

In [31]:
1 np.mean((y_prediction-satislar)**2)

Out[31]: 6668.571428571428

In [32]:
1 RMSE = np.sqrt(np.mean((y_prediction-satislar)**2))
2 RMSE

Out[32]: 81.66132149660223
```

$$Satışlar = \mathbf{0} + \mathbf{10} \times Reklam$$

$$\beta_0 = 0 \longrightarrow RMSE = 81.66$$



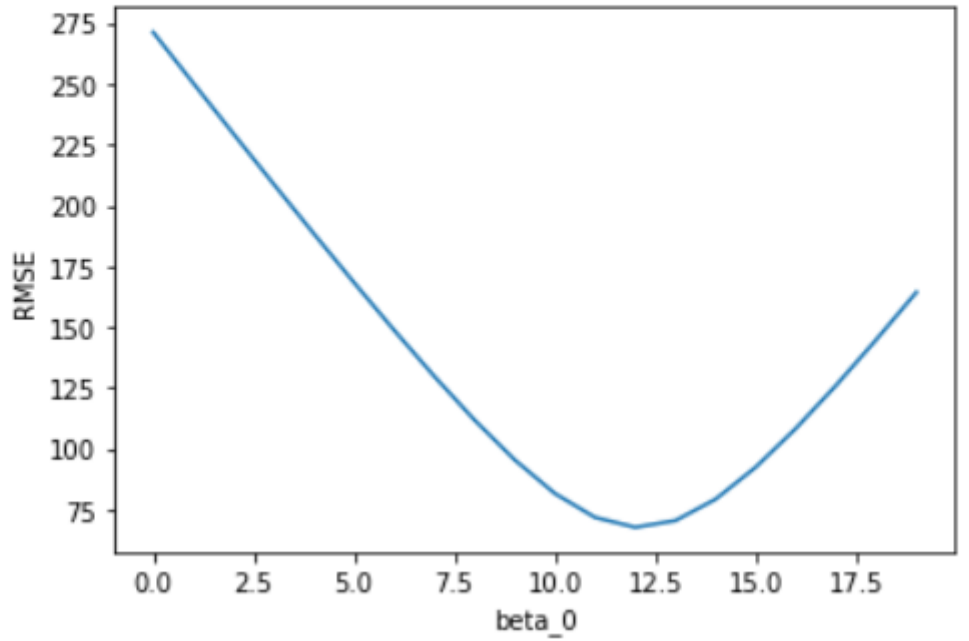
In [62]:
```
RMSE_list = []
beta_0_list = []
for i in range(0, 20, 1):
    y_prediction = i * reklam
    RMSE = np.sqrt(np.mean((y_prediction - satislar)**2))
    RMSE_list.append(RMSE)
    beta_0_list.append(i)
```

In [63]:
```
RMSE_list
```

Out[63]:
```
[271.29846505805585,
 250.3329211841132,
 229.50630243447097,
 208.86017195380126,
 188.45385945333447,
 168.37458240482735,
 148.7548318542964,
 129.80314766159233,
 111.85960588421287,
 95.49420326461107,
 81.66132149660223,
 71.83910594416625,
 67.79380502671317,
 70.52659073002182,
 79.34013576278493,
 92.51254741153456,
 108.4672432448749,
 126.1529683700365,
 144.93742492143684,
 164.4445195195024]
```

In [67]:
```
plt.plot(beta_0_list,RMSE_list)
plt.xlabel('beta_0')
plt.ylabel('RMSE')
```

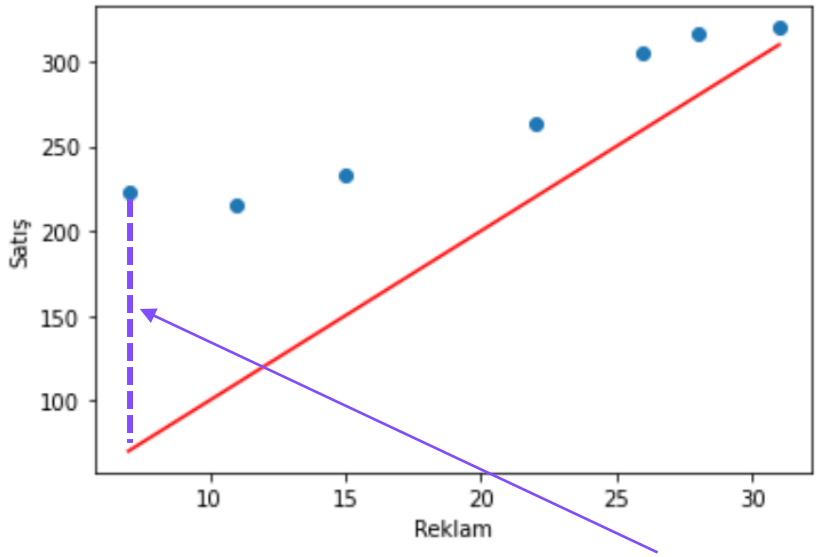
Out[67]:
```
Text(0, 0.5, 'RMSE')
```



In [68]:
```
beta_0_list[np.argmin(np.array(RMSE_list))]
```

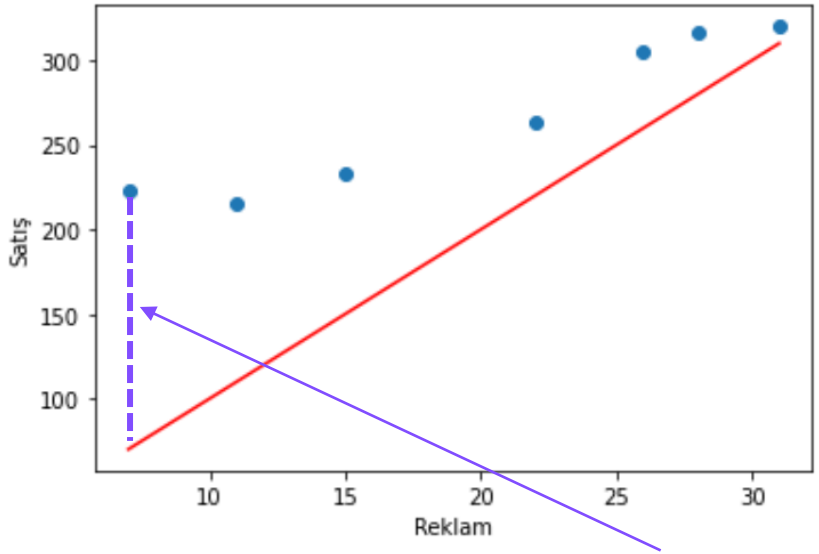
Out[68]:
```
12
```

Adım Boyutu Hesabı



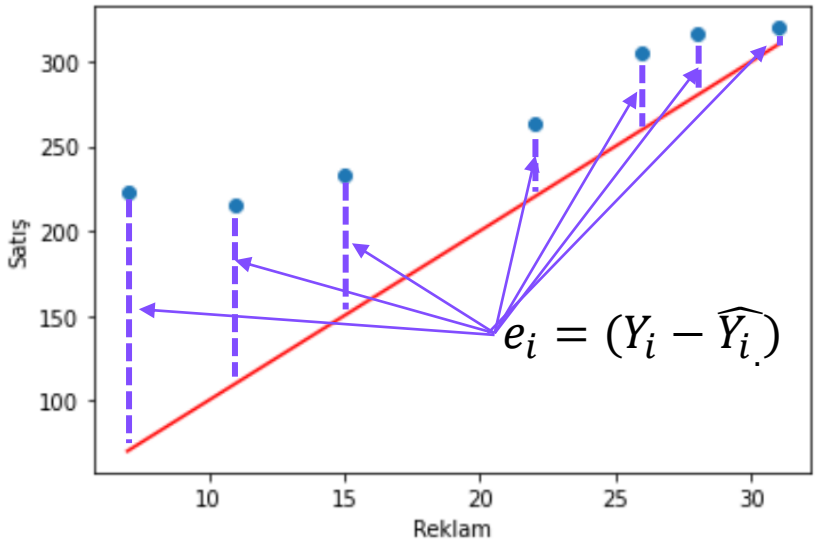
$$e_1^2 = \left(223 - (\beta_0 + \beta_1 * 7)\right)^2$$

Adım Boyutu Hesabı



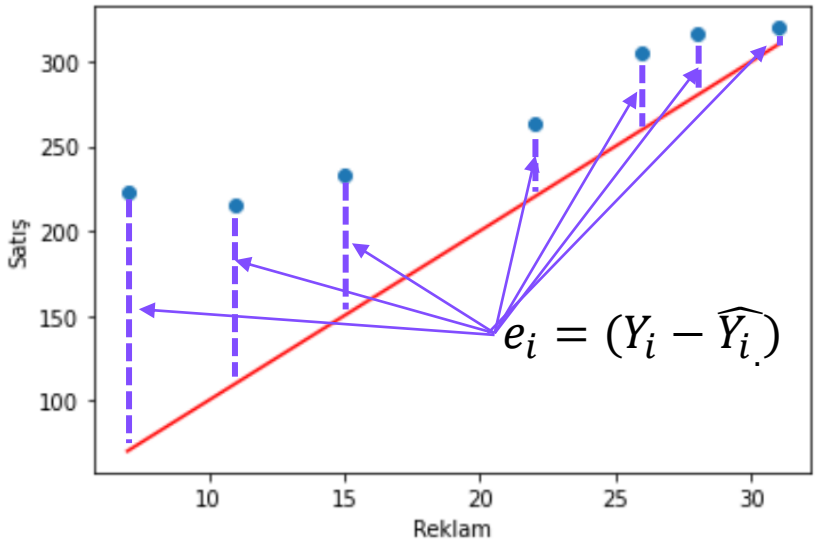
$$e_1^2 = \left(223 - (\beta_0 + 10 * 7)\right)^2$$

Adım Boyutu Hesabı



$$\sum e_i^2 = \big(223 - (\beta_0 + 10 * 7)\big)^2 + \big(215 - (\beta_0 + 10 * 11)\big)^2 + \big(233 - (\beta_0 + 10 * 15)\big)^2 + \big(264 - (\beta_0 + 10 * 22)\big)^2 + \big(305 - (\beta_0 + 10 * 26)\big)^2 + \big(316 - (\beta_0 + 10 * 28)\big)^2 + \big(320 - (\beta_0 + 10 * 31)\big)^2$$

Adım Boyutu Hesabı



$$\sum e_i^2 = \left(223 - (\beta_0 + 10 * 7)\right)^2 +$$

$$\frac{d}{\beta_0}\left(223 - (\beta_0 + 10 * 7)\right)^2$$

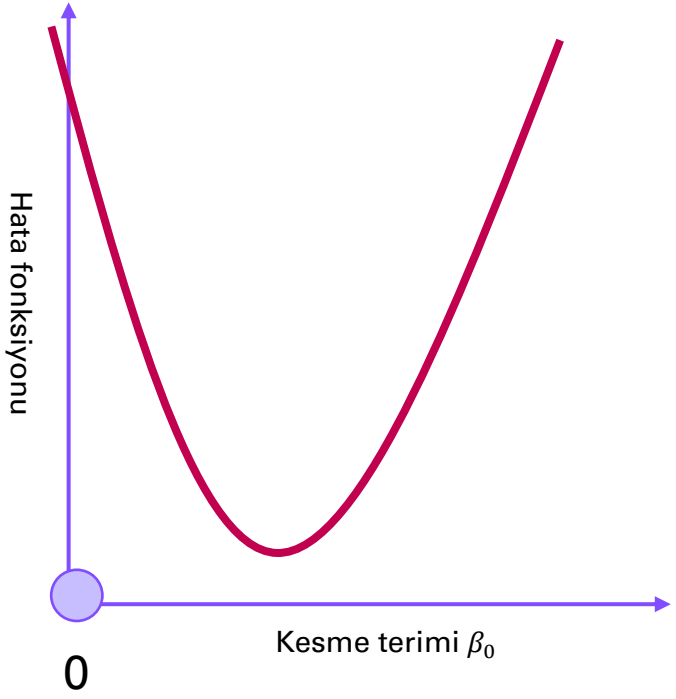
$$\frac{d}{\beta_0}\left(223^2 - 2 * 223(\beta_0 + 10 * 7) + (\beta_0 + 10 * 7)^2\right)$$

$$= 2\beta_0 - 306$$

https://www.derivative-calculator.net/

(223−(x+10*7))^2 + (215−(x+10*11))^2 + (233−(x+10*15))^2+ (264−(x+10*22))^2+ (305−(x+10*26))^2+ (316−(x+10*28))^2 + (320−(x+10*31))^2

Çözüm: 14x-952



$\beta_0 = 0$ olduğunda 14x-952 denkleminden eğim -952 olarak gelir.

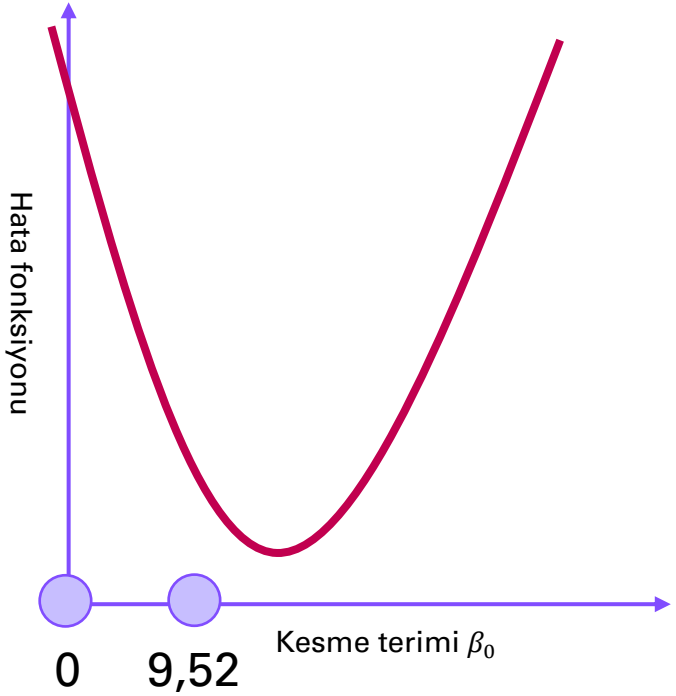
Adım boyu = eğim * öğrenme katsayısı (λ) = -952*0,01=-9,52

Learning rate ile ilgili gösterim
https://developers.google.com/machine-learning/crash-course/reducing-loss/learning-rate

Yeni beta0 önceki beta0 dan adım boyunun çıkarılması ile elde edilir.

$$yeni\beta_0 = eski\beta_0 - (adım\ boyu)$$
$$yeni\beta_0 = 0 - (-9{,}52)$$
$$yeni\beta_0 = 9{,}52$$

Olacaktır.



$\beta_0 = 0$ olduğunda 14x-952 denkleminden eğim -952 olarak gelir.

Adım boyu = eğim * öğrenme katsayısı (λ) = -952*0,01=-9,52

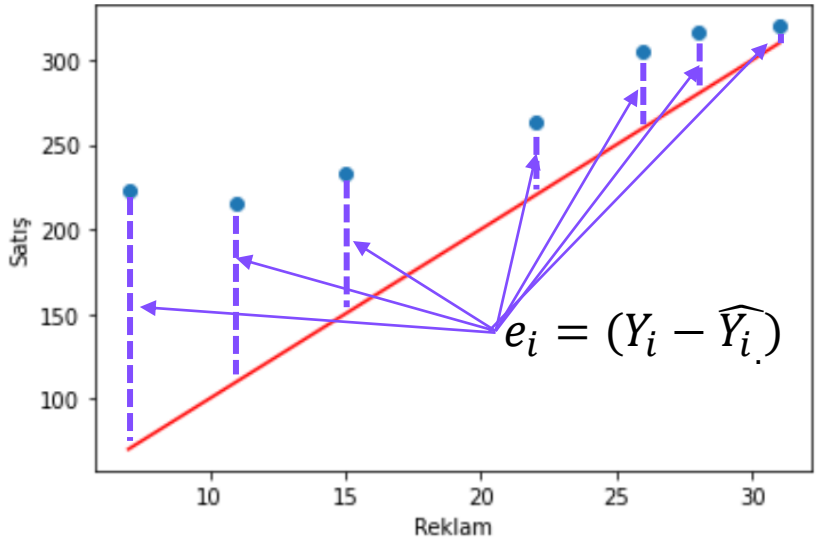
Learning rate ile ilgili gösterim
https://developers.google.com/machine-learning/crash-course/reducing-loss/learning-rate

Yeni beta0 önceki beta0 dan adım boyunun çıkarılması ile elde edilir.

$$yeni\beta_0 = eski\beta_0 - (adım\ boyu)$$
$$yeni\beta_0 = 0 - (-9{,}52)$$
$$yeni\beta_0 = 9{,}52$$

Olacaktır.

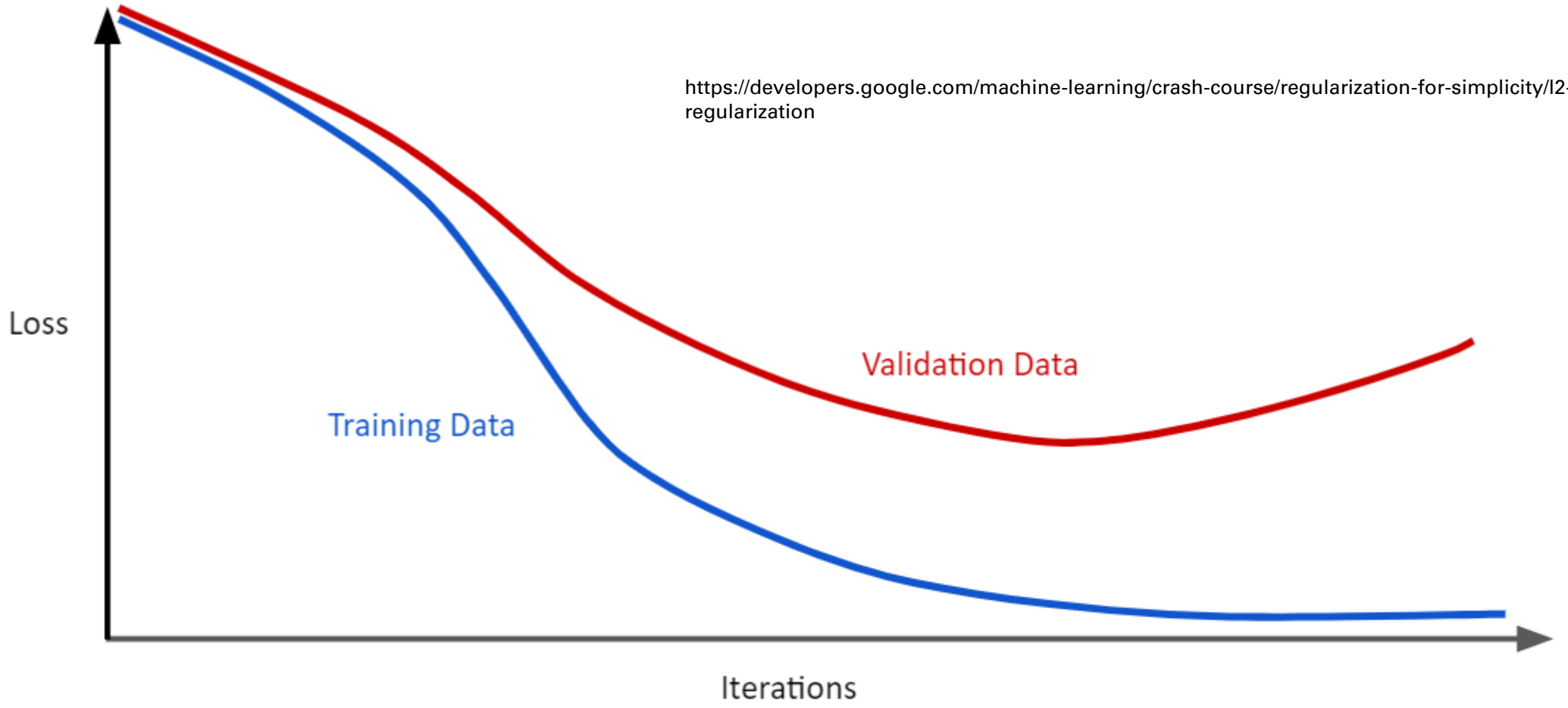
Beta0 = 9,52




```
In [70]:
1 beta_0_952 = 9.52
2 y_prediction_952 = 9.52 + 10*reklam
3 np.mean((y_prediction_952-satislar)**2)
4 RMSE_952 = np.sqrt(np.mean((y_prediction_952-satislar)**2))
5 RMSE_952

Out[70]: 73.92213354991473
```

Aşırı Öğrenme (Overfitting)

https://developers.google.com/machine-learning/crash-course/regularization-for-simplicity/l2-regularization



Ridge ve LASSO Regresyon

Ridge regresyon, aşırı yüksek olan model parametrelerini cezalandırır. Bunu yaparken hata terimlerinin karelerinin toplamına aşağıdaki gibi katsayının karesini ekler. Böylece yüksek katsayılara ceza verilmiş ve loss fonksiyonu yüksek tutulduğunda cezalandırılmış olur. Lambda parametresi cezalandırmanın şiddetini belirler. Lambda sıfır olduğunda ridge regresyon lineer regresyona döner.

$$RSS_{\text{ridge}} = \sum_{i=1}^{n}\left(y_i - x_i^T\beta\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

Diğer formda ise Least Absolute Shrinkage and Selection Operator (LASSO) regresyon bulunur. LASSO da aşağıdaki gibi bir yapıda hataların karelerini cezalandırır.

$$RSS_{\text{lasso}} = \sum_{i=1}^{n}\left(y_i - x_i^T\beta\right)^2 + \lambda\sum_{j=1}^{p}\beta_j$$