Regresyon Analizi

Regresyon; sözcük anlamıyla, **bir şeyi başka bir şeye bağlama** işi ve biçimidir. Bilimsel olarak regresyon terimi bir değişkenle başka bir (ya da birden çok) değişken arasında ilişki kurma işini ve ilişkinin biçimini anlatır.

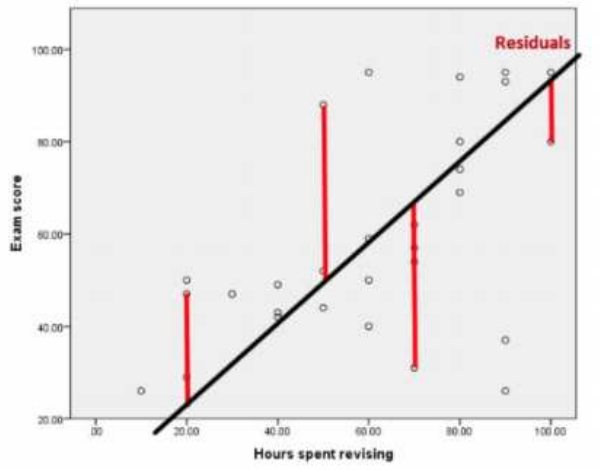
İstatistiksel anlamda iki değişken arasındaki ilişki, bunların değerlerinin **karşılıklı değişmeleri arasında bir bağlılık** şeklinde anlaşılır.

- (X) değişkeninin değerleri değişirken, buna bağlı olarak (Y) değişkeninin değerleri de değişiyorsa, bu iki değişken arasında bir ilişki olduğu söylenebilir.

Regresyon; bir bağımlı değişken (Y), diğeri de bağımsız değişken veya değişkenler (X) olmak üzere, **değişkenler arasındaki en uygun ilişkinin matematiksel bir fonksiyon şeklinde** ifade edilmesidir.

- Regresyonda **bir** bağımsız değişken olması durumunda **basit doğrusal regresyon**, **birden fazla** bağımsız değişken olması durumunda ise **çoklu doğrusal regresyon** adını alır.

Artık Değer (Residual)



Bir regresyon modelinin uygulanmasıyla ortaya çıkan, **model fonksiyonunun açıklayamadığı** değişkenliktir.

Değer olarak, bağımlı değişkenin gözlenen değeri ile tahmin edilen değeri arasındaki **sapmalar** için kullanılır.

Regresyon Varsayımları

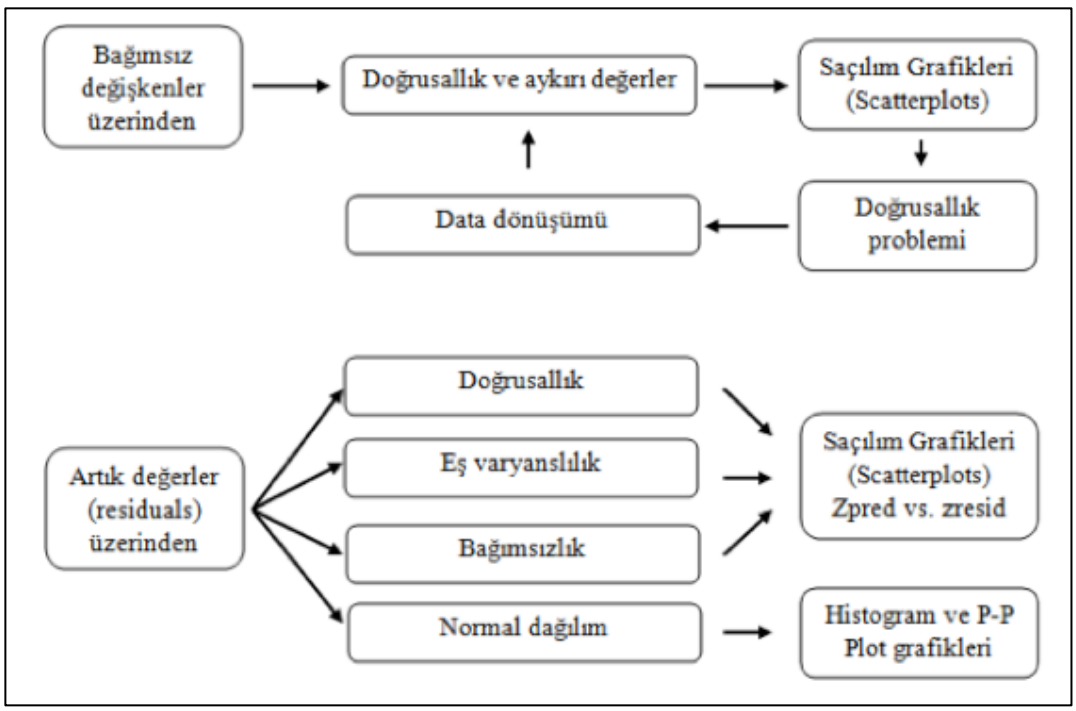
Analiz öncesi, veri seti ve değişkenler üzerinden değerlendirilen varsayımlar:

1. Örneklem yeterince büyüktür.
2. Bağımlı ve bağımsız değişkenler doğru türde olmalıdır.
3. Bağımsız değişkenler normal dağılmalıdır.
4. Bağımlı ve bağımsız değişkenler doğrusaldır (linearity).
5. Uç değerler (outliers) yoktur.

Analiz sonrası, artık değerler (residuals) üzerinden değerlendirilen varsayımlar:

1. Artık değerler bağımsızdır (autocorrelation).
2. Artık değer varyansları sabit ve homojendir (eş varyanslılık - homoscedasticity).
3. Artık değerler normal dağılır.
4. Çoklu doğrusallık (multi-collinearity) yoktur.
5. Modeli yanıltabilecek baskın (uç) değer yoktur.

Önemli varsayımların akış şeması üzerinden kontrolü (Field, 2012)



Analiz Öncesi Varsayım Değerlendirmeleri

(Veri seti ve değişkenler üzerinden yapılan değerlendirmeler)

Örneklem Büyüklüğü ve Değişkenlerin Türü

Örneklem, belli kurallara göre belli bir evrenden seçilmiş ve evreni temsil etmeye yeterli görülen küçük kümedir.

Örneklem büyüklüğü **arttıkça**, yapılan istatistiki analizlerin **güvenilirliği** artacaktır.

Güvenilir bir regresyon analizi için örneklem büyüklüğünün en az aşağıdaki değerden **büyük** olması gerekir (Tabachnick & Fidell, 2007):

N = 50 + 8 * (bağımsız değişken sayısı)

Regresyon analizi uygulanabilmesi için, bağımsız değişken(ler)in en az **kategorik**, bağımlı değişkenin ise **sürekli** olması gerekir.

Normal Dağılım Uyumu: Açıklayıcı ve Teorik Yöntemler

Diğer analizler gibi, regresyon analizinde de **normallik varsayımı** geçerlidir. Bu varsayıma uyulmadığı durumda Tip-1 hata ihtimali artar ve veri üzerinde çalışılarak normalliğe yakınsatmak gerekir.

Normal dağılım kontrolü için açıklayıcı ve teorik yöntemleri içeren, grafiksel ve nümerik metotlar mevcuttur.

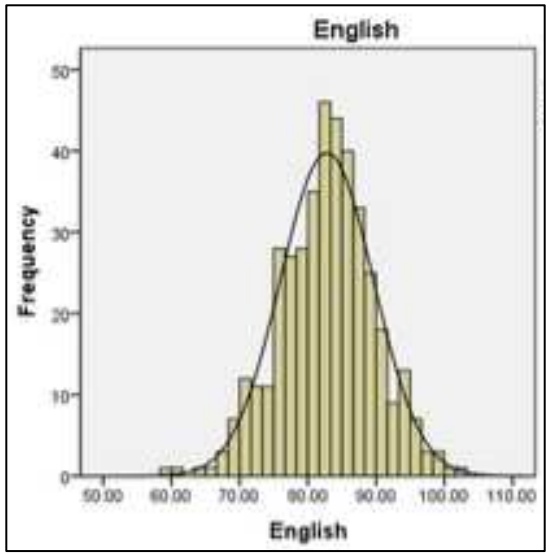
Table 1. Graphical Methods versus Numerical Methods

	Graphical Methods	Numerical Methods
Descriptive	Stem-and-leaf plot, (skeletal) box plot, dot plot, histogram	Skewness Kurtosis
Theory-driven	P-P plot Q-Q plot	Shapiro-Wilk, Shapiro- Francia test Kolmogorov-Smirnov test (Lillefors test)

Açıklayıcı Yöntemler (Normal Dağılım Uyumu)

Grafiksel olarak **histogram grafiği** üzerinden görsel yorumlama yöntemi, en yaygın kullanılan yöntemlerdendir.

SPSS: *Analyze – Descriptive Statistics – Explore: (Dependent List'e incelenecek olan değişkenlerin her analizde yalnız birisi atılır) Plots: "Histogram" / Output: Grafik gözle incelenir*



Eğrilik (skewness) ve **basıklık** (kurtosis) sayısal değerleri de normalliğin yorumlanması için kullanılır.

- İdeal değerleri 0 olmakla birlikte, -2 ile +2 arasındaki değerler de kabul edilebilir değerlerdir.

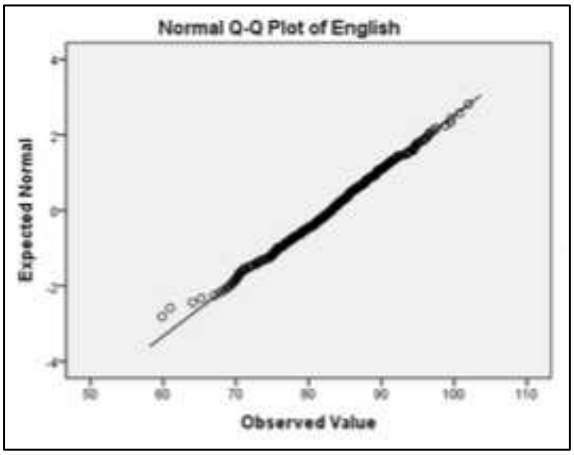
SPSS: *Analyze – Descriptive Statistics – Explore: (Dependent List'e incelenecek olan değişkenlerin her analizde yalnız birisi atılır) / Output: Descriptives tablosu incelenir.*

Descriptives

			Statistic	Std. Error
Height	Mean		68.0318	.26366
	95% Confidence Interval for Mean	Lower Bound	67.5135	
		Upper Bound	68.5501	
	5% Trimmed Mean		67.9687	
	Median		67.5700	
	Variance		28.363	
	Std. Deviation		5.32566	
	Minimum		55.00	
	Maximum		84.41	
	Range		29.41	
	Interquartile Range		6.78	
	Skewness		.230	.121
	Kurtosis		.113	.241

Teorik Yöntemler (Normal Dağılım Uyumu)

Bağımsız değişkenlerin değerlendirilmesinde en sık kullanılan teorik grafiksel yöntem ise **Q-Q grafiğidir** (quantile – quantile).



SPSS: *Analyze – Descriptive Statistics – Explore: (Dependent List'e incelenecek olan değişkenlerin her analizde yalnız birisi atılır) Plots: "Normality plots with tests" / Output: QQ grafiğinde saçılımın çizgiye yakınlığı gözlenir.*

Sayısal yöntemler arsında ise en yaygın yöntemler **Kolmogorov – Smirnov** ve **Shapiro – Wilk** analizleridir.

- Shapiro – Wilk analizi, genellikle **50'den az** sayıdaki örneklemlerde kullanılır ve 2000 örneklem üstünde çalışmaz.
- Her iki analizin de hipotezleri **«H0: Dağılım normal dağılımdan farklılaşmaz.»** şeklinde kurulur.

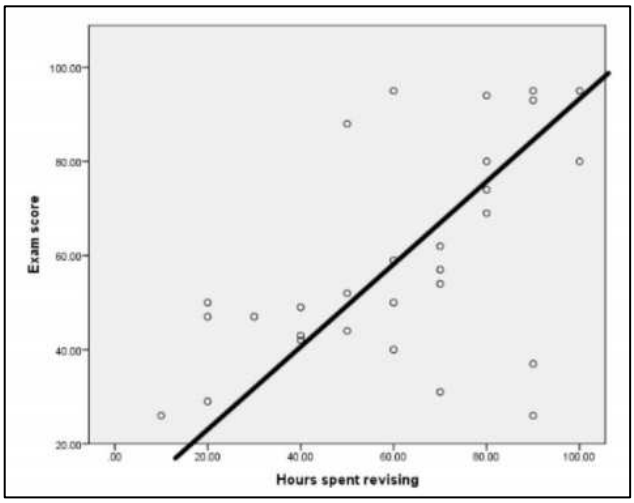
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Normal	.027	500	.200*	.996	500	.168

a. Lilliefors Significance Correction

SPSS: *Analyze – Descriptive Statistics – Explore: (Dependent List'e incelenecek olan değişkenlerin her analizde yalnız birisi atılır) Plots: "Normality plots with tests" / Output: Örneklem sayısına bağlı olarak ilgili analizin Sig. değerleri incelenir. p > 0,050 ise örneklem normal dağılmaktadır.*

Değişkenlerin Doğrusallığı (Linearity)



Regresyon analizi yapılabilmesi için, **bağımsız ve bağımlı değişkenler arasında doğrusal bir ilişki** olması gerekir.

Her bir bağımsız değişkenin, bağımlı değişken ile arasındaki doğrusallık **ayrı ayrı** kontrol edilmelidir.

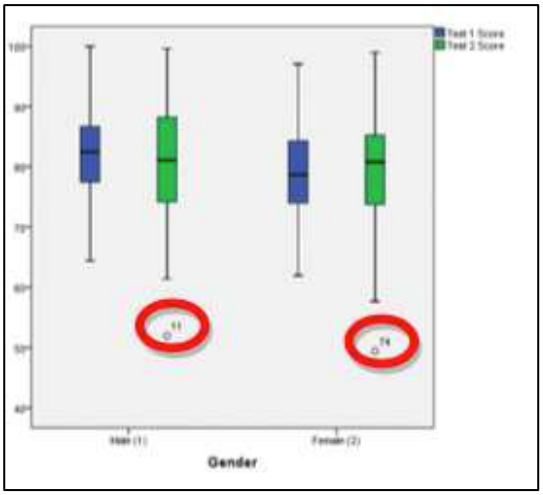
Doğrusallık, **saçılım grafiği** (scatterplot) veya **Pearson Korelasyon katsayısı (r)** ile gözlenebilir ve yorumlanabilir.

Oluşturulan doğrusal çizgiye **değerler ne kadar yakınsa, veri o kadar doğrusaldır**.

Analiz sonrasında, artık değerleri (residuals) de içeren çoklu doğrusallık mutlaka kontrol edilmelidir.

SPSS: *Graphs – Chart Builder: Gallery'den "Scatter / Dot" seçilir ve ilk seçenek üstteki panele sürüklenir – Y eksenine bağımlı, X eksenine bağımsız değişken sürüklenir / Output: Oluşan grafiğin doğrusallığı gözle kontrol edilir.*

Uç Değerlerin (outliers) İncelenmesi



Uç (aykırı) değerler, her türlü istatistiki analizde modelin **hatalı veya yanlı** tahmin yapmasına sebep olabilir.

Analize başlamadan önce, uç değerlerin varlığı **box-plot grafiği** ile tüm değişkenler için kontrol edilebilir.

Bulunan uç değerler, analiz kapsamından **çıkarılmalıdır**.

Her ne kadar analiz öncesinde kontrol yapılsa da, **analiz sonrasında artık değerler üzerinden yeniden kontrol yapılması** doğru olacaktır.

SPSS: *Analyze – Descriptive Statistics – Explore: (Dependent List'e incelenecek olan değişkenlerin her analizde yalnız birisi atılır) Statistics: "Outliers" / Output: Her bir değişken için oluşturulacak grafikte, grafik dışında değerlerin olup olmadığı gözle kontrol edilir.*

Analiz Sonrası Varsayım Değerlendirmeleri

(Artık değerler (residuals) üzerinden yapılan değerlendirmeler)

Artık Değerlerin Bağımsızlığı (Autocorrelation)

Regresyon analizi için, gözlenen değerlerin (bağımsız değişkenler) üzerinden hesaplanan **artık değerlerin arasında bir ilişki olmaması gerekir**.

Artık değerlerin arasında, birbirini izleyecek şekilde oluşabilecek **benzerliklere otokorelasyon** denir. Genellikle zamana bağlı bağımsız değişkenlerin bulunduğu durumlarda gözlenir

Otokorelasyonun tespiti için **Durbin – Watson test istatistiğinden** faydalanılır.

- Durbin – Watson test istatistiği, **0 ile 4** arasındaki değerlerden oluşur.
- Otokorelasyonun olmaması için bu değerin **2'ye yakın** olması gerekir.
- **1-3 arasındaki değerler**, önemli bir otokorelasyon olmadığı, yani **artık değerlerin bağımsız olduğu** yönünde yorumlanabilir.

SPSS: *Analyze – Regression – Linear: (Bağımsız ve bağımlı değişkenler, ilgili alanlara atılır) Statistics: "Durbin-Watson" / Output: Model Summary tablosundaki Durbin-Watson test istatistiği değeri incelenir.*

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.647[a]	.418	.373	17.97120	1.931

a. Predictors: (Constant), Enjoyment of subject, Hours spent revising

b. Dependent Variable: Exam score

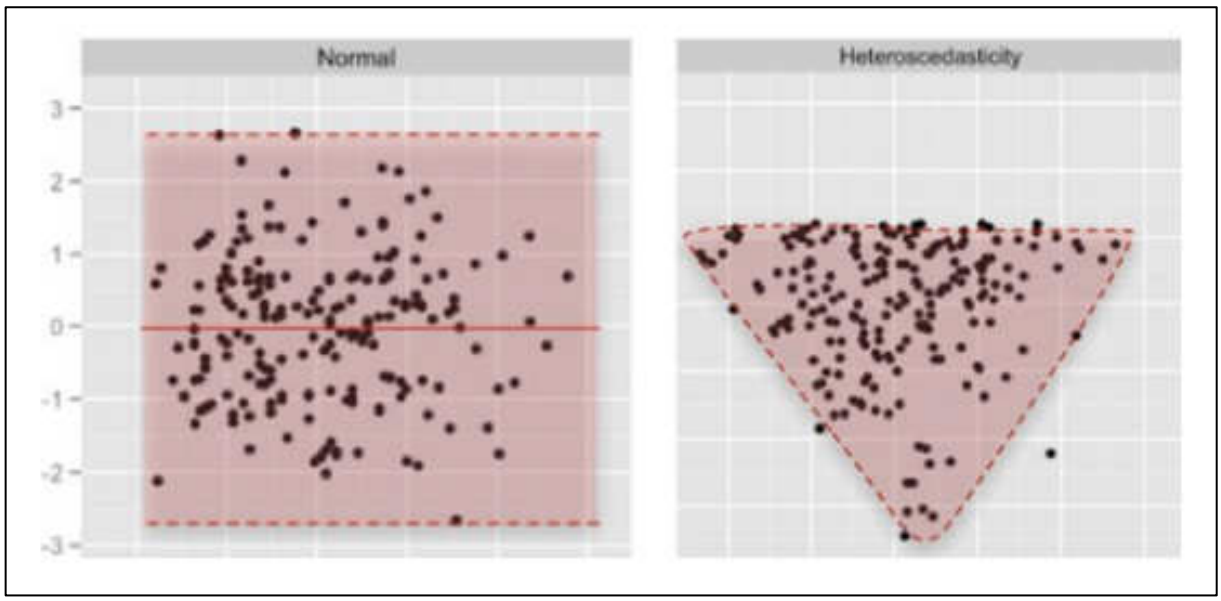
Artık Değer Varyanslarının Sabitliği (Eş Varyanslılık – Varyans Homojenliği – Homoscedasticity)

Regresyon analizinde elde edilen artık değer varyanslarının, **modelin her noktasında benzer** olduğu, bir başka deyişle **artık değer varyanslarının homojen olduğu** kabul edilir.

Artık değerlerin varyans homojenliğinin kontrolü için, **saçılım grafiği** (scatterplot) kullanılır.

- Üst ve alt değerlerin arasındaki mesafe korunuyor ve gözle görülür bir **«hunileşme» yoksa**, eş varyanslılık (yani homojenlik) olduğu söylenebilir.

SPSS: *Analyze – Regression – Linear: (Bağımsız ve bağımlı değişkenler, ilgili alanlara atılır) Plots: ZRESID değeri Y'e, ZPRED değeri X'e atılır. / Output: Scatterplot grafiğinin homojenliği gözle incelenir.*

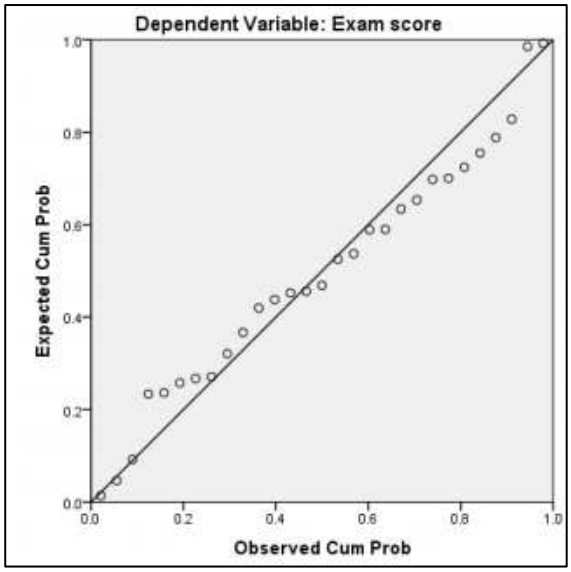


Artık Değerlerin Normalliği

Tıpkı bağımsız değişkenler gibi, regresyon analizinde **artık değerler de normal dağılmalıdır**.

Artık değerlerin normalliği, tıpkı değişkenlerde olduğu gibi **histogram grafiği** ile gözle yorumlanabilir.

Bir başka yöntem ise, **P-P grafiğinin** (probability – probability) kullanılmasıdır.



- P-P grafiği, Q-Q grafiğinden farklı olarak değerler değil, **olasılıklar** üzerinden çalışır.
- Grafikteki **veriler diyagonal çizgiye ne kadar yakınsa**, artık değerler o kadar normal dağılıma yakındır.

SPSS: *Analyze – Regression – Linear: (Bağımsız ve bağımlı değişkenler, ilgili alanlara atılır) Plots: "Normal probability plot" / Output: Normal P-P grafiğinin çizgiye yakınlığı gözle incelenir.*

Çoklu Doğrusallık Kontrolü (Multi-Collinearity)

Bağımsız değişkenlerin arasında **belli bir seviyenin üzerinde ilişki olması**, regresyon modelinin **güvenilirliğini olumsuz yönde** etkiler.

Regresyon analizi sonucunda elde edilen **Pearson Correlation katsayılarının 0.8'in altında olması**, bağımsız değişkenlerin kendi arasında çoklu doğrusallık olmadığını gösterir.

Yine regresyon analizi ile elde edilen Collinearity istatistiği altındaki **«Tolerance»** ve **«VIF»** değerleri, artık değerleri de katarak bir çoklu doğrusallık olup olmadığını gösterir.

- Çoklu doğrusallık olmaması için, **Tolerance'ın 0.2'nin üstünde, VIF değerinin ise 10'un altında** olması gerekir.

SPSS: *Analyze – Regression – Linear: (Bağımsız ve bağımlı değişkenler, ilgili alanlara atılır) Statistics: "Collinearity Diagnostics" / Output: Coefficients tablosunda Tolerance ve VIF değerleri incelenir.*

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	20.647	9.530		2.167	.040		
	Hours spent revising	.295	.154	.333	1.911	.067	.735	1.360
	Enjoyment of subject	.668	.285	.408	2.342	.027	.735	1.360

a. Dependent Variable: Exam score

Uç Değerlerin (outliers) Modeli Yanıltması

Her ne kadar box-plot üzerinden uç değerler analiz öncesinde incelense de, bazı istatistiki değerler ile uç değerlerin model üzerinde etkisi olup olmadığı da gözlenmelidir.

Mahalanobis uzaklığı (Mahal. Distance), **her bir noktanın ortalamadan uzaklığıdır**. Bu anlamda, maksimum Mahalanobis uzaklığı, en uç değeri gösterir.

- En büyük Mahalanobis uzaklığının, **ilgili bağımsız değişken sayısının (df), alfa 0.001 kritik seviyesindeki Ki-Kare değerinden küçük** olması beklenir.

Cook's uzaklığı (Cook's Distance), uç değerlerin modele olan etkisinin gücünü gösterir.

- Cook's Distance değeri **1'den büyük ise, modelde uç değerlerin baskın etkisi olduğu ve modeli yanılttığı** söylenebilir.

SPSS: *Analyze – Regression – Linear: (Bağımsız ve bağımlı değişkenler, ilgili alanlara atılır) Save: "Mahalanobis"ve "Cook's" / Output: Residuals Statistics tablosunda ilgili değerler incelenir, eğer olması gerektiğinden büyükse, SPSS Data View alanına eklenmiş olan ilgili alanlardan büyük değerlere sahip veriler çıkarılarak yeniden analiz yapılır.*

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Mahal. Distance	.066	18.649	2.985	2.972	200
Cook's Distance	.000	.071	.005	.010	200
Leverage Value	.000	.094	.015	.015	200

a. Dependent Variable: Album Sales (Thousands)

Regresyon Varsayımları SPSS Cheat Sheet

Analiz öncesi:

- Analyze – Descriptive Statistics – Explore:
  - İncelenecek her değişkenin Dependent List alanına ayrı ayrı atılarak farklı analizler çalıştırılması, daha doğru sonuç verecek ve karışıklığın önüne geçecektir.
  - Statistics: "Outliers" seçimi ile uç değer kontrolü yapılır
  - Plots: Desctriptive alanından "Stem-and-leaf" seçimi kaldırılır ve "Histogram" seçilir, "Normality plots with tests" seçilir, böylece açıklayıcı ve teorik normallik testleri yapılır.
- Graphs – Chart Builder:
  - Alttaki grafik türü seçimi alanından Scatter / Dot seçilir ve ilk seçenek üstteki "Preview" alanına sürüklenir.
  - Y eksenine bağımlı, X eksenine bağımsız değişkenlerden birisi sürüklenir. Böylece lineerlik kontrolü yapılır.

Analiz sonrası, residuals üstünden:

- Analyze – Regression – Linear:
  - Bağımlı değişken "Dependent", bağımsız değişken(ler) "Independent(s)" alanına atılır.
  - Statistics: "Descriptives", "Collineraity diagnostics" ve "Durbin-Watson" seçilir. Böylece çoklu doğrusallık ve otokorelasyon verileri Output dosyasına yansır.
  - Plots: ZRESID değeri Y, ZPRED değeri X alanına taşınır, "Normal probability plot" seçilir. Böylece varyans homojenliği ve P-P grafiği oluşturulur.
  - Save: Distances altından "Mahalanobis" ve Cook's" seçilir. Böylece hem Output dosyasına uç değerlerle ilgili test istatistikleri eklenir, hem de orijinal verilerin görüldüğü "Data View" sayfasına her değişken için ilgili Mahalanobis ve Cook's değerleri yansır.

Örnek veri: bit.ly/regresyondata