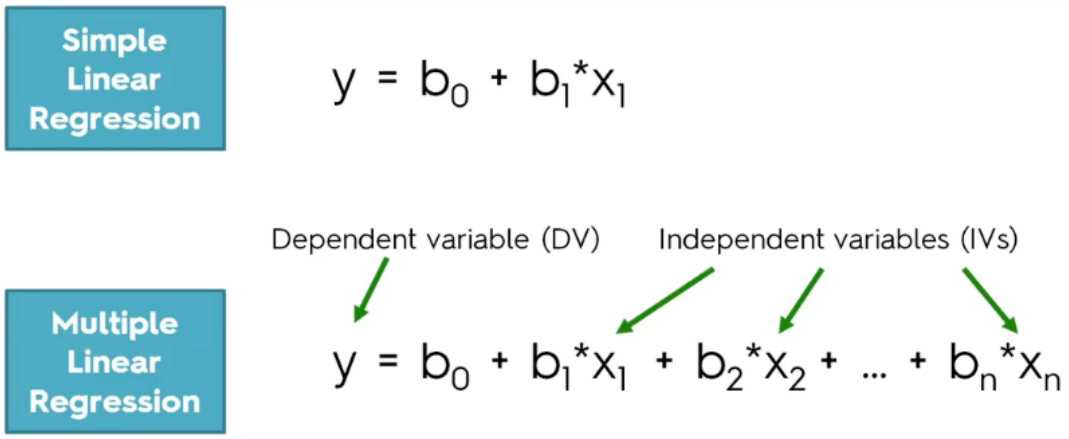
Çoklu Lineer Regresyon Nedir?

Ekonomi ve işletmecilik alanlarında herhangi bir bağımlı değişkeni tek bir bağımsız değişken ile açıklamak mümkün değildir. Ekonomik modeller, genellikle birden fazla sebebin sonucudurlar. Çokfazla sayıda değişken bir araya gelerek bir diğer değişkeni etkileyebilmektedirler. Bu değişkenler aynı zamanda kendi aralarında da birbirlerini etkileyebilmektedir. Bu sebeple, bu tür birden fazla değişkenin kullanılması gereken durumlarda tekli regresyon analizi yapılması mümkün değildir. Birden Fazla bağımsız değişken kullanılarak yapılan regresyon analizinde "çoklu regresyon analizi (multiple regression analysis)" adı verilmektedir.



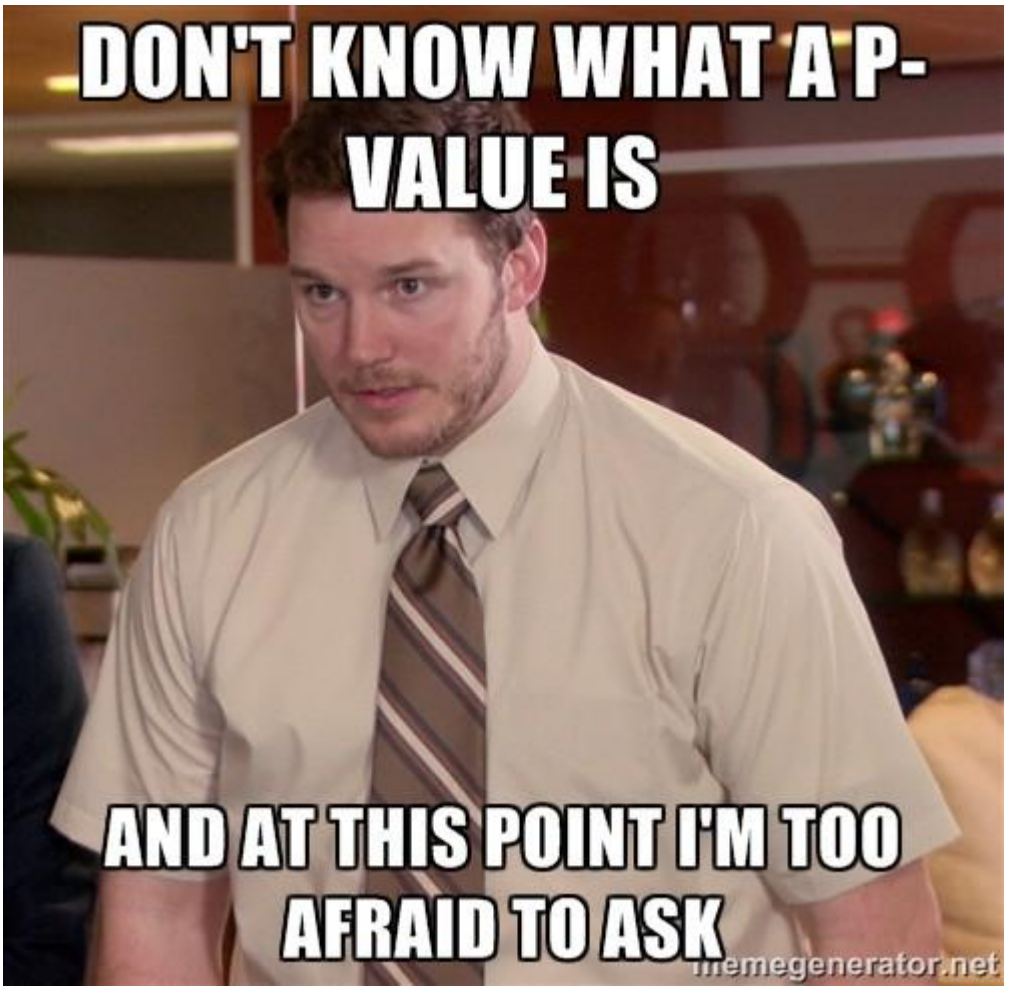
Çoklu lineer regresyonda her bağımsız değişkenin bağımlı değişkene etkileme derecesi birbirinden farklıdır. Bundan dolayı basit lineer regresyon daki denkleme ek olarak her değişkenin katsayısı aynı olmak zorunda değildir. Diğer yandan bu yazı serisinin 3. yazısında kategorik niteliklerin nümerik hale getirilmesi konusunu anlatmıştım. O yazıda kategorik nitelikleri olan değişkende birden fazla çeşit veri olduğu için (Hatay, İstanbul, Karaman gibi), her birine bir numara vermiştik fakat iki çeşit beriden oluşan kategorik değişkenlerde boolean (yani 0 ve 1) mantığını kullanabiliriz. İstatistik biliminde kategorik verilerin yerine bu şekilde kullandığımız sanal verilere "dummy variables" denir.

Dummy Variables

Profit	R&D Spend	Admin	Marketing	State	New York	California
192,261.83	165,349.20	136,897.80	471,784.10	New York	1	0
191,792.06	162,597.70	151,377.59	443,898.53	California	0	1
191,050.39	153,441.51	101,145.55	407,934.54	California	0	1
182,901.99	144,372.41	118,671.85	383,199.62	New York	1	0
166,187.94	142,107.34	91,391.77	366,168.42	California	0	1

$$y = b_0 + b_1*x_1 + b_2*x_2 + b_3*x_3 + b_4*D_1$$

P Değeri Nedir? Ne işe yarar?



P değeri, bizim hipotezlerinin doğru olup olmadığını belirlememize yardımcı olan istatistiksel bir ölçüdür. P değerleri, deney sonuçlarının gözlemlenen olaylar için normal değerler aralığında olup olmadığını belirlemek için kullanılır. Genellikle, bir veri setinin P değeri belli bir önceden belirlenmiş miktarın altında ise (örneğin, 0.05 gibi), bilim insanları deneylerinin "boş

hipotezini" reddedeceklerdir - başka bir deyişle, hipotezi ekarte edeceklerdir. Deney değişkenlerinin sonuçlar üzerinde anlamlı bir etkisinin olmadığı anlaşılmış olacaktır diğer bir deyişle. Eğer bu meseleyi daha detaylı anlamak istiyorsanız aşağıdaki videoyu izlemeniz oldukça yararlı olacaktır.

İşte biz de değişkenlerin çoklu lineer regresyon içerisindek P değerlerini bulup, belirli bir eşiğin üstündeki P değerine sahip olan değişkenleri veri setinden çıkaracağız. Böylece modelimizi optimize edeceğiz.

Çoklu Lineer Regresyon Modeli Oluşturmak

Çoklu lineer regresyon modeli oluştururken dikkat etmemiz gereken şeylerin sayısı basit lineer regresyon da oldukça fazladır çünkü hangi değişkenin önemli olup olmadığını anlamak çoklu lineer regresyon için çok temel ve çok hassas bir noktadır. Bağımlı değişkene oldukça etkiyen bir değişkeni veri setinden kaldırdığımız zaman oluşacak hatalar, yahut gereksiz bir değişkeni veri setinden atmadığımız zaman kaybedeceğimiz verim bize büyük zararlar olarak dönecektir. Bundan dolayı bu noktayı çok dikkatli bir şekilde almamız gerekmekte. İşte bundan dolayı da istatistik bilimiyle uğraşan insanlar bizler için çeşitli yöntemler belirlemişler. Bu yöntemleri sıralamak gerekirse:

1. Hepsi bir arada
2. Geriye doğru eleme (eliminasyon)
3. İleri doğru eleme
4. Çok yönlü eleme
5. Puan Karşılaştırması

Geriye Doğru Eleme

Geriye doğru eleme yöntemi tüm değişkenleri içeren bir modelden, fazlalıklarını atarak daha verimli daha küçük bir model oluşturma içeren algoritmadır. İşleyişi ise şöyledir:

1. Değişkenlerin modelde kalması için önem değeri belirlenir (Significance Level, SL = 0,05 gibi).
2. Modelle alakasız değişkenler modelden çıkarılır.
3. En yüksek P değerine sahip değişkeni bul. Eğer önem değerinden yüksekse (P>SL) bir sonraki adıma ilerle. Eğer düşükse modelin geriye doğru eleme algoritmasını tamamlamıştır.
4. P değeri önem değerinden yüksek değişkeni modelden çıkart ve 3. adıma dön.

İleri Doğru Eleme

İleri doğru eliminasyon (eleme) yöntemi geriye doğru eliminasyon yönteminden farklı olarak hiçbir bağımsız değişken içermeyen bir model olarak başlar yolculuğuna. Sonrasında ise hipoteze en yararlı olduğu tahmin edilen değişkenleri alarak kendisini daha büyük bir model haline getirir. İşleyişi ise şöyledir:

1. Modele giriş için bir önem değeri belirlenir. (Örneğin SL=0,05)
2. Bütün bağımsız değişkenler ve bağımlı değişkenle ayrı ayrı ikişerli modeller oluşturulur. En küçük P değerine sahip değişken seçilir.
3. Seçilen değişken modele eklenir. Bundan sonrasında ise belirlenen önem değerinden düşük olmakla beraber yine en küçük P değerine sahip değişken modele eklenir.
4. Bu işlem seçilen yeni değişkenin P değerinin önem değerinden fazla olmasına kadar devam eder. En küçük P değerine sahip yeni değişkenin P değerinin önem değerinden fazla olduğunu gördüğümüzde daha önce yüklediğimiz değişkenlerin model için yeterli olduğunu anlamalıyız.

Bu yazıda sizleri daha fazla yöntemle karmaşaya boğmak istemiyorum. Çok yönlü eleme yöntemi hakkında bilgi sahibi olmak isterseniz "bidirectional elimination" anahtar kelimesiyle araştırma yapabilirsiniz.