MODEL KURMA HATALARI ve VERİ SORUNLARI

Hüseyin Taştan[1]

[1]Yıldız Teknik Üniversitesi
İktisat Bölümü

Ders Kitabı:
Introductory Econometrics: A Modern Approach (2nd ed.)
J. Wooldridge

14 Ekim 2012

Model Spesifikasyonu ve Veri Sorunları

- Önceki derslerimizde Gauss-Markov varsayımlarından (MLR).5 Sabit varyans varsayımının sağlanmadığı durumu ele aldık.
- Değişen varyans sapma ve tutarsızlığa yol açmadığı için ciddi sorun oluşturmuyordu. Değişen varyansa dirençli standart hatalar kullanarak ya da WLS-FGLS tahmini yaparak geçerli t ve F testleri yapabiliyoruz.
- Şimdi ise daha ciddi bir problemi; "hata terimi u ile x'lerden birinin ya da birkaçının ilişkili olması" durumunu inceleyeceğiz.
- Hata terimiyle ilişkili olan x değişkenine **içsel (endogenous) değişken** denir. Açıktır ki bu durumda MLR.3 Sıfır Koşullu Ortalama varsayımı sağlanmaz.

Model Spesifikasyonu ve Veri Sorunları

- Gerekli bir değişkenin dışlanması durumunda OLS tahmincilerinin sapmalı ve tutarsız olduğunu daha önce görmüştük.
- Dışarıda bırakılan değişken x'lerden birinin bir fonksiyonu ise fonksiyonel biçim spesifikasyon hatası (functional form misspecification) ortaya çıkar.
- Bu bölümde ilk olarak fonksiyonel biçim hatası yapılıp yapılmadığını nasıl anlayacağımızı göreceğiz.
- Bu bölümde ihmal edilmiş değişkenin yol açtığı sapmayı azaltıcı yöntemlerden biri olan temsili değişken (proxy variable) yöntemini inceleyeceğiz.
- Yine bu bölümde ölçme hatalarının yol açtığı problemleri inceleyeceğiz.
- OLS tahmincilerinin sapmalı/tutarsız olduğu durumlarda başka tahmin yöntemlerine ihtiyaç duyulur. Araç Değişkenler, İki Aşamalı En Küçük Kareler yöntemleri gibi.

Fonksiyon Kalıbının Yanlış Kurulması

- Bir regresyonda y ile x'ler arasındaki ilişki doğru formüle edilmezse fonksiyonel biçim hatası (functional form misspecification) ortaya çıkar.
- Örneğin, log-log model yerine level-level model kullanılması, ya da olması gereken bir karesel terimin dışlanması fonksiyonel biçim hatasına, bu ise, $\hat{\beta}$'ların sapmalı ve tutarsız olmasına yol açacaktır.
- Örneğin, ilave bir yıl eğitimin ücrete katkısı cinsiyete göre değişiyorsa ücret regresyonunda $cinsiyet \times egitim$ karşılıklı etkileşim (interaction) terimini kullanmak zorundayız. Bunun nasıl yapılacağını daha önce görmüştük.
- Bu terimi dışlarsak fonksiyon biçimi hatası yapmış oluruz.

Fonksiyon Kalıbının Yanlış Kurulması

- Regresyona eklemek istediğimiz yeni değişken gruplarının (karesel terimler vb) gerekli olup olmadığına F testi (ortak anlamlılık testi) yaparak karar verebiliriz.
- Böylece regresyonumuzun fonksiyonel biçimini daha az hatasız hale getirebiliriz.
- Pek çok ekonomik seride log kullanılması düzey (level) değişken kullanılmasına göre daha iyi sonuç vermektedir. Log kullanarak biçim hatalarını azaltabiliriz.
- Yine, karesel terim eklemek de doğrusal-olmayan (nonlinear) ilişkilerin yakalanmasında önemli bir çözüm oluşturabilmektedir.

Fonksiyonel Biçim Hatasının Testi

- Fonksiyon kalıbının doğru kurmamıza yardımcı olabilecek bir test var mı?
- Aslında çok sayıda test var. Bunların hepsini burada incelemeyeceğiz.
- Bunların içinde en yaygın olarak kullanılan test Ramsey (1969) tarafından geliştirilen "Regression Specification Error Test" ya da kısaca RESET testidir.
- Ramsey RESET testi dışarıda bırakılmış modellenmeyen bir kısım olup olmadığını teşhis etmeye yönelik bir testtir.

RESET Testi

- Çoklu doğrusal regresyon modelinde

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

  MLR.3 x'lerin dışsallığı (sıfır koşullu ortalama) varsayımı sağlanıyor olsun.
- Bu durumda x'lerin kareleri, küpleri, vs. gibi doğrusal olmayan fonksiyonları bu denkleme eklendiğinde anlamsız çıkmalıdır.
- Tıpkı White değişen varyans testinin ilk versiyonunda olduğu gibi x'lerin çok olduğu bir modelde bunların karelerinin, küplerinin, çapraz çarpımlarının eklenmesi serbestlik derecesini önemli ölçüde azaltır.
- Bunun yerine fit edilen değerlerin, $\hat{y}$, karelerinin ve küplerinin, $\hat{y}^2$, $\hat{y}^3$, modele eklenerek F ya da LM testi yapılabilir.

RESET Testi

- RESET testinin yardımcı regresyonu aşağıdaki gibi yazılabilir:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + u$$

- RESET testinin boş hipotezi modelin doğru kurulduğunu söyler:

$$H_0 : \delta_1 = 0, \delta_2 = 0$$

- Büyük örneklemlerde ve Gauss-Markov varsayımları geçerlişken bu boş hipotez altında F kısıt test istatistiği $F(2, n-k-3)$ dağılımına uyar.
- Kritik değerden büyük bir F istatistiği bulunursa bu fonksiyon biçim hatasına işaret eder.
- LM versiyonu da kullanılabilir. LM test istatistiği χ^2_2 dağılımına uyar.

RESET Testi Örnek: Ev fiyatları, hprice1.gdt

- Ev fiyatları denklemi level-level modeli:

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u$$

- Model tahmin sonuçları:

$$\widehat{\text{price}} = \underset{(29.475)}{-21.77} + \underset{(0.0006)}{0.002}\ \text{lotsize} + \underset{(0.013)}{0.123}\ \text{sqrft} + \underset{(9.010)}{13.85}\ \text{bdrms}$$

$$n = 88 \quad R^2 = 0.672$$

- Bu modelden elde edilen $\hat{y}$ değerlerinin kare ve küplerini regresyona eklersek RESET testinin hesaplamakta kullanacağımız yardımcı regresyonu elde etmiş oluruz.

RESET Testi Örnek: Level-level ev fiyatları modeli

```
Auxiliary regression for RESET specification test
OLS, using observations 1-88
Dependent variable: price

              coefficient     std. error     t-ratio   p-value
  ------------------------------------------------------------
  const       166.097         317.433         0.5233    0.6022
  lotsize       0.000153723     0.00520304    0.02954   0.9765
  sqrft         0.0175988       0.299251      0.05881   0.9532
  bdrms         2.17490        33.8881        0.06418   0.9490
  yhat^2        0.000353426     0.00709894    0.04979   0.9604
  yhat^3        1.54557e-06     6.55431e-06   0.2358    0.8142

Test statistic: F = 4.668205,
with p-value = P(F(2,82) > 4.66821) = 0.012
```

GRETL: Model tahmin sonuçlarını gösteren pencere içinde:
TESTS-RAMSEY'S RESET-SQUARES and CUBES
RESET Testinin Sonucu: %5 anlamlılık düzeyinde modelin fonksiyon biçiminin doğru olduğunu söyleyen H_0 reddedilir. Fonksiyon kalıbı yanlıştır.

RESET Testi Örnek: Ev fiyatları, hprice1.gdt

- Ev fiyatları denklemi için başka bir fonksiyon kalıbı düşünelim: log-log modeli:

$$lprice = \beta_0 + \beta_1 llotsize + \beta_2 lsqrft + \beta_3 bdrms + u$$

- Model tahmin sonuçları:

$$\widehat{\text{lprice}} = \underset{(0.651)}{-1.297} + \underset{(0.038)}{0.17}\ \text{llotsize} + \underset{(0.093)}{0.70}\ \text{lsqrft} + \underset{(0.028)}{0.037}\ \text{bdrms}$$

$$n = 88 \quad R^2 = 0.643$$

- Şimdi RESET testini hesaplayalım.

RESET Testi Örnek: Level-level ev fiyatları modeli

```
Auxiliary regression for RESET specification test
OLS, using observations 1-88
Dependent variable: lprice

             coefficient   std. error   t-ratio   p-value
  ---------------------------------------------------------
  const       87.8849      240.974       0.3647   0.7163
  llotsize    -4.18098      12.5952     -0.3319   0.7408
  lsqrft     -17.3491       52.4899     -0.3305   0.7418
  bdrms       -0.925329      2.76975    -0.3341   0.7392
  yhat^2       3.91024      13.0143      0.3005   0.7646
  yhat^3      -0.192763      0.752080   -0.2563   0.7984

Test statistic: F = 2.565042,
with p-value = P(F(2,82) > 2.56504) = 0.0831
```

GRETL: Model tahmin sonuçlarını gösteren pencere içinde:
TESTS-RAMSEY'S RESET-SQUARES and CUBES
RESET Testinin Sonucu: %5 anlamlılık düzeyinde modelin fonksiyon biçiminin doğru olduğunu söyleyen H_0 reddedilemez. Fonksiyon kalıbı doğrudur. Log-log modeli tercih edilmelidir.

RESET Testi

- RESET testinin bir yetersizliği, H_0'ın reddi halinde ne yapacağımız konusunda bize hiçbir şey söylememesidir.
- Bazıları RESET testinin ihmal edilmiş değişken ve değişen varyanstan ileri gelen biçim hatalarını (misspecification) da yakaladığını, dolayısıyla çok genel bir "model kurma hatası testi" olduğunu iddia ederler.
- Bu doğru değildir. İhmal edilmiş değişkenin y ile ilişkisi doğrusal ise RESET testi bunu yakalayamaz.
- Yine, fonksiyonel biçim doğru kurulmuşsa RESET testi değişen varyansı yakalamakta başarısızdır.
- RESET testi sadece bir fonksiyonel biçim testidir, genel bir model kurma hatası testi değildir.

Yuvalanmamış Alternatiflere Karşı Test

- Çok çeşitli fonksiyon kalıbı hatası testleri vardır. Örneğin, açıklayıcı değişkenlerin level ya da log olarak modele dahil etmeye nasıl karar veririz. Şu iki model arasında karar vermek istediğimizi düşünelim:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + u$$

- Bu modeller birbirlerinin özel hali olarak yazılamazlar, yani yuvalanmamış (nonnested) modellerdir.
- Bu modellerin seçiminde standart F testini kullanamayız.
- Bağımlı değişken aynı olduğu sürece iki modeli de kapsayan büyük bir model kurularak F testi yapılabilir. Bu yöntem Mizon-Richard yöntemi olarak bilinir.

Yuvalanmamış Alternatiflere Karşı Test

- Bir başka yöntem ise literatürde Davidson-MacKinnon testi olarak bilinir. Modellerin tahmininden elde edilen $\hat{y}$ değerlerinin diğer modelde açıklayıcı değişken olarak kullanılmasına dayanır.
- Bu testlerin daha fazla ayrıntısına girmeyeceğiz.
- Yuvalanmamış modellerin testiyle ilgili birçok problem bulunmaktadır.
- Bu testler alternatiflerin hangisinin doğru olduğuna her zaman karar veremeyebilir. Testin sonucuna göre her iki model yanlış ya da doğru çıkabilir.
- Alternatiflerden birinin reddi diğerinin doğru olduğu anlamına gelmez. Doğru model başka bir spesifikasyona sahip olabilir.
- Bağımlı değişken farklıysa, örneğin birinde y diğerinde $\log(y)$ ise bu testler uygulanamaz. Bu durum için geliştirilmiş daha karmaşık testler vardır.

Gözlenemeyen Açıklayıcı Değişkenler Yerine Temsili (Proxy) Değişken Kullanılması

- Ölçülemeyen, gözlenemeyen ya da veri bulunamayan ve bu nedenle de model dışında bırakılmış bir değişken yerine temsili bir değişken kullanabilir miyiz?
- Eğer gözleyemediğimiz değişken modelde yer alması gereken önemli bir değişkense bu değişkeni ihmal etmek OLS tahmincilerinin sapmalı ve tutarsız olmasına yol açar.
- Öyleyse soruyu şöyle de sorabiliriz: İhmal edilmiş bir değişkenin yol açtığı sapmayı temsili bir değişken yoluyla azaltabilir miyiz?
- Temsili değişken (proxy variable): ihmal edilmiş değişken ile ilişkili olan, onun yerine kullanabileceğimiz bir değişken.
- Örnek: ücret denkleminde doğuştan gelen yetenek gözlenemiyordu. Acaba IQ test puanı doğuştan gelen yetenek için bir temsili değişken olabilir mi?
- IQ ile doğuştan gelen yeteneğin aynı şey olmadığını biliyoruz. Ancak aynı şeyi ölçmeleri gerekmez, sadece ilişkili olmaları yeterlidir.

Temsili (Proxy) Değişken Kullanılması

- Modelimiz şu olsun:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$$

  y :log(wage), x_1:educ, x_2: exper, x_3^*:ability (unobserved)
- x_3^*: gözlenemeyen değişken; x_3: temsili değişken
- Temsili değişkenin gözlenemeyen değişken ile ilişkisi:

$$x_3^* = \delta_0 + \delta_3 x_3 + \nu_3$$

- Bu değişkenler tam olarak ilişkili olmadıklarından ν_3 hata terimi yer almaktadır.
- Bu değişkenler pozitif ilişkili olacaklarından $\delta_3 > 0$ olmasını bekleriz.
- $\delta_3 = 0$ ise x_3 uygun bir proxy olamaz.

Temsili (Proxy) Değişken Kullanılması

- x_3^* yerine x_3'ün modele konarak tahmin edilmesine "ihmal edilmiş değişkenlerin yerine koyma yöntemiyle çözümü" denir (plug-in solution to the omitted variables problem)
- Bu durumda OLS en azından tutarlı tahminciler verir mi?
- Bunu ortaya koyabilmek için u ve ν_3 hakkında bazı varsayımlar yapmamız gerekir.
- Hata terimi, u, açıklayıcı değişkenler x_1, x_2 ve x_3^* ile ilişkisiz olmalı. Bu zaten standart MLR.3 varsayımı.
- Buna ek olarak u, x_3 ile de ilişkisiz olmalı. x_3 proxy değişken olduğundan popülasyon modeli için gereksiz bir değişkendir. y'yi doğrudan etkileyen x_3^*'dır, x_3 değil.

$$\mathsf{E}(u|x_1, x_2, x_3^*, x_3) = \mathsf{E}(u|x_1, x_2, x_3^*) = 0$$

Temsili (Proxy) Değişken Kullanılması

- ν_3 hata terimi, açıklayıcı değişkenler x_1, x_2 ve x_3 ile ilişkisiz olmalı.
- Bu varsayımı şöyle ifade edebiliriz:

$$\mathsf{E}(x_3^*|x_1, x_2, x_3) = \mathsf{E}(x_3^*|x_3) = \delta_0 + \delta_3 x_3$$

- Buna göre x_3 kontrol edildikten sonra x_3^*'ın koşullu beklenen değeri x_1 ve x_2'ye bağlı değildir. Bu x_3'ün iyi bir proxy olması için şarttır.
- Örneğin ücret denkleminde IQ puanının ability için iyi bir temsili değişken olması için

$$\mathsf{E}(ability|educ, exper, IQ) = \mathsf{E}(ability|IQ) = \delta_0 + \delta_3 IQ$$

  olmalıdır.
- Yani ortalama doğuştan gelen yetenek düzeyi sadece IQ ile değişmeli, eğitim ve tecrübe değişkenleri ile ilişkisiz olmalıdır.
- Bunun aşağı yukarı doğru olduğu söylenebilir.

Temsili (Proxy) Değişken Kullanılması

- $x_3^* = \delta_0 + \delta_3 x_3 + \nu_3$ modelde yerine yazılıp yeniden düzenlenirse

$$y = (\beta_0 + \beta_3\delta_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3\delta_3 x_3 + u + \beta_3\nu_3$$

- Bileşik hata terimine $e = u + \beta_3\nu_3$ dersek

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha_3 x_3 + e$$

  burada, $\alpha_0 = (\beta_0 + \beta_3\delta_0), \alpha_3 = \beta_3\delta_3$
- Açıktır ki yukarıda incelediğimiz varsayımlar sağlanıyorsa yeni hata terimi e modelde yer alan değişkenlerle ilişkisiz olacaktır. Bu durumda $\alpha_0, \beta_1, \beta_2, \alpha_3$ katsayıları tutarlı bir şekilde tahmin edilebilir.
- α_3 katsayısı IQ testindeki bir puanlık artışın ücretler üzerindeki etkisini gösterecektir.

Temsili (Proxy) Değişken Kullanılması Örnek: Wage2.gdt

- Bu veri setinde 935 çalışana ilişkin ücret, eğitim, tecrübe, kıdem ve IQ test puanları ile çeşitli demografik özelliklerine (evli olup olmadığı, yaşadığı yer, etnik köken) ilişkin bilgiler yer almaktadır.
- IQ test skorlarının modele eklenmesiyle aşağıdaki sonuçlara ulaşılmıştır:

Model 1: OLS, using observations 1–935
Dependent variable: lwage

	Coefficient	Std. Error	t-ratio	p-value
const	5.17644	0.128001	40.4407	0.0000
educ	0.0544106	0.00692849	7.8532	0.0000
exper	0.0141458	0.00316510	4.4693	0.0000
tenure	0.0113951	0.00243938	4.6713	0.0000
married	0.199764	0.0388025	5.1482	0.0000
south	−0.0801695	0.0262529	−3.0537	0.0023
urban	0.181946	0.0267929	6.7908	0.0000
black	−0.143125	0.0394925	−3.6241	0.0003
IQ	0.00355910	0.000991808	3.5885	0.0004

Mean dependent var	6.779004	S.D. dependent var	0.421144
Sum squared resid	122.1203	S.E. of regression	0.363152
R^2	0.262809	Adjusted R^2	0.256441

Ücret Denklemi, bağımlı değişken: log(wage)

Independent Variables	(1)	(2)	(3)
educ	.065 (.006)	.054 (.007)	.018 (.041)
exper	.014 (.003)	.014 (.003)	.014 (.003)
tenure	.012 (.002)	.011 (.002)	.011 (.002)
married	.199 (.039)	.200 (.039)	.201 (.039)
south	−.091 (.026)	−.080 (.026)	−.080 (.026)
urban	.184 (.027)	.182 (.027)	.184 (.027)
black	−.188 (.038)	−.143 (.039)	−.147 (.040)
IQ	———	.0036 (.0010)	−.0009 (.0052)
educ·IQ	———	———	.00034 (.00038)
intercept	5.395 (.113)	5.176 (.128)	5.648 (.546)
Observations *R*-Squared	935 .253	935 .263	935 .263

Bağımlı Değişkenin Gecikmesinin Proxy Olarak Kullanımı

- Ücret denkleminde olduğu gibi bazı durumlarda hangi gözlenemeyen faktörü kontrol etmek istediğimiz hakkında bir fikrimiz vardır.
- Bazı durumlarda ise ihmal edilmiş değişken ya da değişkenlerin modeldeki değişkenlerle ilişkili olduğunu biliriz ancak bu ihmal edilmiş değişken için proxy bulmakta zorlanırız.
- Bu gibi durumlarda bağımlı değişken y'nin önceki zaman periyotlarındaki değerini, y_{-1}, proxy olarak kullanabiliriz.
- Kesit veri analizinde bağımlı değişkenin gecikmesini kullanabilmek için bu değişkenin geçmişte aldığı değerleri bilmemiz gerekir. Bu değişkeni modele ekleyerek bağımlı değişkendeki yavaş hareket eden kısmı (inertia) kontrol etmiş oluruz.
- Örneğin şehirlerdeki suç oranını açıklayan bir modelde işsizlik ve güvenlik harcamalarının yanı sıra geçmişteki suç oranını da modele ekleyerek suç oranındaki yavaş değişimleri kontrol etmiş oluruz.

Bağımlı Değişkenin Gecikmesinin Proxy Olarak Kullanımı

- Örnek: CRIME2.gdt, 46 şehir için 1987 yılı verileri, ayrıca suç oranı için 1982 verisi de mevcut
- Suç oranının gecikmesini içermeyen model:

$$\widehat{\text{l\_crmrte87}} = \underset{(1.251)}{3.34} - \underset{(0.032)}{0.029}\,\text{unem87} + \underset{(0.173)}{0.203}\,\text{l\_lawexpc87}$$

$$n = 46 \quad R^2 = 0.057$$

- Suç oranının gecikmesini içeren model:

$$\widehat{\text{l\_crmrte87}} = \underset{(0.821)}{0.076} + \underset{(0.02)}{0.009}\,\text{unem87} - \underset{(0.109)}{0.140}\,\text{l\_lawexpc87} + \underset{(0.132)}{1.194}\,\text{l\_crmrte82}$$

$$n = 46 \quad R^2 = 0.680$$

- İlk modelde işsizlik arttıkça suç oranı azalıyor. Bu teorik beklentilerle ve sezgilerle uyumlu değil.
- 5 yıl önceki suç oranını kontrol ettikten sonra unem katsayısı pozitif ancak anlamlı değil.
- Cari dönem suç oranının bir önceki dönem suç oranına göre esnekliği kaçtır?

Ölçme Hataları

- Bazı uygulamalarda iktisadi davranışı etkileyen değişkenin gerçek değeri tam olarak gözlenemeyebilir ya da ölçülemeyebilir.
- Değişkenin gerçek değeri gözlenemiyorsa ölçme hataları oluşur.
- Örneğin, hanehalklarının beyan ettiği gelir ya da tüketim düzeyi gerçek değerlerden farklı olabilir.
- Bu gibi durumlarda OLS tahmincilerinin özelliklerini inceleyeceğiz.
- Ölçme hataları iki kısımda incelenebilir: (1) Bağımlı değişkendeki ölçme hataları ve (2) Açıklayıcı değişkenlerdeki ölçme hataları
- Bu ölçme hatalarının hangi durumlarda OLS tahmincilerinin tutarsızlığına neden olduğunu inceleyeceğiz.

Bağımlı Değişkendeki Ölçme Hataları

- y^* açıklamaya çalıştığımız bağımlı değişkenin (gözlenemeyen) gerçek değeri olsun. Örneğin hanehalkı tasarruf düzeyi.

$$y^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

- y gözlenen (ya da beyan edilen) değer olsun. Bu iki değer arasındaki fark popülasyondaki ölçme hatası olarak tanımlanır:

$$e_0 = y - y^*$$

- Buradan $y^* = y - e_0$ yazılabilir. Bunu modelde yerine yazarsak tahmin edilebilir bir modele ulaşırız:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u + e_0$$

- Burada bileşik hata teriminin $u + e_0$ olduğuna dikkat ediniz. Ölçme hatası artık regresyon hata teriminin içinde yer almaktadır. Hangi koşul altında OLS tahmincileri tutarlı olur?

Bağımlı Değişkendeki Ölçme Hataları

- OLS ile tahmin edilebilir model şu şekildeydi:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \underbrace{u + e_0}$$

- Eğer ölçme hatası, e_0, her bir x_j ile ilişkisiz ile tutarlı tahmin mümkün olur. Ölçme hatası açıklayıcı değişkenlerden bağımsız ise OLS hem sapmasız hem de tutarlıdır.
- Hata terimi, u ile ölçme hatası e_0 bağımsız ise (genellikle böyle varsayılır):

$$\text{Var}(u + e_0) = \text{Var}(u) + \text{Var}(e_0) = \sigma_u^2 + \sigma_0^2 > \sigma_u^2$$

- Öyleyse bağımlı değişkende ölçme hatası olması durumunda regresyonun varyansı, olmadığı duruma göre daha yüksek olacaktır.
- Sonuç olarak OLS tahmincilerinin varyansı, ve tabii ki standart hataları daha yüksek çıkacaktır. Bu durumda daha fazla "kaliteli" verinin toplanması gerekebilir.

Bağımlı Değişkendeki Ölçme Hataları: Örnek

- Aşağıdaki tasarruf fonksiyonunu düşünelim:

  $tasarruf^* = \beta_0 + \beta_1 gelir + \beta_2 buyukluk + \beta_3 egitim + \beta_4 yas + u$

  $tasarruf^*$: gerçek hanehalkı tasarruf düzeyi, $tasarruf$: beyan edilen (verilerimizdeki) tasarruf düzeyi, $gelir$: yıllık gelir düzeyi, $buyukluk$: hanehalkındaki kişi sayısı, $egitim$: hanehalkı reisinin eğitim düzeyi, yas: hanehalkı reisinin yaşı.
- Bu modelde ölçme hatası ($tasarruf - tasarruf^*$) hangi durumda problem oluşturur?
- Bu ölçme hatasının gelir, büyüklük, eğitim ve yaş ile ilişkisiz olduğu düşünülebilir.
- Ancak yüksek gelir düzeyine sahip aileler daha düşük tasarruf düzeyi beyan etme eğiliminde iseler OLS tahmincileri tutarlı olmayabilir. Benzer şekilde, yüksek eğitimli ailelerin tasarruf düzeylerini daha gerçeğe yakın beyan edecekleri söylenebilir.
- Ölçme hatasını gözleyemediğimiz için gelir ya da eğitimle ilişkili olup olmadığını hiç bir zaman belirleyemeyiz.

Açıklayıcı Değişkenlerde Ölçme Hataları

- x'lerdeki ölçme hataları y'deki ölçme hatasına göre daha ciddi problemlere neden olabilmektedir.
- Hangi durumlarda ölçme hatalarının OLS'nin tutarsızlığına yol açtığını görebilmek için basit regresyon modelini düşünelim:

$$y = \beta_0 + \beta_1 x_1^* + u$$

İlk dört Gauss-Markov varsayımı sağlanıyor.

- Burada x_1^* gerçek değeri ifade etmektedir. Gözlenen değeri x_1 ile göstereceğiz.
- Bu durumda ölçme hatası aşağıdaki gibi tanımlanır:

$$e_1 = x_1 - x_1^*$$

- Bu ölçme hatasının koşulsuz beklenen değeri sıfırdır: $\mathsf{E}(e_1) = 0$

Açıklayıcı Değişkenlerde Ölçme Hataları

- Hata terimi u'nun hem x_1^* hem de x_1 ile ilişkisiz olduğunu düşünelim. Bunu şöyle ifade edebiliriz:

$$\mathsf{E}(y|x_1^*, x_1) = \mathsf{E}(y|x_1^*)$$

- Bu koşullu beklenti şöyle yorumlanabilir: x_1^* kontrol edildikten sonra artık x_1'e gerek yoktur.
- x_1^* yerine x_1 kullanırsak OLS tahmincilerinin özellikleri nelerdir? Tutarlı olurlar mı?
- Bu özellikler ölçme hatalarına ilişkin yapacağımız varsayıma bağlı.
- Ekonometri literatüründe iki varsayım yapılmaktadır. (1) ölçme hatası gözlenen x_1 değişkeni ile ilişkisizdir.
- (2) ölçme hatası gözlenemeyen gerçek değer, x_1^*, ile ilişkisizdir.

(1) e_1 ile x_1 ilişkisizdir

- Bu varsayım şu şekilde ifade edilebilir:

$$\mathsf{Cov}(x_1, e_1) = 0$$

- Bu durumda, $e_1 = x_1 - x_1^*$ olduğundan, ölçme hatası e_1, x_1^* ile ilişkili olmak zorundadır.
- Bu varsayım altında, $x_1^* = x_1 - e_1$, modelde yerine yazılırsa:

$$y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1)$$

- Bileşik hata teriminin beklenen değer ve varyansı

$$\mathsf{E}(u - \beta_1 e_1) = 0, \quad \mathsf{Var}(u - \beta_1 e_1) = \sigma_u^2 + \beta_1^2 \sigma_{e_1}^2$$

- Yukarıdaki modelde hata terimi ile x_1 varsayım gereği ilişkisiz olduğundan OLS tahmincileri tutarlı olur. Ancak tahmin varyansı yükselir.

(2) e_1 ile x_1^* ilişkisizdir (CEV varsayımı)

- Bu varsayıma "Klasik Değişkenlerde Hata" varsayımı denir (Classical Errors in Variable - CEV). Açıklayıcı değişkendeki ölçme hataları denilince bu varsayım akla gelir.
- Bu varsayım şu şekilde ifade edilebilir:

$$\mathsf{Cov}(x_1^*, e_1) = 0$$

- Gözlenen değeri gerçek değer ve ölçme hatasının toplamı olarak yazabiliriz:

$$x_1 = x_1^* + e_1$$

- Açıktır ki eğer x_1^* ile e_1 ilişkisiz ise x_1 ile e_1 ilişkili olmak zorundadır:

$$\mathsf{Cov}(x_1, e_1) = \mathsf{E}(x_1 e_1) = \mathsf{E}(x_1^* e_1) + \mathsf{E}(e_1^2) = 0 + \sigma_{e_1}^2 = \sigma_{e_1}^2$$

- CEV varsayımı altında x_1 ile e_1 arasındaki kovaryans ölçme hatasının varyansına eşittir.

(2) $\text{Cov}(x_1^*, e_1) = 0$: e_1 ile x_1^* ilişkisizdir (CEV varsayımı)

- Hatırlarsak modeli şu şekilde yazmıştık:

$$y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1)$$

- e_1 bileşik hata teriminde yer aldığına göre bunun x_1 ile ilişkili olması problem yaratacaktır.
- Bileşik hata terimi ile x_1 arasındaki kovaryansı hesaplarsak:

$$\text{Cov}(x_1, u - \beta_1 e_1) = -\beta_1 \text{Cov}(x_1, e_1) = -\beta_1 \sigma_{e_1}^2$$

- Bu kovaryans 0 olmadığına göre, CEV varsayımı altında OLS tahmincileri sapmalı ve tutarsız olacaktır.
- Tutarsızlığın boyutunu hesaplayabiliriz.

(2) $\text{Cov}(x_1^*, e_1) = 0$: e_1 ile x_1^* ilişkisizdir (CEV varsayımı)

- OLS tahmincisinin tutarsızlığının boyutunu ölçebiliriz. Basit regresyon modelinde eğim katsayısının OLS tahmincisinin formülünden hareketle olasılık limitini aşağıdaki gibi yazabiliriz:

$$\begin{aligned}
\text{plim}(\hat{\beta}_1) &= \beta_1 + \frac{\text{Cov}(x_1, u - \beta_1 e_1)}{\text{Var}(x_1)} \\
&= \beta_1 - \frac{\beta_1 \sigma_{e_1}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2} \\
&= \beta_1 \left(1 - \frac{\sigma_{e_1}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2} \right) \\
&= \beta_1 \left(\frac{\sigma_{x_1^*}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2} \right)
\end{aligned}$$

(2) $\text{Cov}(x_1^*, e_1) = 0$: e_1 ile x_1^* ilişkisizdir (CEV varsayımı)

- OLS tahmincisinin olasılık limiti:

$$\text{plim}(\hat{\beta}_1) = \beta_1 \underbrace{\left(\frac{\sigma_{x_1^*}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2} \right)}_{\leq 1} \neq \beta_1$$

- Parantez içindeki ifade her zaman 1'den küçük çıkacaktır. Ancak ve ancak $\sigma_{e_1}^2 = 0$ olduğunda 1 olur.
- Bunun anlamı şudur: $\hat{\beta}_1$ her zaman 0'a doğru değer β_1'den daha yakındır. Buna küçültme sapması (attenuation bias) adı verilir.
- $\beta_1 > 0$ ise $\hat{\beta}_1$ limitte bu değerden daha küçük bir sayıya yaklaşacaktır (underestimation). Aksi durumda ise daha büyük bir sayıya yaklaşacaktır (overestimation).

(2) $\text{Cov}(x_1^*, e_1) = 0$: e_1 ile x_1^* ilişkisizdir (CEV varsayımı)

- OLS tahmincisinin olasılık limiti:

$$\text{plim}(\hat{\beta}_1) = \beta_1 \underbrace{\left(\frac{\sigma_{x_1^*}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2} \right)}_{\leq 1} \neq \beta_1$$

- Eğer x_1^*'ın varyansı e_1'in varyansına kıyasla büyükse $\text{Var}(x_1^*)/\text{Var}(x_1)$ oranı 1'e yakın çıkacağı için OLS tutarsızlığı çok büyük olmayabilir. Ancak bunu bilmek çoğunlukla mümkün değildir.
- Şimdiye kadar tek açıklayıcı değişken modeli çerçevesinde CEV varsayımının sonuçlarını gördük. Çok değişkenli modelde CEV ölçme hatası varsa durum çok daha karmaşık hale gelmektedir.
- Genel olarak bir değişkendeki ölçme hatası tüm değişkenlerin katsayılarının OLS tahmincilerinin tutarsız olmasına yol açmaktadır.

(2) Cov$(x_1^*, e_1) = 0$: e_1 ile x_1^* ilişkisizdir (CEV varsayımı)

- Örnek olarak aşağıdaki üniversite başarı modelini düşünelim:

$$colGPA = \beta_0 + \beta_1 faminc^* + \beta_2 hsGPA + \beta_3 SAT + u$$

$faminc$: aile geliri, $hsGPA$: lise not ortalaması, SAT: üniversite giriş sınav notu

- $faminc^*$ gerçek aile geliridir. Bu modeli tahmin etmek için bir anket çalışması yapılıyorsa muhtemelen öğrenciden aile gelirini beyan etmesi istenecektir.
- Öğrenci kayıtlarından hareketle hsGPA ve SAT değerleri alınabilir. Ancak aile geliri için bunu yapamayız.
- Beyan edilen gelir gerçek gelirden farklıysa ve CEV varsayımları geçerliyse, yani gerçek gelir ile ölçme hatası ilişkisiz ise, β_1 sapmalı ve tutarsız tahmin edilecektir.
- Ailenin gelirinin başarı üzerindeki etkisi olduğundan çok daha düşük düzeyde tahmin edilecektir.

Veri Problemleri

- Ölçme hataları bir veri problemi olarak düşünülebilir. Sonuçta verilerin gerçek değerlerini gözleyemiyoruz.
- Daha önce karşılaştığımız diğer bir veri problemi yüksek (tam olmayan) çoklu doğrusallık (multicollinearity) idi. Çoklu doğrusallık durumunda x değişkenleri yüksek dereceden doğrusal ilişkiliydi. Fakat bu sapma ya da tutarsızlığı yol açmıyordu, sadece regresyonun ve katsayı tahminlerinin varyansını yükseltiyordu.
- Bunların dışında verilerde şu problemler ortaya çıkabilir:
- Kayıp ya da eksik gözlemler (missing data)
- Rassal olmayan örneklemler (nonrandom samples)
- Uç gözlemler (outliers)

Kayıp/Eksik Gözlemler (Missing Data)

- Kayıp/eksik gözlemler ya da verilerde boşluklar çok çeşitli şekillerde ortaya çıkabilir. En yaygın biçimi, bazı deneklerin anket sorularının bir kısmını yanıtlamaması halidir.
- Bu gözlemler analizde kullanılamazlar. Ekonometri paket programları boşluğa denk gelen gözlemleri otomatik olarak dışlamaktadır. Dolayısıyla, veri boşluğu örnek hacmini küçültmektedir.
- Veri boşluğunun daha ciddi istatistiki sorunlara yol açıp açmayacağı boşluğun nedeni ile ilgilidir. Eğer boşluklar rasgele oluşmuşsa, bu, örnek hacmini küçültmenin dışında sapma ve tutarsızlık sorunları doğurmaz, MLR.2 Rassal Örneklem varsayımı hala geçerlidir. Boşluklar, rasgele değil de sistematik ise sorun ciddidir.
- Veride boşluklar rasgele-olmayan örneklemede daha ciddi sorun yaratır. Örneğin, doğumda bebek ağırlıkları veri setinde EDUC değişkenindeki boşluklar eğitim düzeyi ortalamanın altında olan anne-babalarda daha yaygın ise, bu sistematik bir olaydır.

Rassal Olmayan Örneklemler (Nonrandom Sampling)

- Verilerde boşluklar sistematik olarak oluşuyorsa bu rassal olmayan bir örneklem ortaya çıkmasına neden olabilir.
- Örneğin, ücret denklemine IQ test sonuçlarını eklemek istediğimizi düşünelim. Bu regresyonu tahmin ederken IQ puanı olmayan çalışanları dışlamak zorunda kalırız.
- Eğer yüksek IQ düzeyine sahip olanlar için IQ test sonuçlarını elde etmek daha kolaysa ortaya çıkan örneklem anakütleyi temsil edemez.
- Bu durumda MLR.2 Rassal Örnekleme varsayımı sağlanamayacağı için OLS tahmin sonuçları yanıltıcı olabilir.
- Bazı tür rassal olmayan örnekleme biçimleri ise sapma ve tutarsızlığa yol açmaz.

Rassal Olmayan Örneklemler (Nonrandom Sampling)

- Eğer örneklem açıklayıcı değişkenlere göre seçiliyorsa bu istatistiksel problemlere yol açmaz.
- Buna **dışsal örneklem seçimi** (exogenous sample selection) denir.
- Örneğin aşağıdaki tasarruf modelini düşünelim:

$$saving = \beta_0 + \beta_1 income + \beta_2 age + \beta_3 size + u$$

- Eğer veri setimiz 35 yaşın üzerinde kişilerle yapılan bir ankete dayanıyorsa rassal olmayan bir örneklem elde ederiz.
- Diğer varsayımlar sağlanıyorsa OLS hala sapmasız ve tutarlıdır. Bunun nedeni gelir, yaş ve büyüklük değişkenleri kontrol edildiğinde ortalama tasarruflar anakütlenin her kesiminde aynıdır.

$$\mathrm{E}(saving|income, age, size)$$

Rassal Olmayan Örneklemler (Nonrandom Sampling)

- Eğer örneklem açıklayıcı değişkenlere göre değil de bağımlı değişken y'ye göre seçiliyorsa bu problem yaratacaktır.
- Buna **içsel örneklem seçimi** (endogenous sample selection) denir.
- Örneğin aşağıdaki servet regresyonunu düşünelim:

$$wealth = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 age + u$$

- Eğer örnekleme sadece serveti 250,000$'dan az olan bireyleri seçersek içsel örneklem seçimi yapmış oluruz.
- Bu durumda tüm OLS tahmincileri sapmalı ve tutarsız olur. Bunun nedeni şudur: Serveti 250,000$'dan az olanların koşullu beklenen değeri

$$\mathrm{E}(wealth|educ, exper, age)$$

bu değerden yüksek servete sahip olanlar için koşullu beklenen değerle aynı değildir.

Uç Gözlemler (Outliers - Extreme Observations)

- Çok yüksek ya da çok düşük değerler alan gözlemlere uç gözlemler denir.
- Özellikle küçük örneklemlerde uç değerler OLS sonuçlarını çok fazla etkileyebilir.
- Bir gözlem değerinin uç değer olduğunu nasıl anlarız?
- Eğer o gözlemi veri setinden çıkardığımız regresyon sonuçları önemli ölçüde değişiyorsa o gözlem uç gözlemdir.
- OLS kalıntı kareleri toplamını (SSR) minimize ettiği için mutlak olarak büyük kalıntılar (eksi ya da artı) kareleri alındıklarında daha da büyümekte ve tahmine egemen olmaktadırlar.
- Başka bir ifadeyle, uç değerler örnekte çok büyük ağırlık almaktadır.

Uç Gözlemler (Outliers - Extreme Observations)

- Uç değerler maddi bir hatadan ya da populasyonun dağılımından kaynaklanır.
- Verileri analize hazırlarken bir değerin yanına yanlışlıkla fazladan bir sıfır koyarsak maddi bir hata yapmış oluruz. Bu hata OLS sonuçlarını önemli ölçüde etkileyecektir.
- Bu nedenle regresyon analizine geçmeden önce değişkenlerin özet istatistiklerini, ortalama, mod, medyan, minimum, maksimum, vs., incelemek faydalı olacaktır.
- Eğer uç değerler değişkenin dağılımının bir özelliği ise ne yapılacağı çok açık değildir.
- Genellikle böyle bir durumda model uç değeri içeren ve içermeyen verilerle iki kere tahmin edilerek sonuçlar sunulur.

Uç Gözlemler: Örnek

- Araştırma-Geliştirme (R&D) harcamaları ve firma performansı:

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 profmarg + u$$

$rdintens$: AR-GE harcamalarının satışlara oranı; $sales$: satışlar (milyon \$); $profmarg$: kar/satışlar, %

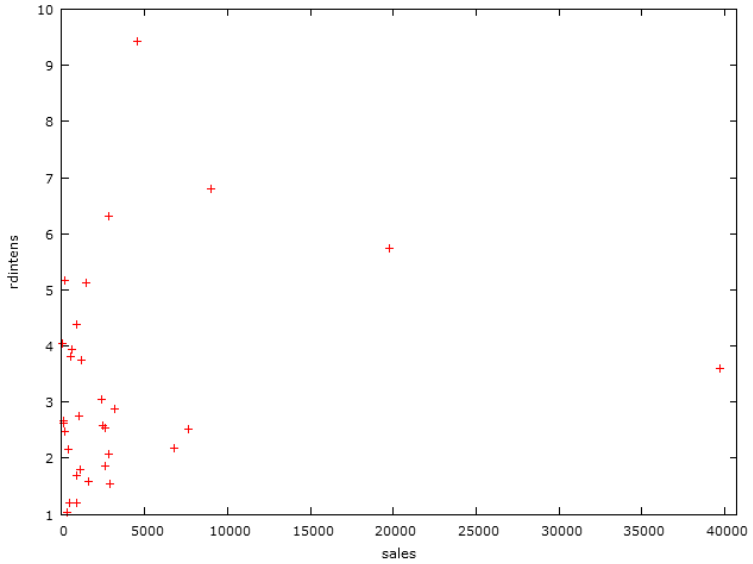
- Veri seti: RDCHEM.gdt, tahmin sonuçları

$$\widehat{\text{rdintens}} = \underset{(0.585)}{2.62} + \underset{(0.00004)}{0.00005}\,\text{sales} + \underset{(0.046)}{0.045}\,\text{profmarg}$$

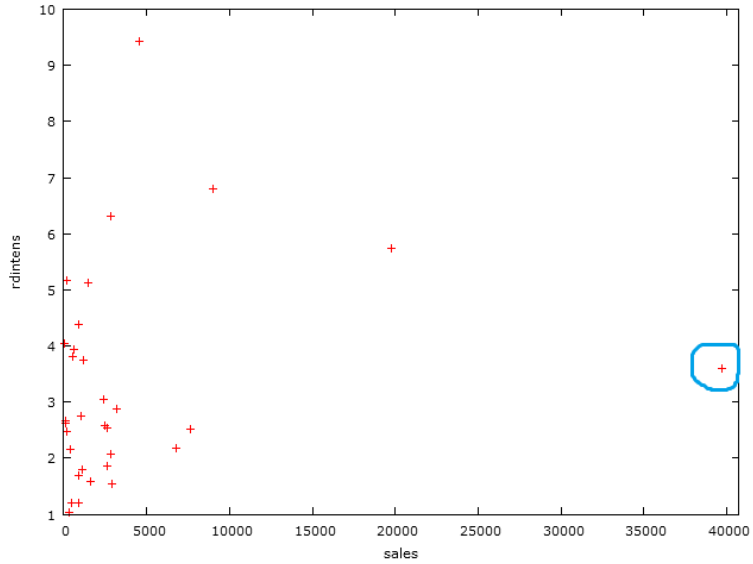
$$n = 32 \quad R^2 = 0.076$$

- Satışlar ve kar marjı değişkenlerinin her ikisi de %10 düzeyinde bile anlamsız.
- Acaba uç değerler var mı? Bunu görmenin en kolay yolu ikili serpilme çizimlerini ve özet istatistikleri incelemektir.

Uç Değerler: Örnek



Uç Değerler: Örnek



Uç Değerler: Örnek

- 32 firmadan 31'inin yıllık satışları 20 milyar dolardan daha azdır. 1 firmanın satışları ise 40 milyar dolara yakındır.
- Bu gözlem bir uç değer olabilir. Bu değeri dışlayarak regresyonu yeniden tahmin edersek:

$$\widehat{\text{rdintens}} = \underset{(0.592)}{2.297} + \underset{(0.000084)}{0.000186}\,\text{sales} + \underset{(0.0445)}{0.0478}\,\text{profmarg}$$

$$n = 31 \quad R^2 = 0.1728$$

- Satışların katsayısı yaklaşık 3 kat arttı. Ayrıca t istatistiği anlamlı.
- Firma büyüklüğü ile AR-GE yoğunluğu arasında anlamlı bir ilişki mevcut.
- Kar marjının katsayısı ise fazla değişmedi ve anlamsız.

Uç Değerler

- Doğal logaritma dönüştürmesi uç değer sorununu hafifletebilir.

$$\log(rd) = \beta_0 + \beta_1 \log(sales) + \beta_2 profmarg + u$$

rd: AR-GE harcamaları, milyon \$

- $n = 32$ uç değer dahil model sonuçları:

$$\widehat{\log(\text{rd})} = \underset{(0.468)}{-4.378} + \underset{(0.060)}{1.084}\,\log(\text{sales}) + \underset{(0.013)}{0.023}\,\text{profmarg}$$

$$n = 32 \quad R^2 = 0.918$$

- $n = 31$ uç gözlem dışlanırsa:

$$\widehat{\log(\text{rd})} = \underset{(0.511)}{-4.404} + \underset{(0.067)}{1.088}\,\log(\text{sales}) + \underset{(0.013)}{0.0218}\,\text{profmarg}$$

$$n = 31 \quad R^2 = 0.9037$$

- Sonuçlar pratik olarak aynı. AR-GE harcamalarının satış esnekliği 1 olarak kabul edilebilir mi?