İSTATİSTİK

7. Korelasyon ve Regresyon

Korelasyon

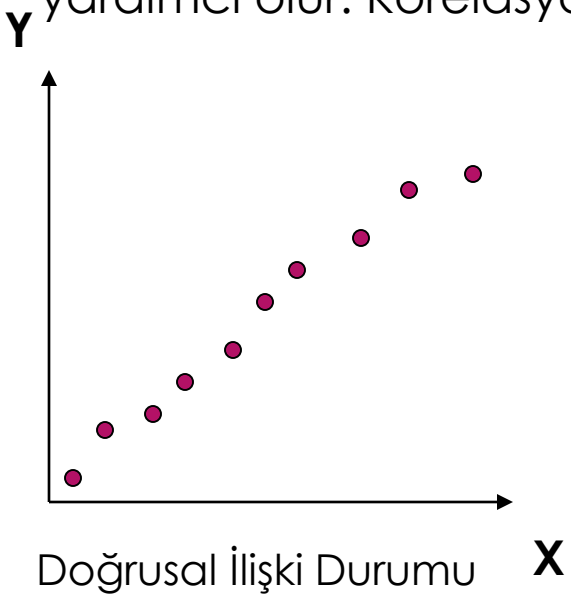
- Korelasyon iki sayısal değişken arasında ilişki olup olmadığını, ilişki varsa bu ilişkinin yönünü ve büyüklüğünü görmek için kullanılır.

- İki değişken arasındaki ilişkinin derecesi olan Pearson korelasyon katsayısı ‘r’ ile gösterilir.

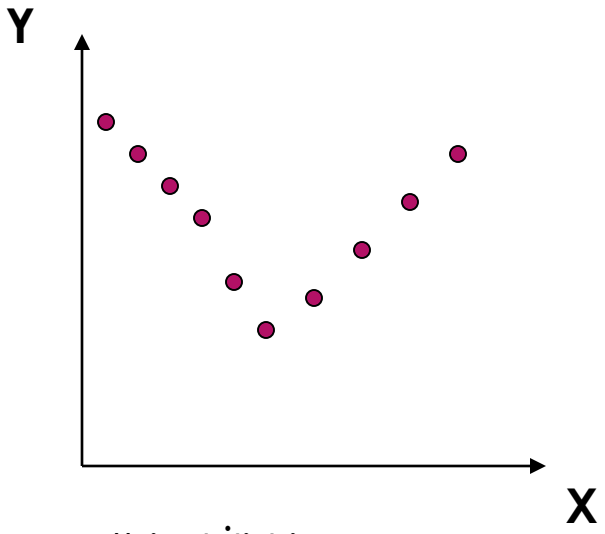
- Korelasyon katsayısı değişkenler arasında neden-sonuç ilişkisi kurmaz, iki değişkenin değişimlerinde ne dereceye kadar uygunluk olduğunu belirler.

SERPİLME DİYAGRAMI

Değişkenler arasındaki ilişki tipinin belirlenmesine yardımcı olur. Korelasyon doğrusal ilişkinin bir ölçüsüdür.

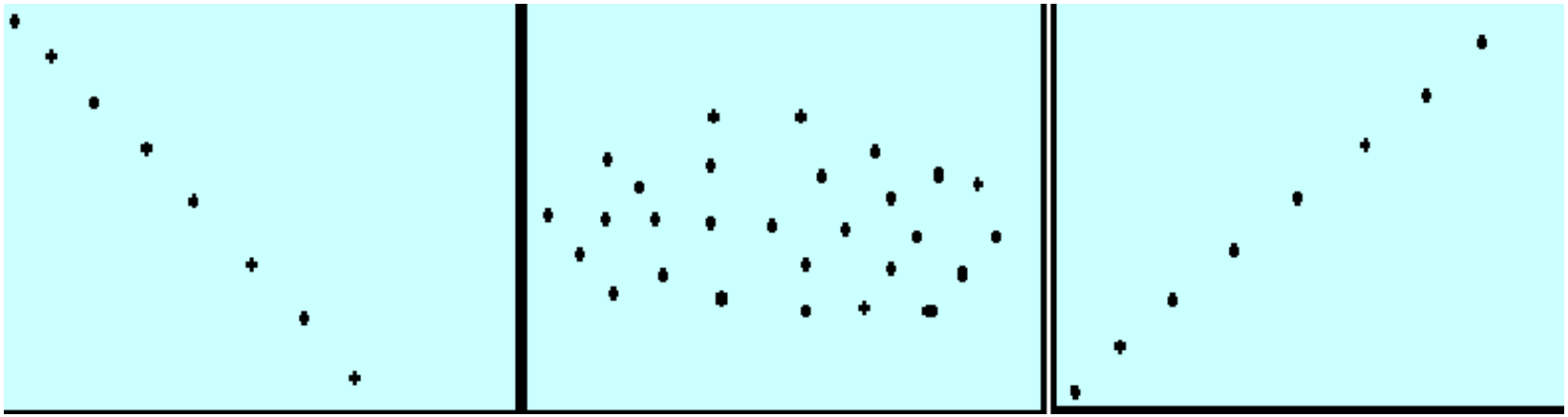


Doğrusal İlişki Durumu



Eğrisel İlişki Durumu

$-1 \leq r \leq +1$



Korelasyon katsayısının yorumlanması

- 0.00 - 0.25 Çok zayıf ilişki
- 0.26 - 0.49 Zayıf ilişki
- 0.50 - 0.69 Orta ilişki
- 0.70 - 0.89 Yüksek ilişki
- 0.90 - 1.0 Çok yüksek ilişki

Regresyon Nedir?

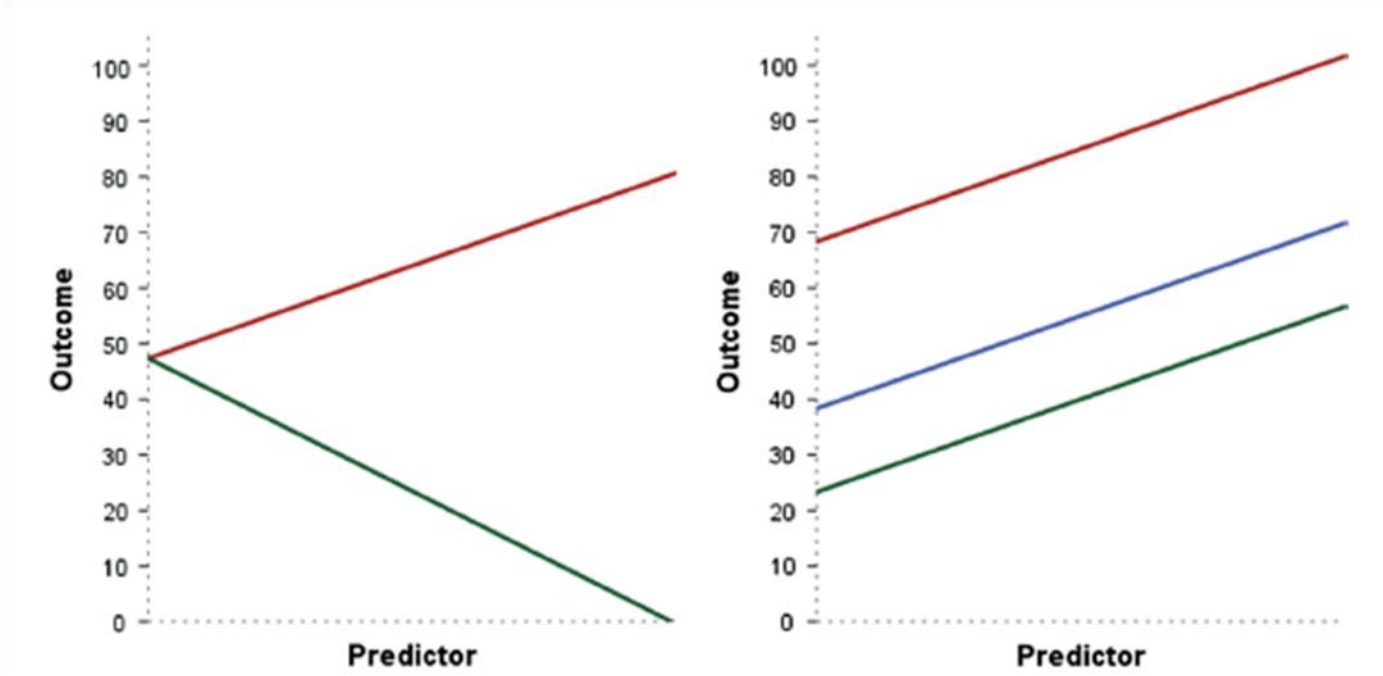
- Bir değişken kullanarak başka bir değişkenin değerini tahmin etme yöntemidir.
  - İki değişken arasındaki ilişkiye ait bir model oluşturur.
  - Bu model doğrusal bir modeldir.
  - Dolayısıyla bu ilişki bir doğru denklemi kullanılarak ifade edilir.

Doğru Denklemi

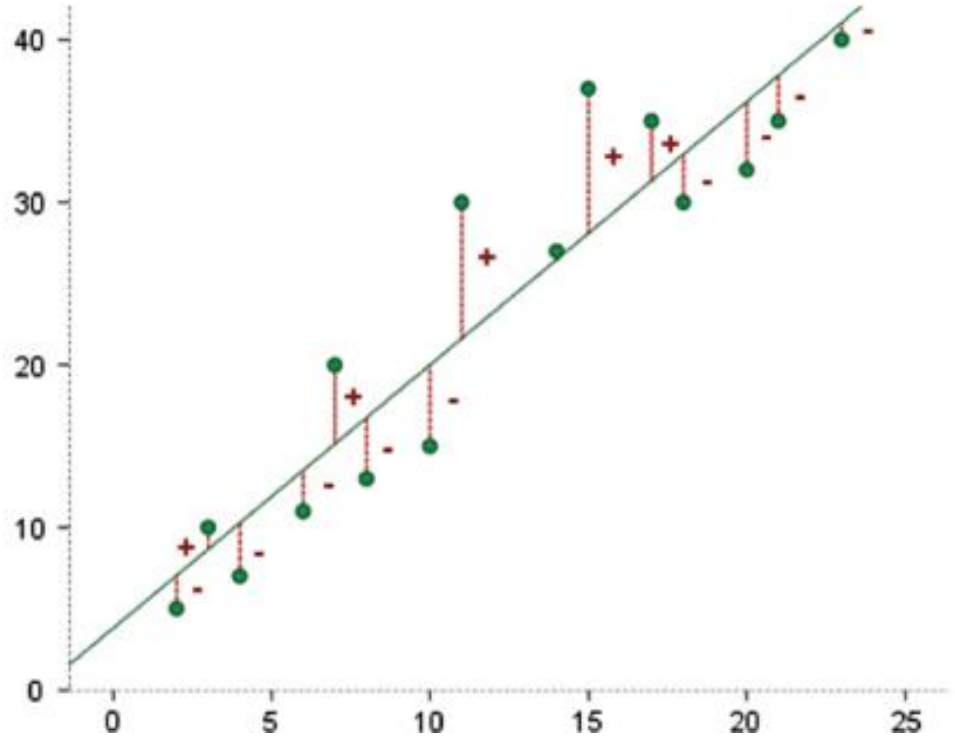
$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

- b_1
  - Bağımsız (tahminleyici) değişkenin katsayısıdır
  - Regresyon doğrusunun eğimidir.
  - İlişkinin yönünü / büyüklüğünü belirler
- b_0
  - Y eksenini kesme noktasıdır (X = 0 iken Y'nin aldığı değerdir)

Kesme noktası ve Eğimler

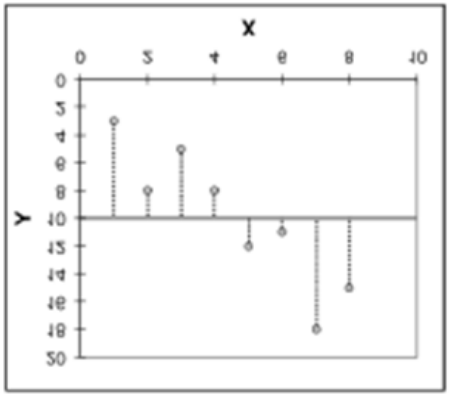


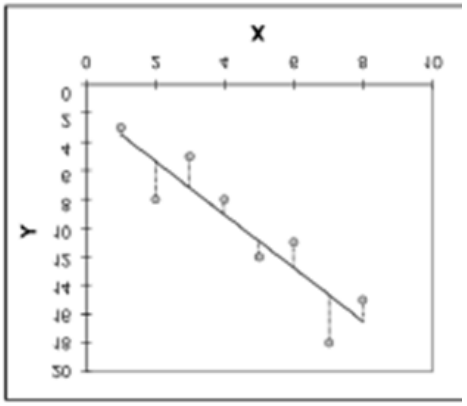
En Küçük Kareler Yöntemi

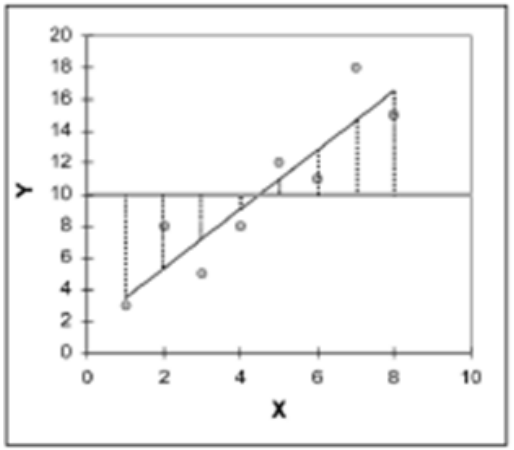


Regresyon Modeli Ne Kadar İyi?

- Regresyon modeli sadece veriyi temsil ettiği düşünülen bir modeldir.
- Bu model gerçeği yansıtmayabilir.
  - Dolayısıyla modelimizin gözlemlenen veriye ne kadar iyi uyduğunu test etmemiz gerekir.







- SS_T
  - Toplam değişenlik (gözlemlenen skorlar ve ortalama arasındaki değişkenlik).
- SS_R
  - Residual/Hata değişkenliği (Regresyon modeli ile gerçek veriler arasındaki değişkenlik).
- SS_M
  - Model değişkenliği(Regresyon modeli ile ortalama modellerinin değişkenlikleri arasındaki fark).

Model Testi

- Eğer modelin tahmini ortalama kullanılarak yapılan tahminden daha iyi ise SS_M 'nin SS_R'den daha büyük olmasını bekleriz

Model Testi: ANOVA

$$F = \frac{MS_M}{MS_R}$$

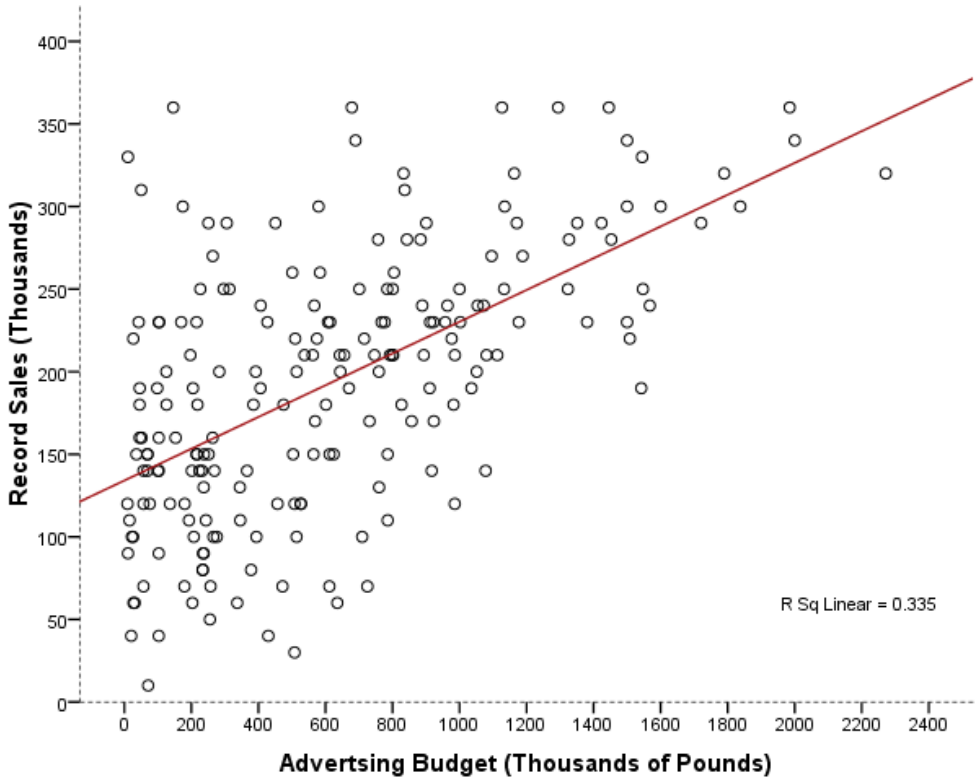
$$R^2 = \frac{SS_M}{SS_T}$$

R^2 regresyon modeli tarafından (bağımsız değişkenler tarafından açıklanan) varyans. Korelasyon katsayısının karesidir.

Örnek

- Bir müzik şirketi albüm satışlarına reklamın etkisini ölçmek istemektedir.
- Veri
  - 200 farklı albüme ait
- Çıktı değişkeni:
  - Satışlar (Albüm çıktıktan sonra bir hafta içerisindeki)
- Tahmin değişkeni:
  - Albüm satışa çıkmazdan önce reklama harcanan para.

Grafik


Record Sales (Thousands)
400
350
300
250
200
150
100
50
0
0
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
2400
Advertsing Budget (Thousands of Pounds)
R Sq Linear = 0.335

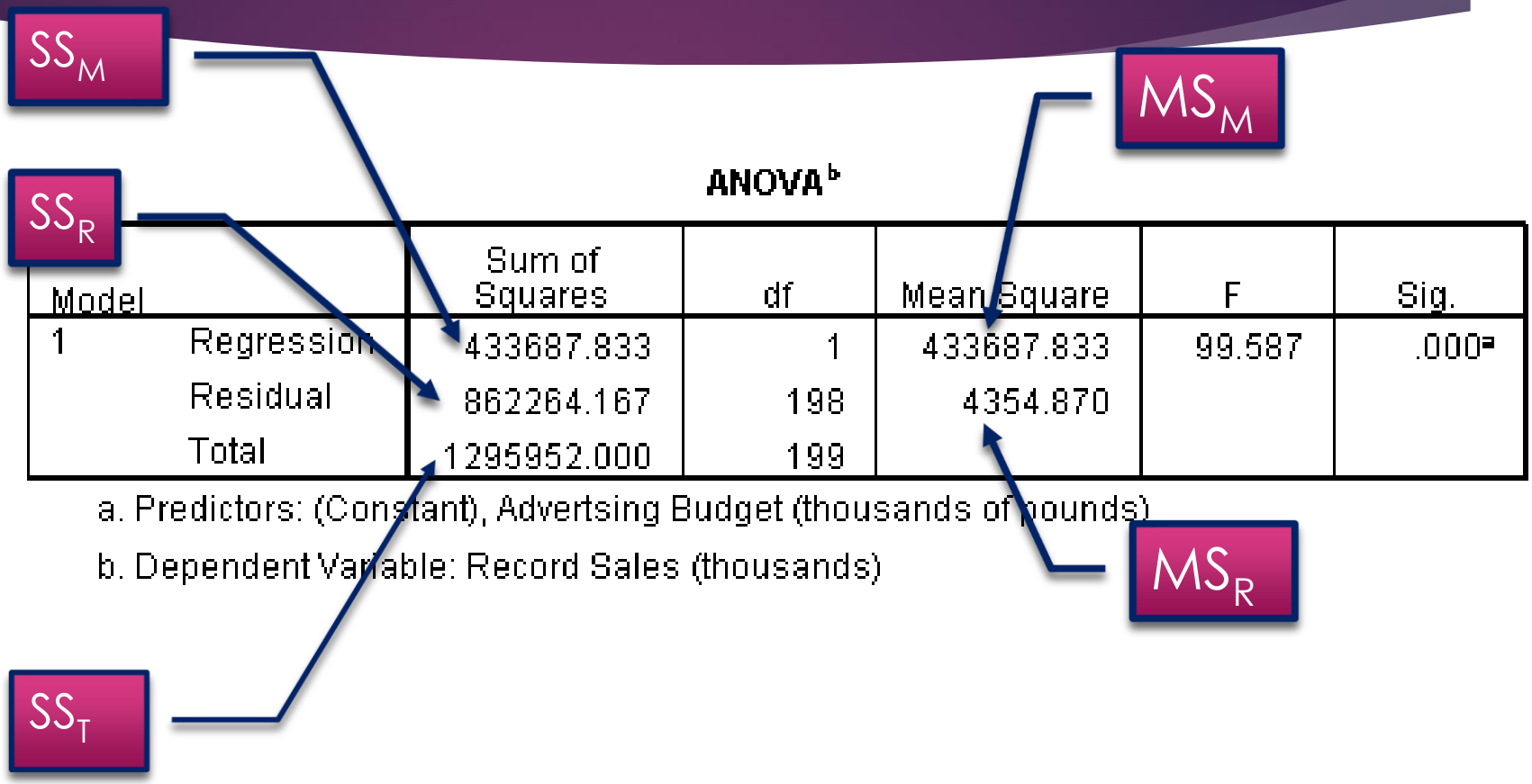
Çıktı: Model Özeti

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.578[a]	.335	.331	65.9914

a. Predictors: (Constant), Advertising Budget (thousands of pounds)

Çıktı: ANOVA



ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	433687.833	1	433687.833	99.587	.000[a]
	Residual	862264.167	198	4354.870		
	Total	1295952.000	199			

a. Predictors: (Constant), Advertsing Budget (thousands of pounds)

b. Dependent Variable: Record Sales (thousands)

SPSS Çıktısı: Model Parametreleri

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	134.140	7.537		17.799	.000
	Advertsing Budget (thousands of pounds)	.096	.010	.578	9.979	.000

a. Dependent Variable: Record Sales (thousands)

Model

$$\begin{aligned}\text{Record Sales}_i &= b_0 + b_1 \text{Advertising Budget}_i \\ &= 134.14 + \left(0.09612 \times \text{Advertising Budget}_i\right)\end{aligned}$$

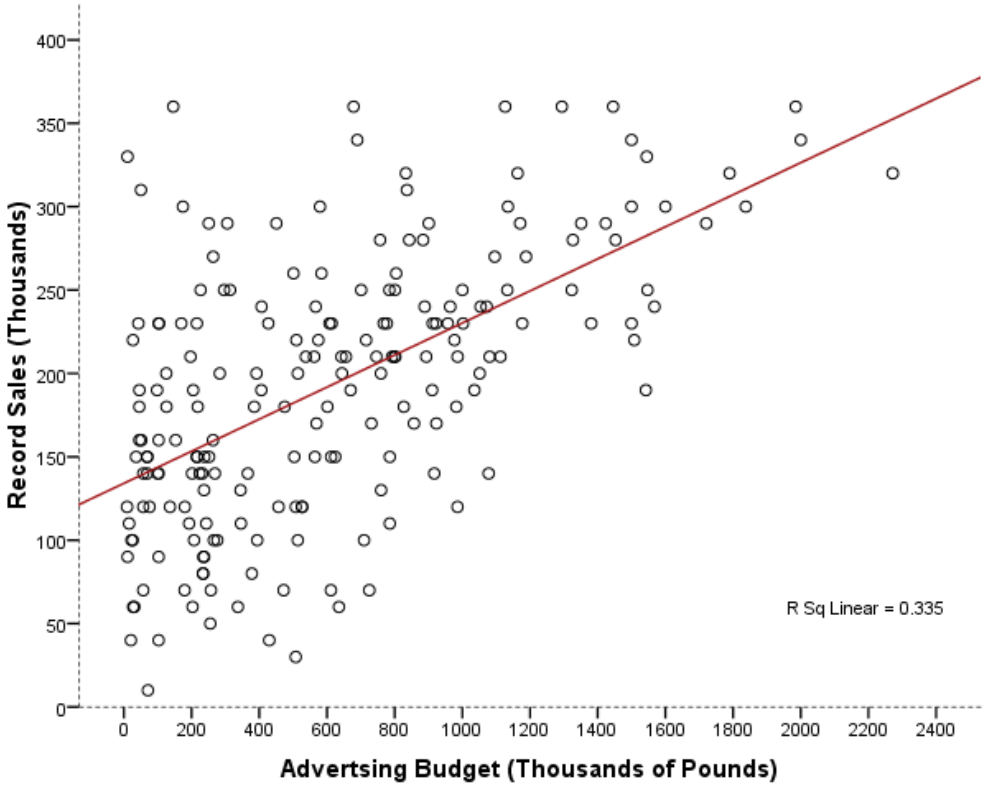
$$\begin{aligned}\text{Record Sales}_i &= 134.14 + \left(0.09612 \times \text{Advertising Budget}_i\right) \\ &= 134.14 + \left(0.09612 \times 100\right) \\ &= 143.75\end{aligned}$$

Çoklu Regresyon Nedir?

- Çoklu regresyonda bir çıktı değişkeninin değeri çok sayıdaki tahminleyici değişken kullanılarak tahmin edilir.
- Örnek:

- Bir müzik şirketi albüm satışlarına reklamın etkisini ölçmek istemektedir.
- Veri
  - 200 farklı albüme ait
- Çıktı değişkeni:
  - Satışlar (Albüm çıktıktan sonra bir hafta içerisindeki)
- Tahmin değişkenleri
  - Albüm satışa çıkmazdan önce reklama harcanan para
  - Radyoda albümün kaç kere çalındığı (yeni değişken)

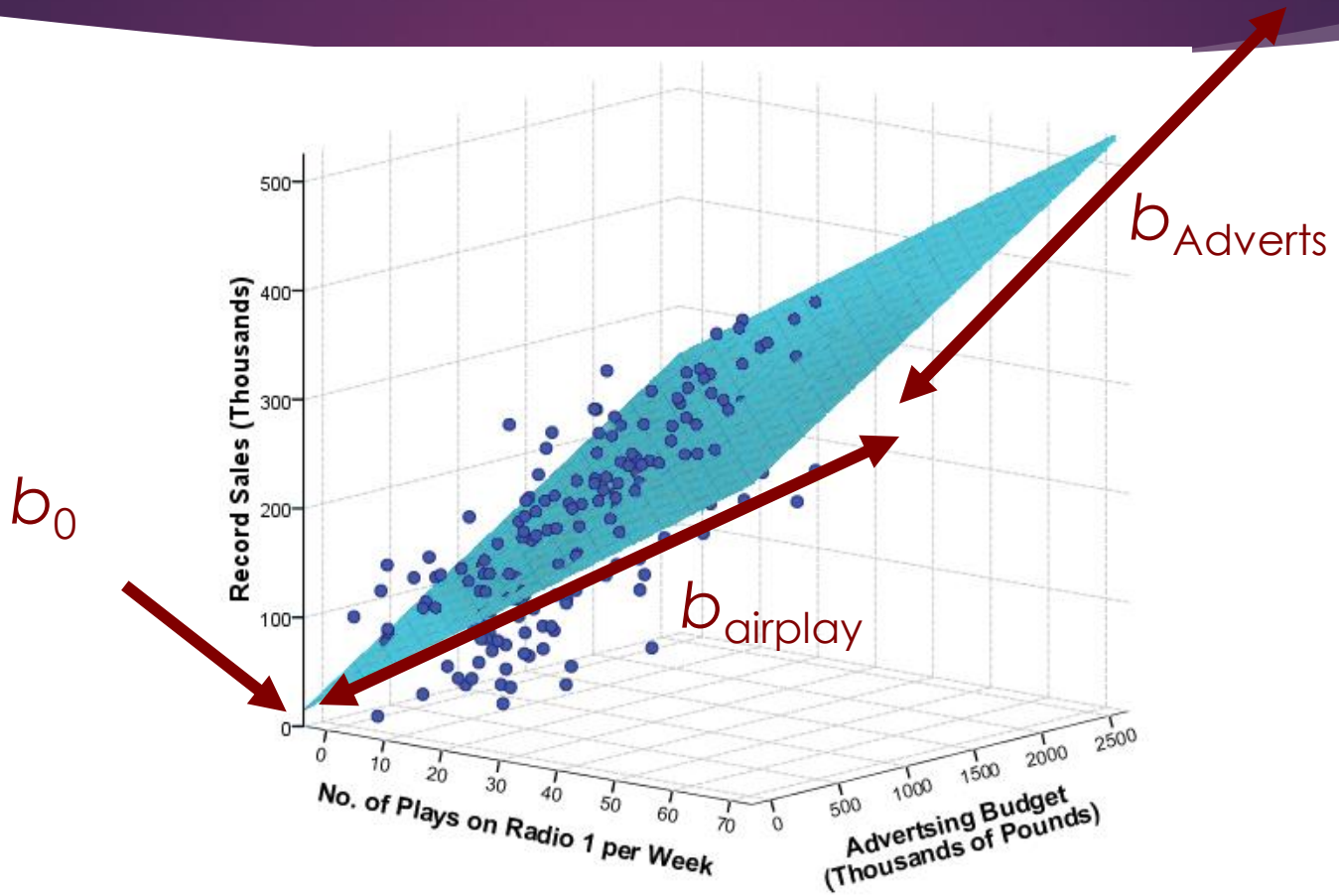
Tek değişkenli model



Çoklu Regresyon Denklemi

$$y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_n X_n + \varepsilon_i$$

İki değişkenli model grafiği



Katasayıları yorumlamak

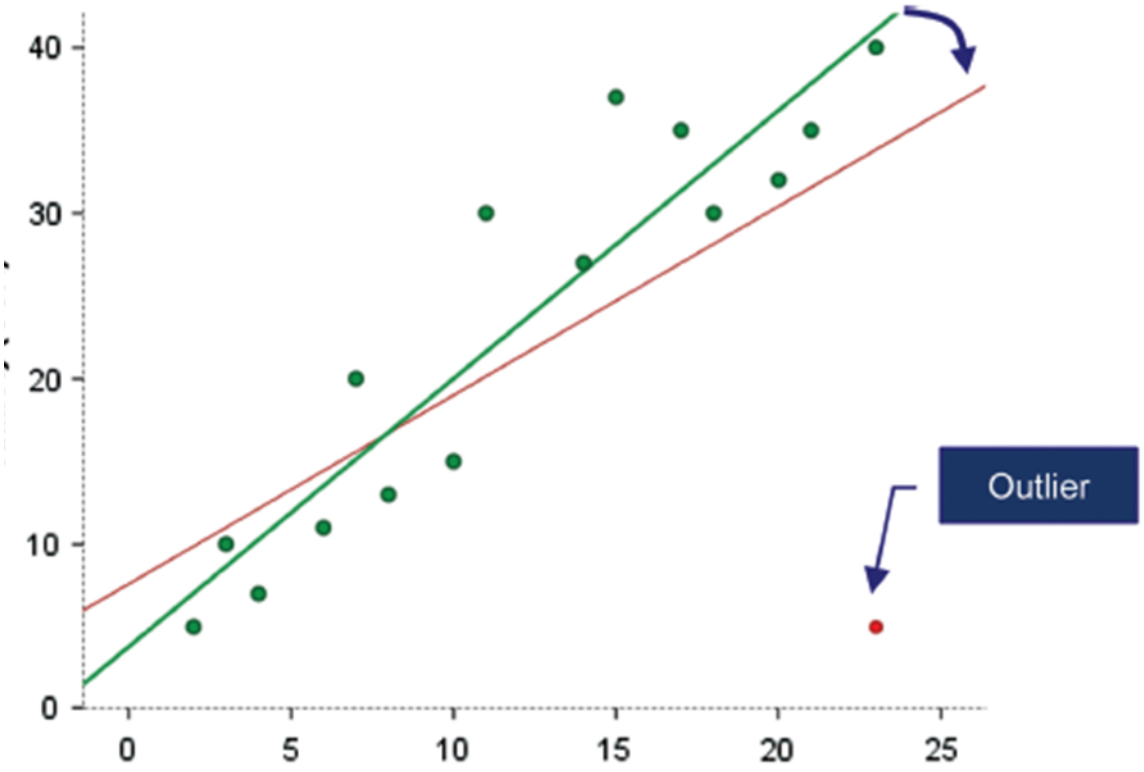
- B değerleri:
  - bağımsız değişkendeki bir birim değişme, bağımlı değişkende b kadar değişime neden olur.
- Beta değerleri:
  - Yukarıdaki değer standart sapma cinsinden ifade edilir.

Model'in doğruluğu için dikkat edilecek konular

- Residual istatistiklerine bakılması gerekir
  - Ortalama olarak standardized residual'ların 95% 'i ± 2 arasında olmalıdır.
  - Ortalama olarak standardized residual'ların 99 % 'i ± 2.5 arasında olmalıdır.
  - Mutlak standardized residual değeri 3 yada daha fazla olan veriler uç değer olabilir.

- Yüksek etkili verilere dikkate etmek gerekir.
  - Cook's distance: Bir verinin tek başına modelin bütünü üzerindeki etkisini ölçer.
  - Mutlak değeri 1'den büyük olan veriler sorun oluşturabilir

Uç Değerler



Regresyon Varsayımları

- Hesaplanan regresyon modelinin genelleştirilebilmesi için bazı varsayımların sağlanması gerekir. Bu varsayımlar sağlanmazsa örnekleme dayalı oluşturduğumuz regresyon modelini ana kütleye genelleştiremeyiz.

1. Çıktı değişkeni sayısal (sürekli) olmalı
2. Tahmin değişkenleri sürekli yada kategorik (iki kategorili) olabilir.
3. Tahmin değişkenlerinin varyansı sıfırdan farklı olmalı
4. Modelimiz doğrusal olmalı
5. Bağımlı değişkenin ölçümleri birbirinden bağımsız olmalı
6. Bağımsız değişkenler arasında yüksek oranda korelasyon olmamalı (Multicollinearity). Bu amaçla Tolerance ve VIF değerlerine bakılabilir. Tolerans değeri 0.2'den büyük olmalı veya VIF değeri 10'dan küçük olmalı

Varsayımlar…

Coefficients[a]

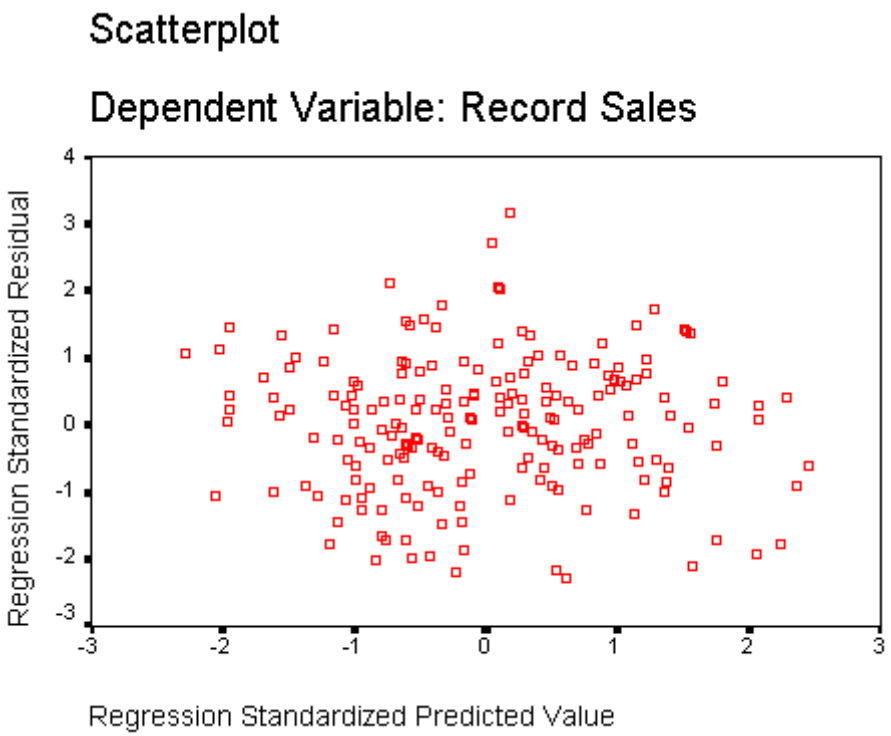
Model		Correlations			Collinearity Statistics	
		Zero-order	Partial	Part	Tolerance	VIF
1	Advertsing Budget (thousands of pounds)	.578	.578	.578	1.000	1.000
2	Advertsing Budget (thousands of pounds)	.578	.659	.507	.986	1.015
	No. of plays on Radio 1 per week	.599	.655	.501	.959	1.043
	Attractiveness of Band	.326	.309	.188	.963	1.038

a. Dependent Variable: Record Sales (thousands)

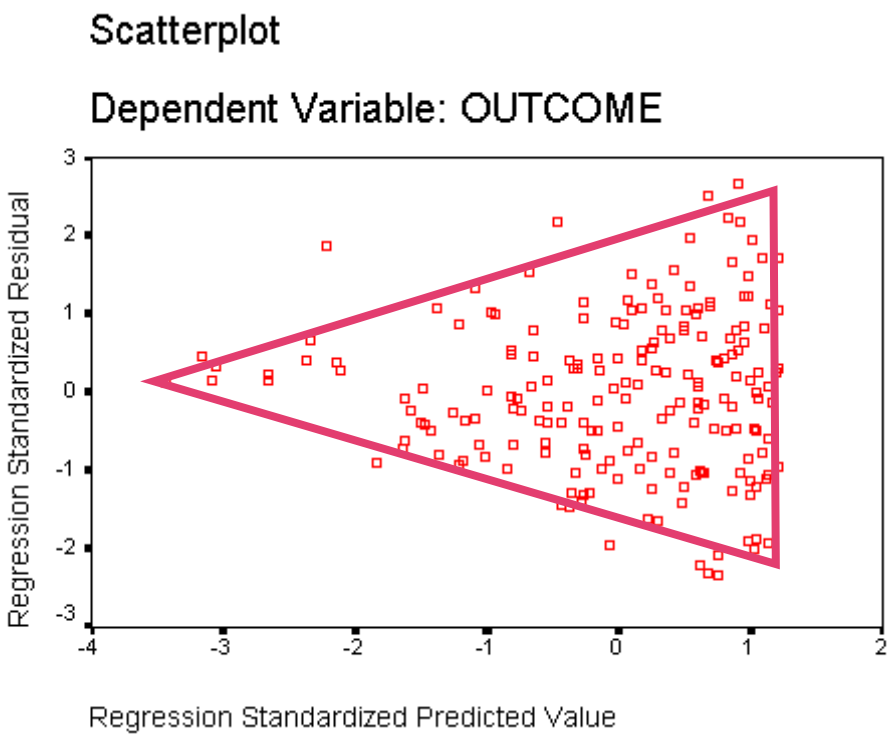
Varsayımlar…

7. Hata terimleri arasında korelasyon olmamalı. Bu amaçla Durbin-Watson testi yapılabilir. Test sonuçları 0-4 arası değişir ve 2 civarında ise sorun yok demektir.
8. Bağımsız değişken değerleri boyunca hata varyansı sabit olmalı (Homoscedasticity). Ayrıca hata verileri normal dağılmalı. Bu amaçla ZRESID xZPRED grafiğine bakabiliriz. Hataların normal dağılması için normal dağılımla ilgili grafiklere bakabiliriz

Homoscedasticity: ZRESID x ZPRED

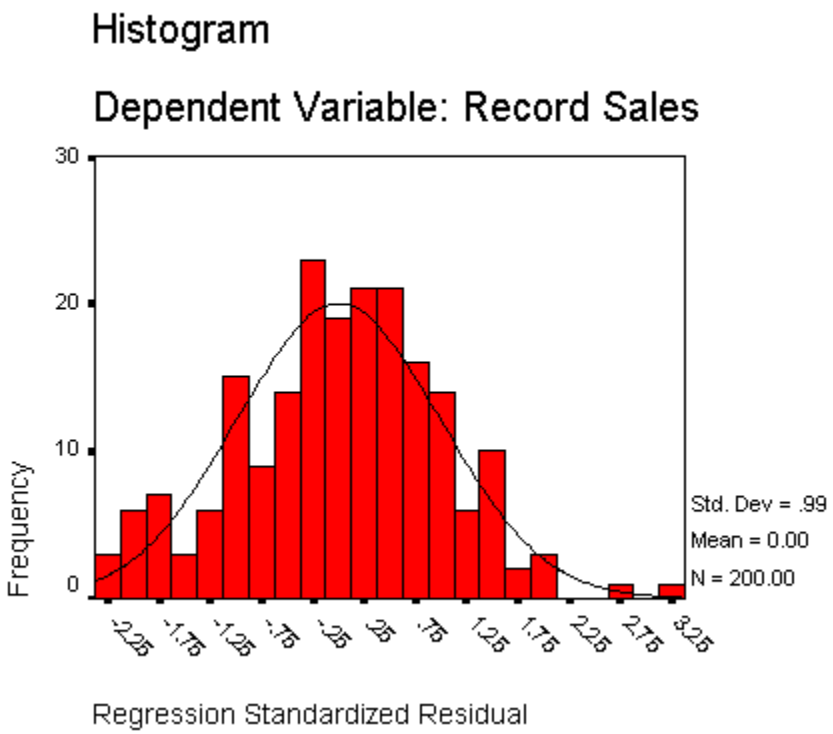


İyi

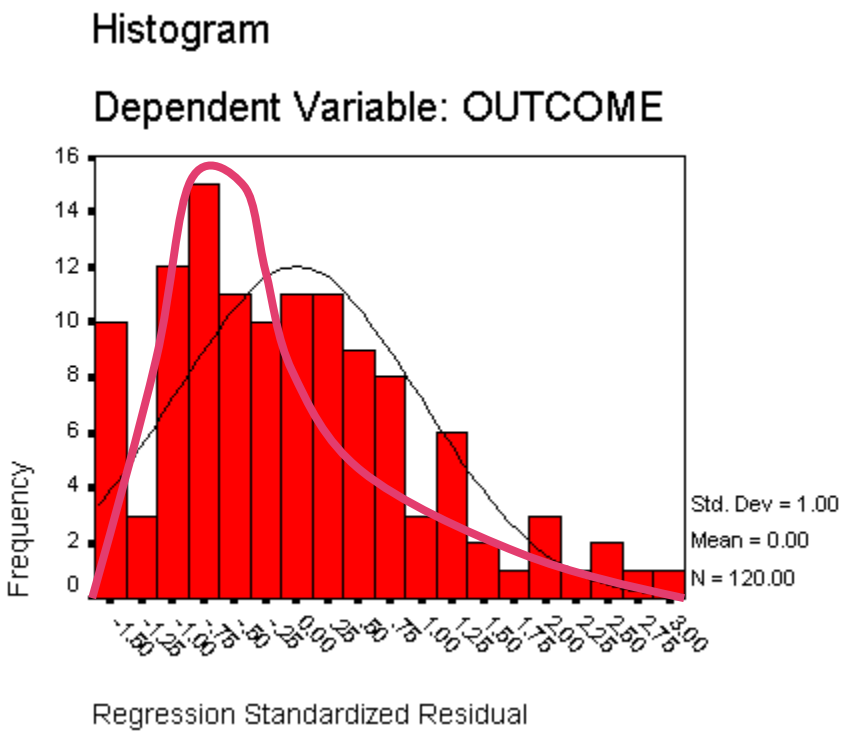


Kötü

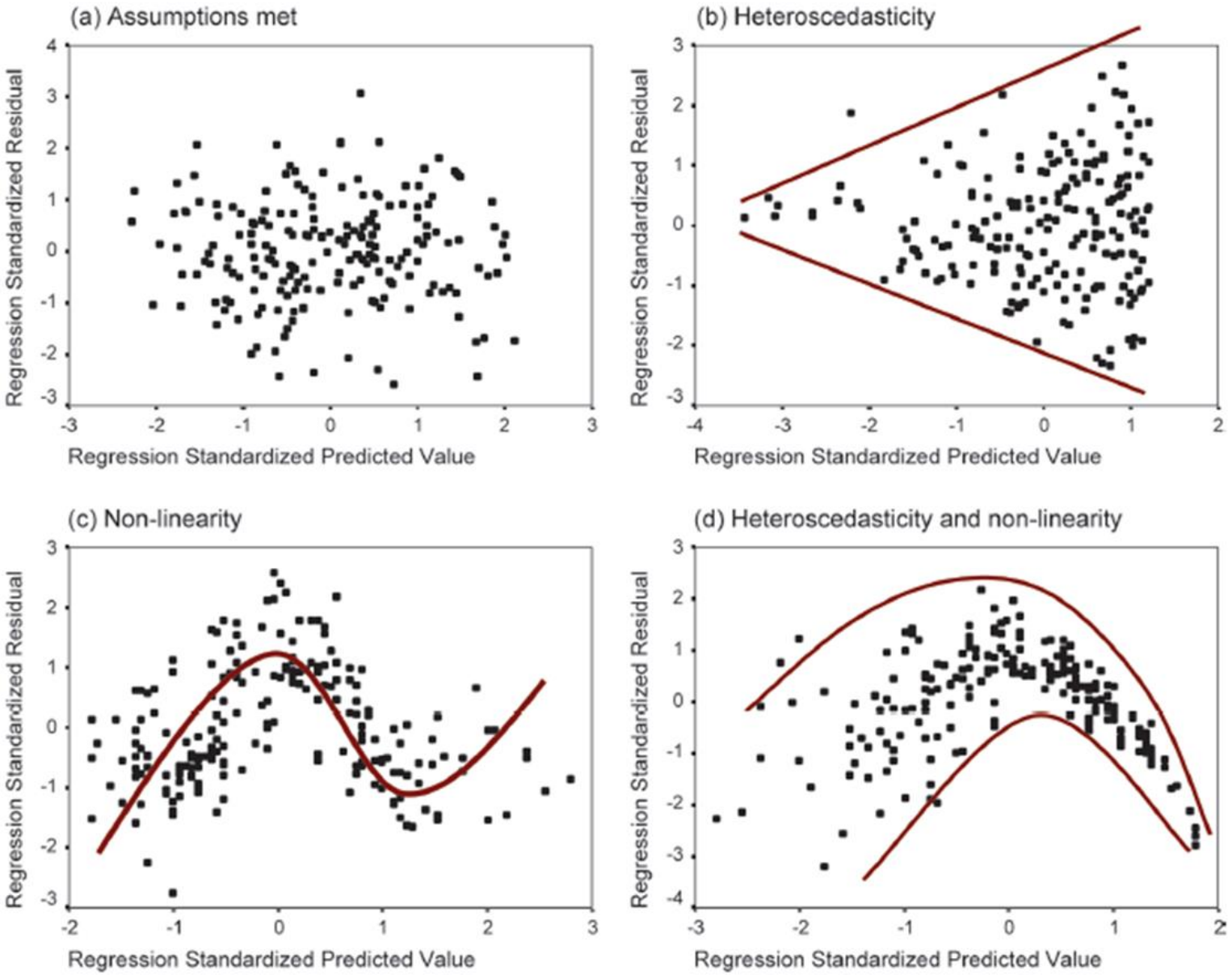
Hata terimleri dağılımı: Histogram



İyi



Kötü


(a) Assumptions met
Regression Standardized Residual
Regression Standardized Predicted Value
(b) Heteroscedasticity
Regression Standardized Residual
Regression Standardized Predicted Value
(c) Non-linearity
Regression Standardized Residual
Regression Standardized Predicted Value
(d) Heteroscedasticity and non-linearity
Regression Standardized Residual
Regression Standardized Predicted Value

Regresyon ve Örneklem Büyüklüğü