

Regresyon

Regresyon analizi, bağımlı bir değişken ile bir veya daha fazla bağımsız değişken arasındaki ilişkilerin tahmin edilmesi için kullanılan bir dizi istatistiksel yöntemdir. Örneğin, ne kadar yediğiniz ile kilonuz arasında bir bağlantı olduğunu tahmin edebilirsiniz; regresyon analizi bunu ölçmeye yardımcı olabilir.

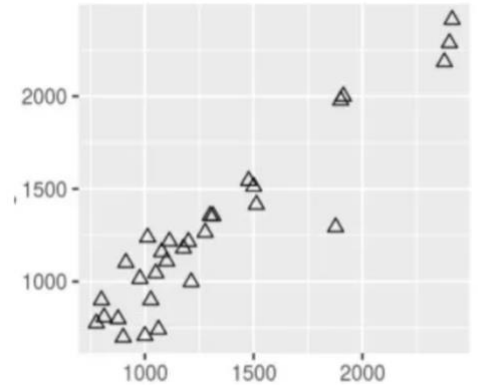
Regresyon analizi, doğrusal, çoklu doğrusal ve doğrusal olmayan gibi çeşitli varyasyonları içerir. En yaygın modeller doğrusal ve çoklu doğrusaldır. Doğrusal olmayan regresyon analizi genellikle, bağımlı ve bağımsız değişkenlerin doğrusal olmayan bir ilişki gösterdiği daha karmaşık veri kümeleri için kullanılır. Aşağıda bir regresyon tablosu örneği verilmiştir. Burada maaşın, tecrübe edilen yıl aralığı verilmiştir. Görüldüğü üzere tecrübe arttığında maaş da buna orantılı olarak artmıştır. Bizim çizdiğimiz çizgi bu oranlara en yakın şekilde konumlandırılmalıdır ki tahmin yapıldığında hata oranı en düşük olsun.

Bahsedeceğimiz regresyon türleri:

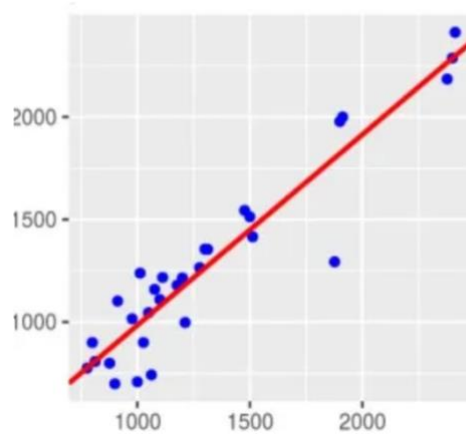
- Basit Lineer Regresyon
- Polinom Regresyonu
- Lojistik Regresyon
- Ridge Regresyonu
- Lasso Regresyon
- Elastic Net Regresyon

Basit Lineer Regresyon (Doğrusal Regresyon)

Aslında yukarıda verilen regresyon örneği burada bahsedeceğimiz doğrusal regresyon örneğidir. İki veya daha fazla değişken arasındaki ilişkinin bir doğru ile gösterilmesine lineer regresyon denir. Tek bir bağımsız değişken olduğunda regresyon modeline **Basit Doğrusal Regresyon**, birden fazla bağımsız değişken olduğunda ise **Çoklu Doğrusal Regresyon** modeli denir. Aşağıdaki verilen tabloya bazı veriler işlenmiştir. Buradan bir tahmin yapılabilir ama veriler düzenli dağılmadığı için sağlıklı bir tahmin elde etmek daha zordur.



Ama lineer regresyon bize daha doğru tahminde bulunabilmemiz için aşağıdaki gibi bir doğru çiziyor. Çizilen doğru sayesinde tahmin yapılabilmesi kolaylaşmıştır.



Peki lineer regresyonun nasıl ifade edilir? Basit lineer regresyon, aşağıdaki denklem kurularak ifade edilebilir.

$$Y = a + bX + \epsilon$$

Y: Bağımlı (sonuç) değişken olup belli bir hataya sahip olduğu varsayılır. Tahmin edilen değerdir. Gerçek y değeri ile arasındaki fark ne kadar az ise tahmin o kadar gerçeğe yakındır

X: Bağımsız (sebebi) değişken olup hatasız ölçüldüğü varsayılır

a: Sabit olup $X=0$ olduğunda Y'nin aldığı değerdir

b: Çizilen doğrunun eğimidir. Regresyon katsayısı olup, X'in kendi birimi cinsinden 1 birim değişmesine karşılık Y'de kendi birimi cinsinden meydana gelecek değişme miktarını ifade eder.

ϵ : Tesadüfi hata terimi olup ortalaması sıfır varyansı σ^2 olan normal dağılım gösterdiği varsayılır. Bu varsayım parametre tahminleri için değil katsayıların önem kontrolleri için gereklidir.

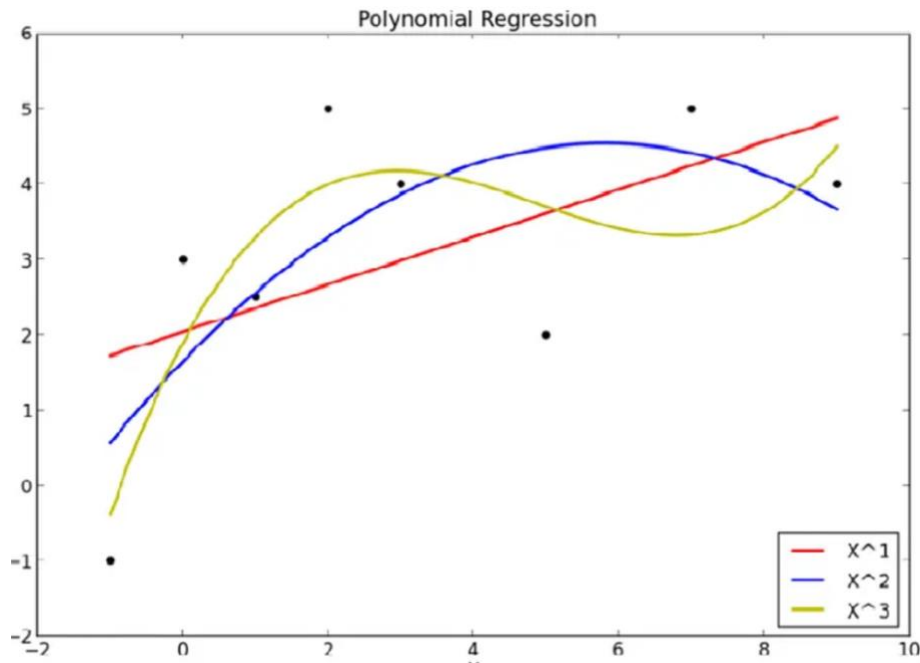
En Küçük Kareler Yöntemi (Least Square Method): Bu yöntem lineer regresyonun temel kavramıdır denilebilir. Yukarıdaki modellerde tahmin edilen değer ile gerçek değer arasındaki uzaklığa (residual) bakacak olursak bunun aslında modelin tahmininin hata payı olduğunu görebiliriz. Her tahmin değerinin bir residual değeri varsa ve bu 0'dan sonsuza kadar ilerliyorsa bizim tüm tahminlerin residual değerlerinin karelerini (eğer gerçek değer tahmin edilenden küçükse residual negatif değer alır bu yüzden karelerini alıyoruz) toplamamız bize hata paylarının toplamına denk gelen bir sayı verecektir. Öyle bir çizgi (linear regresyon çizgisi) oluşturacak lineer regresyon modeli yapmalıyız ki, bu çizgi residualları minimize etsin, yani bize en küçük hata payı toplamını versin . Kısaca bu yöntem hataları minimum yapmak üzerine kuruludur.

Polinom Regresyonu (Polynomial Regression)

Polinomsal regresyon basit lineer regresyona benzerdir. Fakat burada regresyon doğrusal bir şekilde ilerlemez. Bir doğru yerine eğriden (curve) oluşan bir regresyon söz konusudur. Kısacası değişkenler arasındaki ilişki doğrusal olmadığı durumlarda başvurulacak bir yöntemdir.

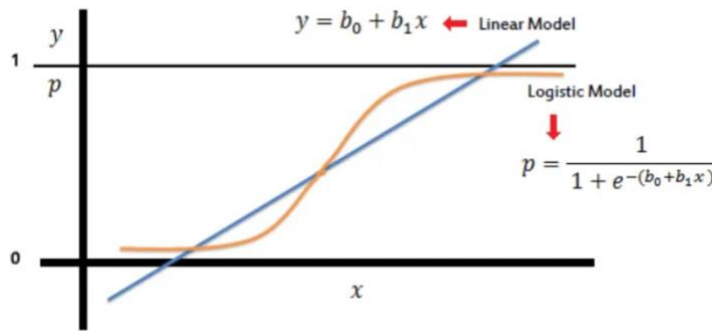
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_h X^h + \epsilon,$$

Bu denklemde 'h' polinom derecesini ifade eder. Aşağıdaki şekilde farklı derecelere ait polinom regresyonu verilmiştir. Hata payını azaltmak için bu yöntemle başvurulabilir.



Lojistik Regresyon (Logistic Regression)

Lojistik Regresyon her ne kadar bir regresyon çeşidi de olsa, sınıflandırma işlemi için kullanılır. Kategorik ya da sayısal veriler bu regresyonla sınıflandırılabilir. Burada bağımlı değişken (sonuç) sadece iki farklı değer alabilir (Doğru/Yanlış, Evet/Hayır...). Örneğin; bir e mailin spam veya spam olmadığını tahmin etmek için kullanılabilir. Aşağıdaki şekilde hem basit lineer regresyon hem de lojistik regresyon verilmiştir. Lojistik regresyonda benzer olarak düzlem üzerinde verileri yakalamaya çalışır ama farklı olarak görüldüğü üzere doğrusal değildir. Kendi formülü gereği logaritmik bir eğri üzerinde verileri yakalar. Bu da inişli çıkışlı verilerde daha yüksek tahmin başarısı sağlamaktadır.



Ridge Regresyonu (Ridge Regression)

Ridge regresyonu çok değişkenli verileri analiz etmede kullanılır. Amaç hata kareler toplamını minimize eden katsayıları, bu katsayıları bir ceza uygulayarak bulmaktır. Over-fittinge karşı dirençlidir. Çok boyutluluğa çözüm sunar. Tüm değişkenler ile model kurar, ilgisiz değişkenleri çıkarmaz sadece katsayılarını sıfıra yaklaştırır. Modeli kurarken alpha (ceza) için iyi bir değer bulmak gerekir. Yani modeli tekrar tekrar çalıştırdığımızda farklı katsayılar elde edeceğiz. Gerçek değere en yakın bulma işlemini ridge regresyonu ile sağlıyoruz.

Lasso Regresyon (Lasso Regression)

Lasso (Least Absolute Shrinkage and Selection Operator) Ridge regresyonuna benzer. En temel fark Ridge L2 penalty kullanırken, Lasso L1 penalty kullanmaktadır. Yani ridge katsayıların karesini alırken lasso katsayıların mutlak değerini alır. Lasso regresyonu sadece aşırı öğrenmeyi azaltmak için değil aynı zamanda öznelik seçimi (feature selection) konusunda da önemli bir rol oynar.

Elastic Net Regresyon (Elastic Net Regression)

Burada da amaç Lasso ve Ridge ile aynıdır. Farkı bu iki regresyonu birleştirmesidir. Ridge'den cezalandırma, Lasso'dan değişken seçimi yapar.