**Regresyon Analizinin Varsayımları**

İki veya daha fazla değişken arasındaki ilişkileri ölçmek için kullanılır. Hem tanımlayıcı hem de sonuç çıkarıcı istatistik sağlar.

Bağımlı değişken bir, bağımsız değişken bir ise yöntem Basit Regresyon Analizi,  
Bağımlı değişken bir, bağımsız değişken iki veya daha fazla ise yöntem Çoklu Regresyon Analizi,  
Bağımlı değişken iki yada daha fazla ise yöntem Çok Değişkenli Regresyon Analizidir.

Regresyon modelini kurarken her zaman bir bağımlı değişkenden ve bir ya da birçok bağımsız (tahmin değişkeni) değişkeni söz konusudur.

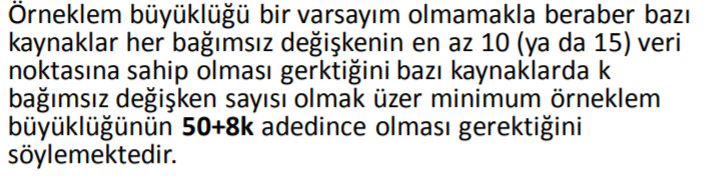
**BAĞIMSIZ DEĞİŞKENLER / TAHMİN DEĞİŞKENLERİ**

Birçok tahmin de ğişkeni olan karma şık bir model kurmak istiyorsak hangi tahmin de ğişkenlerini seçeceğimize nasıl karar vereceğiz? • Hangi tahmin değişkenlerinin seçildi ği ve bunların modele nasıl girildi ği önemli. • Rastgele çok sayıda tahmin değişkeni seçmek do ğru değil. • Peki tahmin değişkenlerini modele nasıl gireceğiz?

Hiyerarşik ekleme (Blockwise entry): (Önceki çalışmaların sonuçlarına dayanarak) en önemli tahmin değişkeni önce girilir. • Zorla ekleme (Enter): Tüm tahmin değişkenleri eş zamanlı olarak modele girilir • Adım adım ekleme: Tamamen matematiksel ölçütlere göre girilir. Bilgisayar bağımlı değişkendeki değişimi en fazla açıklayan tahmin değişkenini bulur, sonra geri kalan değişimi en fazla açıklayan tahmin değişkenini bulur, vs. (adım adım eklemede backward yöntemini seçmek daha uygun).

İki önemli soru: – Model gözlenen verilere ne kadar iyi uyuyor veya model az sayıda vakadan etkileniyor mu? – Model diğer örneklemlere genellenebilir mi? • İlk sorunun yanıtı için uç de ğerlere (outliers) ve model tarafından tahmin edilen de ğerlerle gözlenen de ğerler arasındaki farka (residuals) bakılmalı. • Uç de ğerler regresyon e ğrisini gerçekte oldu ğundan farklı gösterir.

**Örnek Kütle Büyüklüğü**



**Varsayımları**

1. İki değişken de sürekli veri niteliğinde olmalıdır.

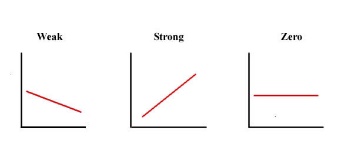
Tüm tahmin değişkenleri aralıklı/oranlı (nicel) veya kategorik olmalı (iki kategori), çıktı/sonuç değişkeni nicel, sürekli olmalı (çıktı 1-10 arasında değişiyorsa sürekli ama toplanan veri 3-7 arasındaysa veri kesikli demektir).

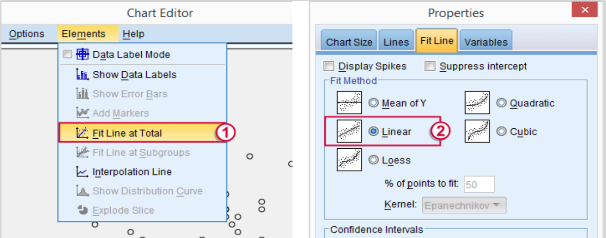
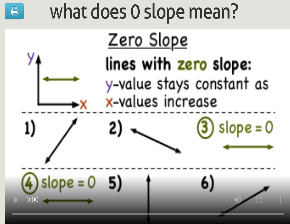
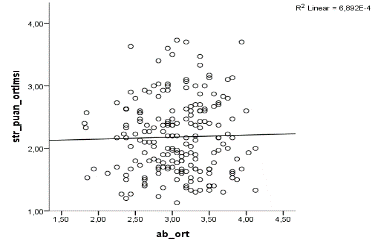
1. İki değişken arasında doğrusal bir ilişki olmalıdır. SPSS’te nokta dağılım grafiği-saçılım (scatterplot) grafiği ile incelenir. Çizginin iki tarafında yer alan noktalar 45 derece açıyla yükselmeli veya ilişki negatif ise 45 derece açıyla aşağıya doğru düşen bir eğilim göstermelidir. Eğer ilişkiler doğrusal değil, eğrisel ise non-linear regression analysis yöntemlerinden polynomial regression yönteminden yararlanılır. Veya verilerin dönüştürme yoluna başvurulur. Dönüştürme yöntemi de sağlıklı bir sonuç vermiyorsa STATA yazılımındaki Kernel Regression analizinden yararlanılır.

Graphs 🡪 Scatterplot 🡪 Simple Scatter X ve Y değişkeni alanlarına bağımlı (sonuç değişkeni) ve bağımsız değişkenlerin (tahmin değişkenleri) atanması

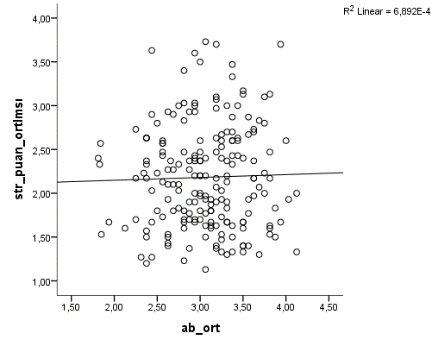
If all points fall directly on a straight line, we have a perfect linear relationship between our two variables.

Fit line at total immediately adds the desired regression line to our scatterplot. We don't have to change any of the default settings; we can just Close the dialog



Doğrusal bir ilişki olmadığı anlaşılmıştır. Doğrusal ilişkide korelsasyon katsayıları yüksek çıkardı.  If no linear relationship exists, then linear regression is an inaccurate representation of the data



**Şekil 1.** Saçılım grafiği ile doğrusallık testi.

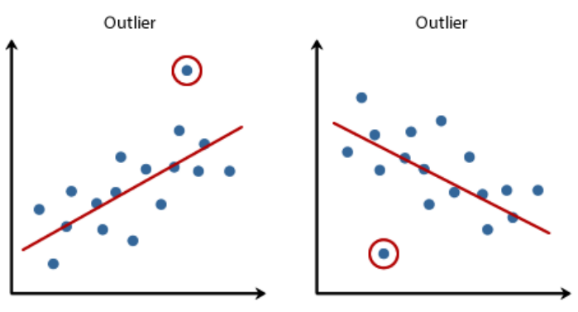
I guess you mean by non-linear that in a scatterplot a straight line does not well describe the relationship between two variables!? In that case there are plenty of options to model such a non-linear pattern (in R e.g. nls, GAMs, glm). GLMs are often applied because the measurement scale of your response variable and related error structure can not be adequately described with a normal distribution (e.g. count or binary data), not necessarily because the data is "non-linear" (as defined above). So the choice of a specific GLM results from the kind of response variable you have.

**Violations of linearity or additivity** are extremely serious: if you fit a linear model to data which are nonlinearly or nonadditively related, your predictions are likely to be seriously in error, especially when you extrapolate beyond the range of the sample data.

**How to diagnose**: nonlinearity is usually most evident in a plot of**observed versus predicted** **values** or a plot of **residuals versus predicted values**, which are a part of standard regression output. The points should be symmetrically distributed around a diagonal line in the former plot or around horizontal line in the latter plot, with a roughly constant variance.  (The residual-versus-predicted-plot is better than the observed-versus-predicted plot for this purpose, because it eliminates the visual distraction of a sloping pattern.)  Look carefully for evidence of a "bowed" pattern, indicating that the model makes systematic errors whenever it is making unusually large or small predictions. In multiple regression models, nonlinearity or nonadditivity may also be revealed by systematic patterns in plots of the **residuals versus individual independent variables**.

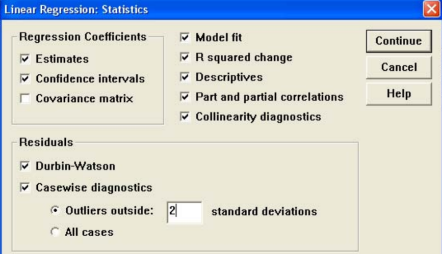
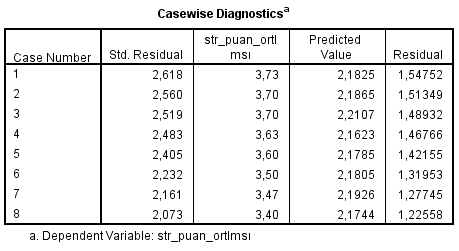
**How to fix:** consider applying a *nonlinear transformation*to the dependent and/or independent variables *if* you can think of a transformation that seems appropriate. (Don’t just make something up!) For example, if the data are strictly positive, the log transformation is an option.  (The logarithm base does not matter--all log functions are same up to linear scaling--although the natural log is usually preferred because small changes in the natural log are equivalent to percentage changes.  See [these notes](http://people.duke.edu/~rnau/411log.htm) for more details.)  If a log transformation is applied to the dependent variable only, this is equivalent to assuming that it grows (or decays) exponentially as a function of the independent variables.  If a log transformation is applied to *both* the dependent variable and the independent variables, this is equivalent to assuming that the effects of the independent variables are *multiplicative* rather than additive in their original units. This means that, on the margin, a small *percentage* change in one of the independent variables induces a proportional *percentage* change in the expected value of the dependent variable, other things being equal.  Models of this kind are commonly used in modeling price-demand relationships, as illustrated on the [beer sales example](http://people.duke.edu/~rnau/regex3.htm) on this web site.

1. LR-Üçüncü varsayım değişkenlerde ayrık değer bulunmamasıdır. Ayrık değerler nokta dağılım grafiğinde diğer noktalardan oldukça uzak konuma düşen vakalardır. Ayrık değerler regresyon analizinde negatif etkiye sahiptir. Tahminin gücünü ve etkisini azaltır.



Ayrık değerleri tespit etmenin 100 farklı yöntemi olduğundan söz ediliyor.

Ayrık değer regresyon analizinin kendi içinde de tespit edilebilir. Case’leri sort etmiş olduğumdan vaka numaraları arka arkaya geldi. 2 standart sapmadan hareket edersek Ortalaması 3,40’ın üzerindeki bütün değerleri ayrık değer olarak kabul etmemiz gerekiyor.

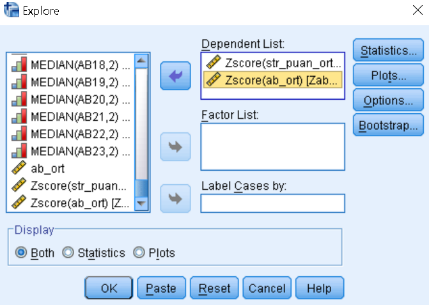
Örnekte 8 vakada ayrık değer olduğu görülmektedir. Bunları analizden çıkarabiliriz. Fakat 8 vaka çok değil. Ayrıca 3 standart sapmayla çalışırsak bu 8 vaka bütünüyle ortadan kalkmaktadır. Çeşitliliği temsil etmesi açısından bu vakalar bu haliyle kalabilir veya katı kriterlerle çalışmak istiyorsak çıkarılabilir. Ayrık değerler konusunda ne yapacağımıza birkaç hesaplama yaptıktan sonra karar vermemiz daha doğru olabilir. Burada üç arklı yöntem üzerinde durulmuştur. Regresyon analizinin kendi ayrık değer hesaplaması, Todd’un önerdiği birinci yaklaşım ve ikinci yaklaşım. Birinci yaklaşımın hesaplanması daha kolay gözüküyor.

Veya aşağıdaki yöntemle hesaplanır.

Dr. Todd Grandenin Youtube kaynağı: İki yöntem.

1 --- Birincisi kompozit ölçek değişkenlerini Descriptive/Descriptive menüsüyle *z puanlarına* dönüştürme ve çıkan değerlerin -2,68 ve 2,68 arasında kalıp kalmadığını inceleme. Bağımlı ve bağımsız değişkenlerin her ikisini de bağımlı değişken penceresine alıyoruz. Sonuçları ascending veya descending ile sıralamaya sokuyoruz ve uç değer içeren vakaları tespit ediyoruz. Bunları çıkarabiliriz veya ortalama puanlarında değişiklik yapabiliriz. Çıkarma yerine değişiklik yapma daha doğru. Böylece vakamız azalmamış oluyor.

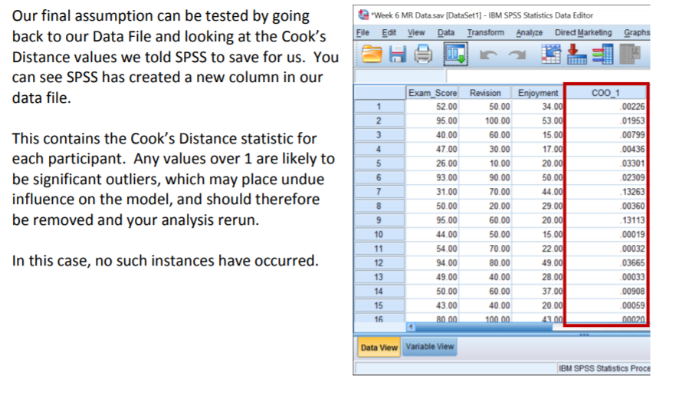
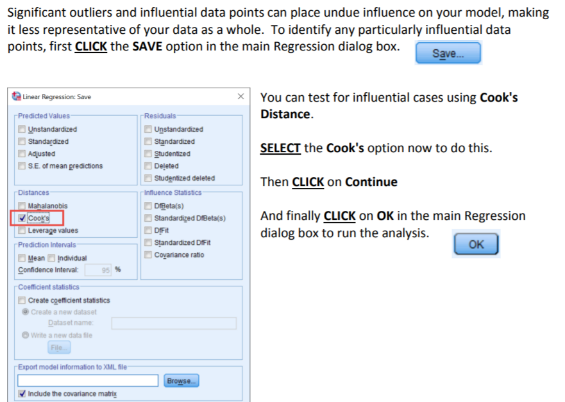
2 --- İkinci yöntem Descriptive>Explore> Box Plot yöntemiyle ayrık değerleri görme.



Bağımlı ve bağımsız değişkenlerinin her ikisini de bağımlı değişken alanına alıyoruz ve Box plot grafiklerini inceliyoruz.

<https://www.youtube.com/watch?v=ltRobzmOzsw>

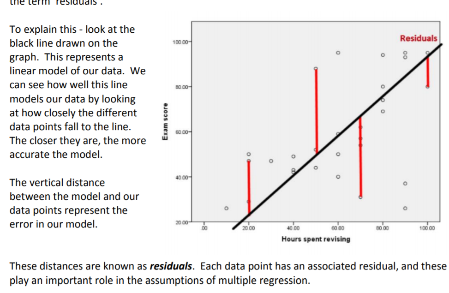
**Dördüncü yöntem**: Bir başka ayrık değer hesaplaması

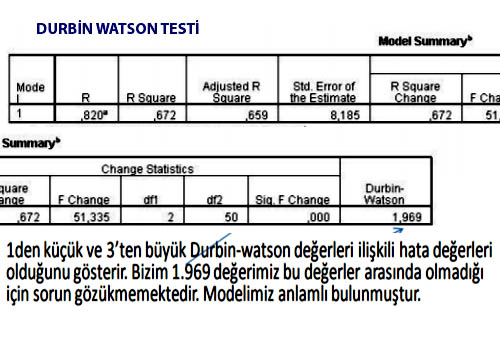


4. LR- Gözlem verilerinde saptanan hataların birbirinden bağımsız olması. Kısaca, “hatalarının bağımsız olması” olarak adlandırılır.  (**independence** of the errors (no serial correlation)

Regresyon analizi içindeki “statistics” düğmesi altında bulunan Durbin-Watson istatistiği ile belirlenir. **Durbin**-**Watson** statistic to test the assumption that our **residuals** are independent (or uncorrelated). Esas olarak “gözlemlerin bağımsız olmasını ölçen” bir ön kabul. Durbin Watson test is used to check for auto-correlation in the residuals of regression analysis.

Residuals



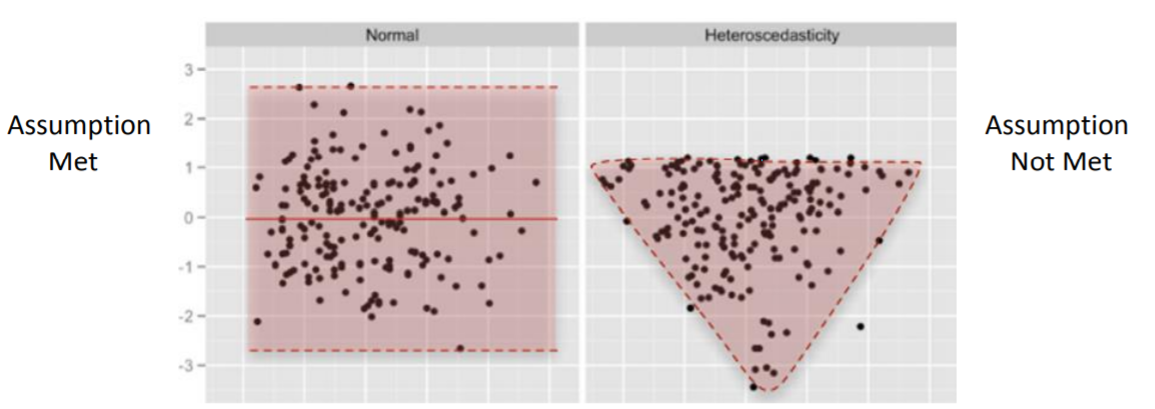


A rule of thumb is that test statistic values in the range of 1.5 to 2.5 are relatively normal. Values outside of this range could be cause for concern. Field(2009) suggests that values under 1 or more than 3 are a definite cause for concern.

The Durbin-Watson statistic is always between 0 and 4. A value of 2 means that there is no autocorrelation in the sample. Values from 0 to less than 2 indicate positive autocorrelation, whereas values from more than 2 to 4 indicate negative autocorrelation. If observations are made over time, it is likely that successive observations are related. If there is no autocorrelation (where subsequent observations are related), the Durbin-Watson statistic should be between 1.5 and 2.5. The Durbin-Watson statistic is 2.39 which is between 1.5 and 2.5 and therefore the data is not autocorrelated.

**Violations of independence** are also very serious in *time series regression*models: serial correlation in the residuals means that there is room for improvement in the model, and extreme serial correlation is often a symptom of a badly mis-specified model, as we saw in the auto sales example. Serial correlation is also sometimes a byproduct of a violation of the linearity assumption--as in the case of a simple (i.e., straight) trend line fitted to data which are growing exponentially over time.

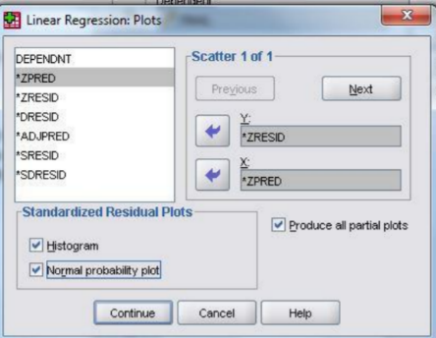
5. Değişkenlere ait residuals (artık değer-hata) varyanslarının her iki bölgede sabit olması. Denk varyanslılık… Residual’ların her iki bölgede sabit veya birbirine benzer uzaklıkta olması anlamına geliyor… Veya modeldeki hata miktarının modelin her iki tarafındaki noktalarda birbirine benzerlik göstermesidir. **Homoscedasticity.** Karşılanmasa bile analize devam edilir.



**To check is homoscedasticity. The scatterplot of the residuals…..**

**SPSS’te scatterplot grafiğinin çizimi: Regresyon analizi / Plot/ Y=ZRESİD, X=ZPRED/**



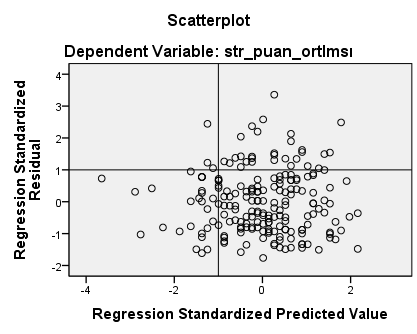


Üç grafik elde edilir. Histogram, P-P grafiği ve asıl Saçılım grafiği. Histogram grafiğini seçmeyebiliriz de. Histogram verilerin Normal dağılım özelliği göstermesi açısından önemli.

PP grafiği (Normal Probability Plot) bir sonraki 6. Varsayımda kullanılacaktır. O açıdan önemli.

Bizi bu aşamada **Homoscedasticity** varsayımı açısından asıl ilgilendiren “Nokta dağılım grafiği” veya “saçılım/Scatterplot” grafiğidir.

Scatterplot Grafiğin yorumlanması: (1) Noktaların bir desen oluşturmaması….. (2) yatay ve dikey düzlemde noktaların dağılımlarının dengeli olması. Burada çok dengeli bir dağılım görülmüyor Noktalar daha çok sağ alt hücrede toplanmış gibi. Toplanma “desen oluşumuna” işaret olarak değerlendirilebilir. Eş varyanslılık koşulu tam sağlanmamış gibi….

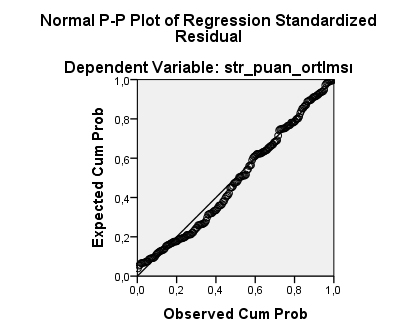


**Violations of homoscedasticity** make it difficult to gauge the true standard deviation of the forecast errors, usually resulting in confidence intervals that are too wide or too narrow. In particular, if the variance of the errors is increasing over time, confidence intervals for out-of-sample predictions will tend to be unrealistically narrow. Heteroscedasticity may also have the effect of giving too much weight to small subset of the data (namely the subset where the error variance was largest) when estimating coefficients.

6. LR- Artık değerlerin (hataların) yaklaşık olarak normal dağılım özelliği göstermesi gerekir. (**Normality** of the error distribution)

Bunu P-P Plot grafiği ile test ederiz.

Hatalar normal dağılmalı (yani regresyon modeli ile gözlenen veriler arasındaki farklar sıfır ya da sıfıra yakın olmalı)



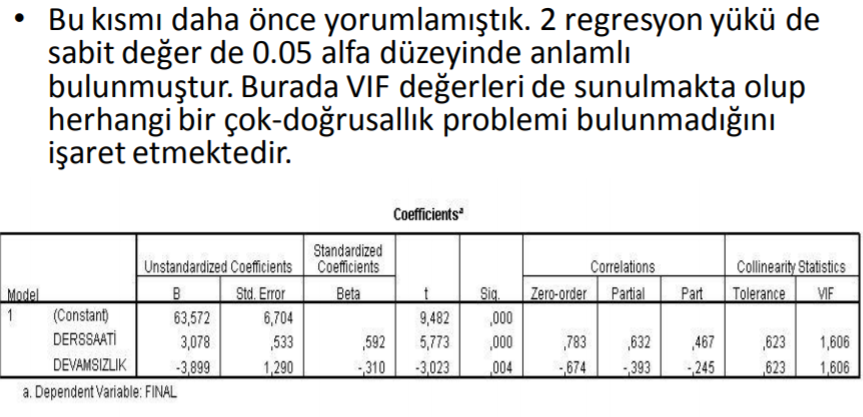
You will see a diagonal line and a bunch of little circles. Ideally, your plot will look like the two leftmost figures below. If your data is not normal, the little circles will not follow the normality line.

It seems the test is relatively robust for moderate nonnormality and moderately large sample size. These effects are also found to have insignificant influence from the regressors.

**Violations of normality** compromise the estimation of coefficients and the calculation of confidence intervals. Sometimes the error distribution is "skewed" by the presence of a few large outliers. Since parameter estimation is based on the minimization of*squared* error, a few extreme observations can exert a disproportionate influence on parameter estimates. Calculation of confidence intervals and various signficance tests for coefficients are all based on the assumptions of normally distributed errors. If the error distribution is significantly non-normal, confidence intervals may be too wide or too narrow.

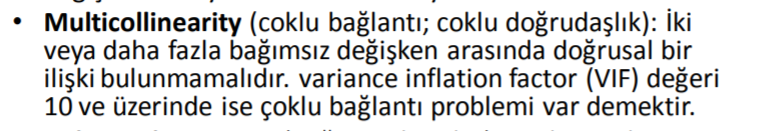
How to fix: violations of normality often arise either because (a) the distributions of the dependent and/or independent variables are themselves significantly non-normal, and/or (b) the linearity assumption is violated. In such cases, a nonlinear transformation of variables might cure both problems. In some cases, the problem with the residual distribution is mainly due to one or two very large errors. Such values should be scrutinized closely: are they genuine (i.e., not the result of data entry errors), are they explainable, are similar events likely to occur again in the future, and how influential are they in your model-fitting results?

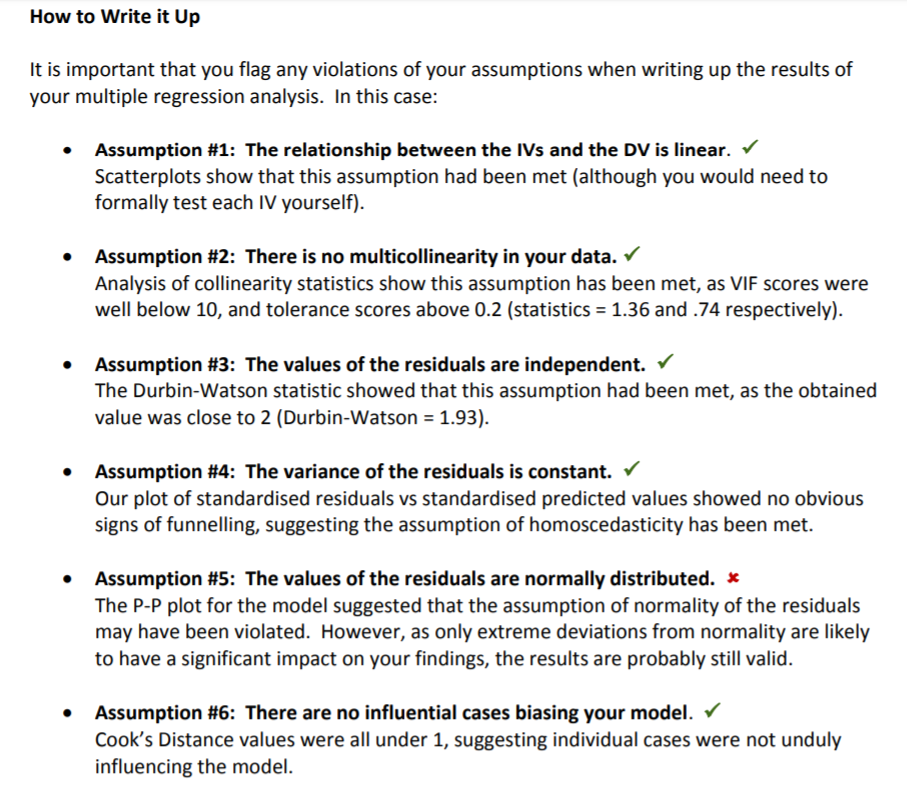
**7. Çoklu bağlantı varsayımı** (Çoklu regresyon analizi yapılırsa bağımsız değişkenlerin kendi aralarında korelasyon olmaması demektir). Regresyon analizi sonucu elde edilen SPSS çıktılarındaki VIF ve Tolerans değerlerine bakılarak karar verilir.



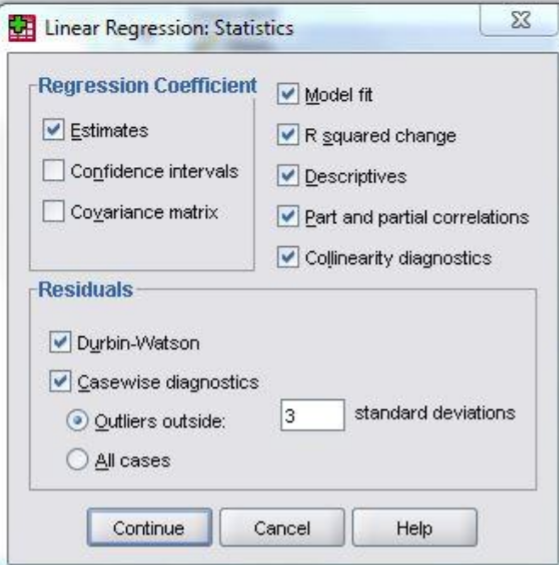
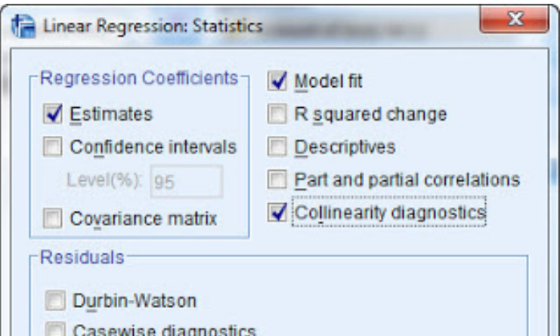
**VIF values: Each value is below 10, indicating that the assumption is met.**

**Tolerans değeri açısından yüzde 10’dan büyük bir değer elde edilmelidir.  >0.10 for Tolerance.**



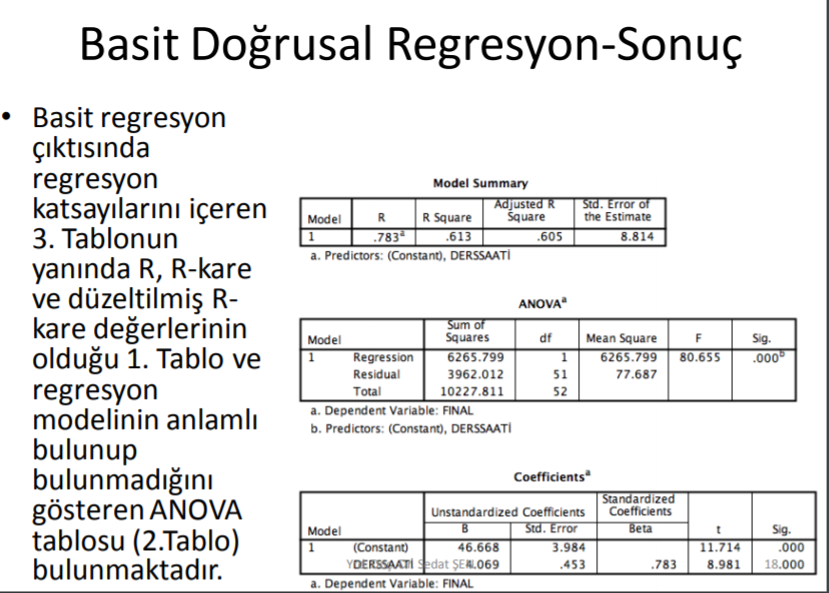


**REGRESYON ANALİZİNİN HESAPLANMASI**

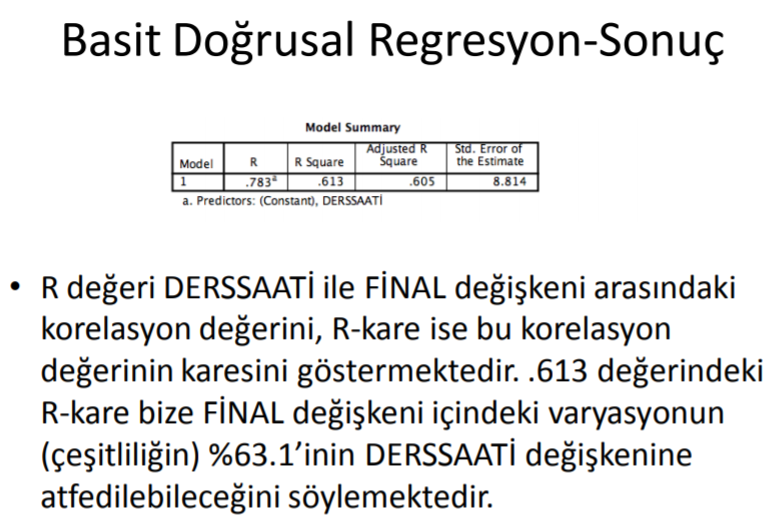


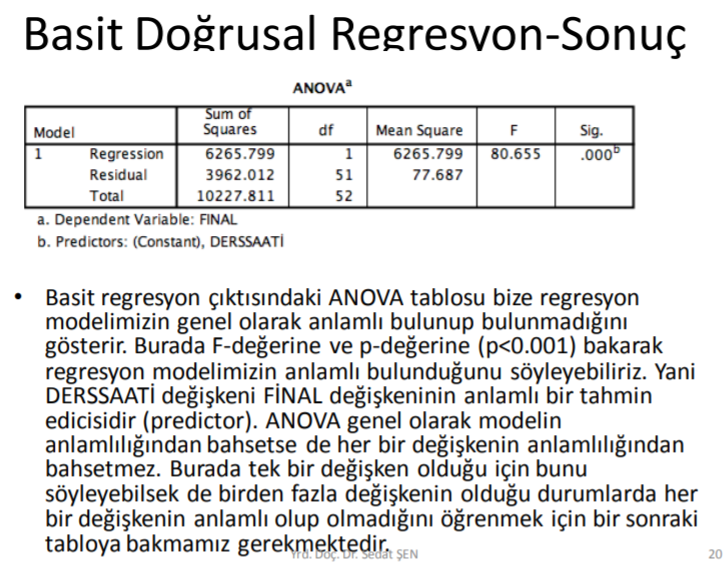
Tahmin değişkenleri arasında mükemmel doğrusal ilişkiler olmamalı (o zaman aralarında doğrusal ilişki olan değişkenler için b değeri aynı olur; R’nin büyüklüğünü sınırlar; hangi değişkenin önemli olduğunu söyleyemeyiz

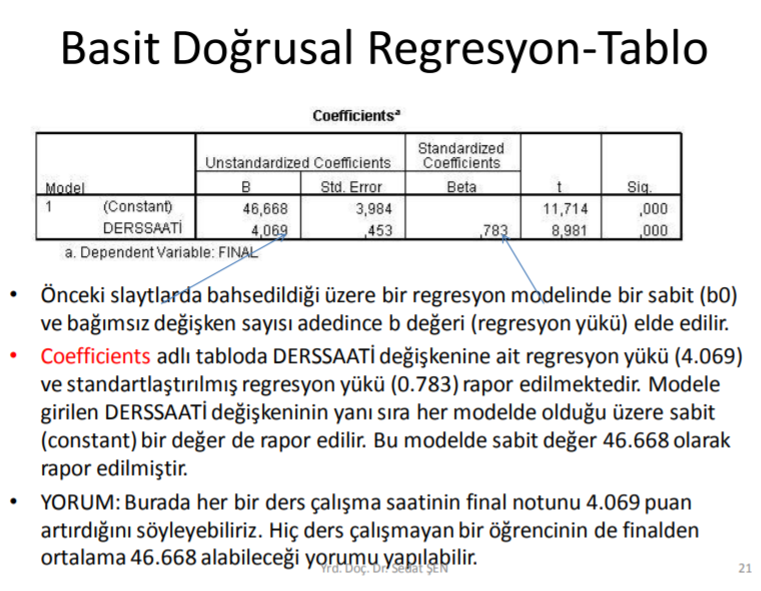
---------------------------------------------------------------------------------------------------



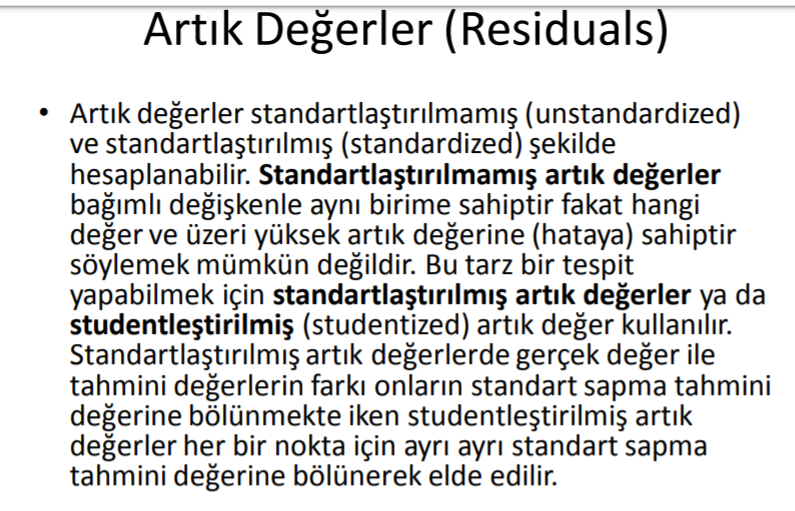
Model summary tablosunda R kare (R square) değerini görmekteyiz, bu değer korelasyon değerinin (R) karesidir. Bu değer bağımsız değişkenin bağımlı değişken üzerinde ne kadar etkili olduğu hakkında bilgi vermektedir. R kare değeri ne kadar büyük ise tutumun başarı üzerinde o kadar çok etkili olduğu yorumunu yapabiliriz.

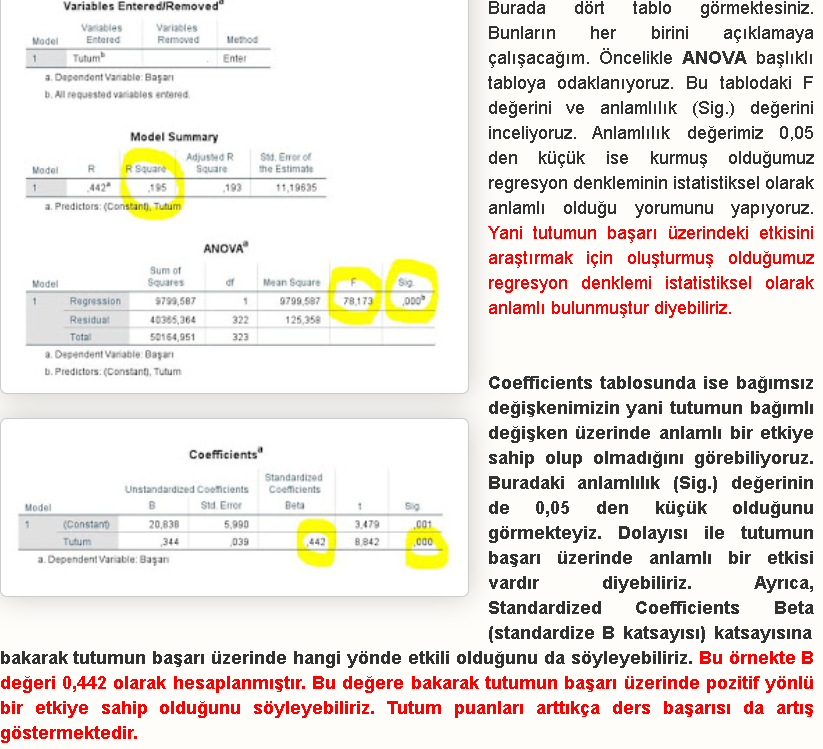




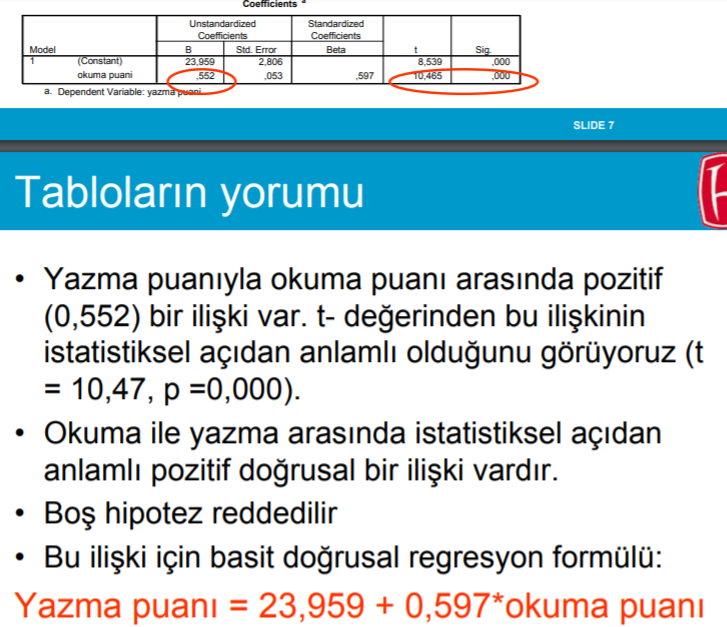




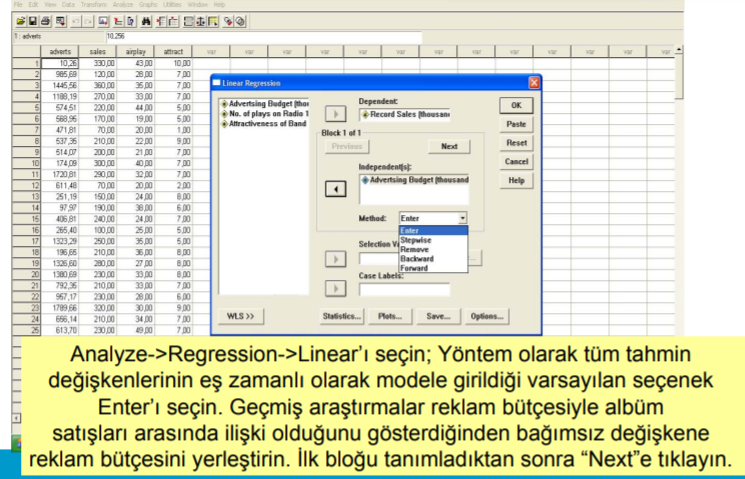




**REGRESYON ANALİZİNİN ÇIKTILARI**

****

**Çoklu Regresyon Analizi**

****