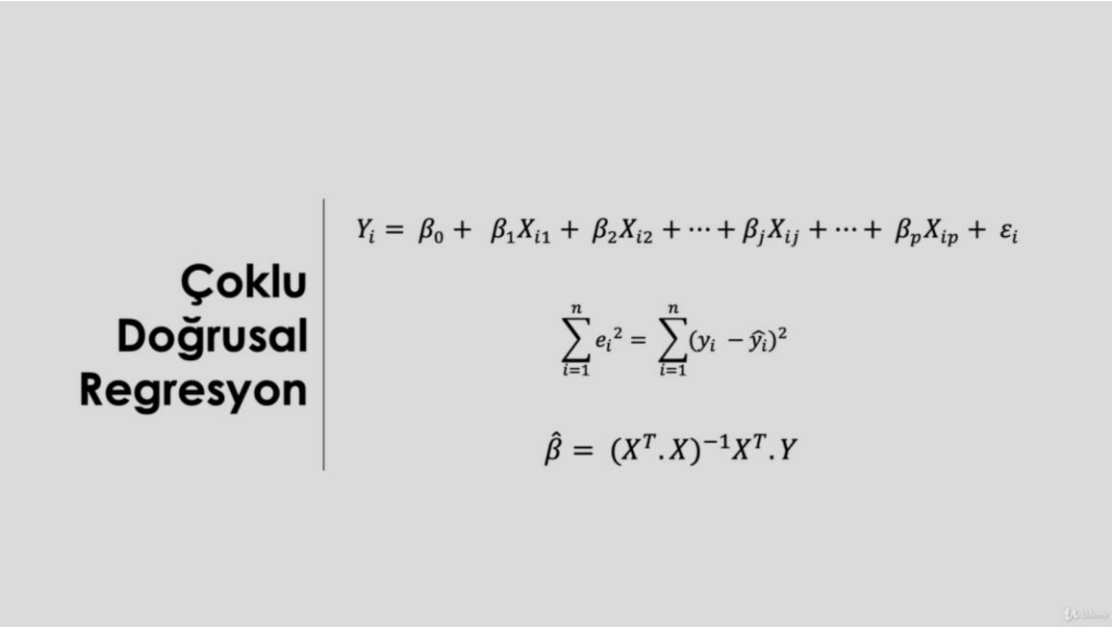
Udemy - (50 Saat) Python A-Z™: Veri Bilimi ve Machine Learning

10. Bölüm - Makine Öğrenmesi II – Doğrusal Regresyon ve Kuzenleri (264. Video) - Vahit Keskin anlattı.

Multiple Linear Regression (Çoklu Doğrusal Regresyon)

Temel amacımız basit doğrusal regresyonda olduğu gibi bağımlı ve bağımsız değişkenler (sadece buraya ler eklemiş olduk) arasındaki ilişkiyi ifade eden doğrusal fonksiyonu bulmaktır. Nasıl? Hata kareler toplamını minimize edecek şekilde kat sayı tahminlerini bulmaya çalışarak.

Çoklu Doğrusal Regresyon yapan veri bilimcinin genelde iki amacı vardır. Bunların birincisi; bağımlı değişkeni etkilediği belirlenen değişkenler aracılığıyla bağımlı değişkenin değerlerinin tahmin edilmesi, ikincisi ise; bağımlı değişkeni etkilediği düşünülen bağımsız değişkenlerden hangisinin veya hangilerinin bağımlı değişkeni ne yönde, ne şekilde etkilediğini tespit edebilmek, aralarındaki ilişkiyi tanımlamaya çalışmak. Örneğin bir değişken negatif yönlü etkiliyor olabilir ki kilometre artışı araç fiyatını bu şekilde negatif etkiliyor olacaktır, vites durumu pozitif etkiliyor olacaktır gibi bu değişkenlerin bağımlı değişkene olan etkileri, yönü ve bunların şiddetleriyle ilgili yorumlanabilir modeller çıkarmaktır.



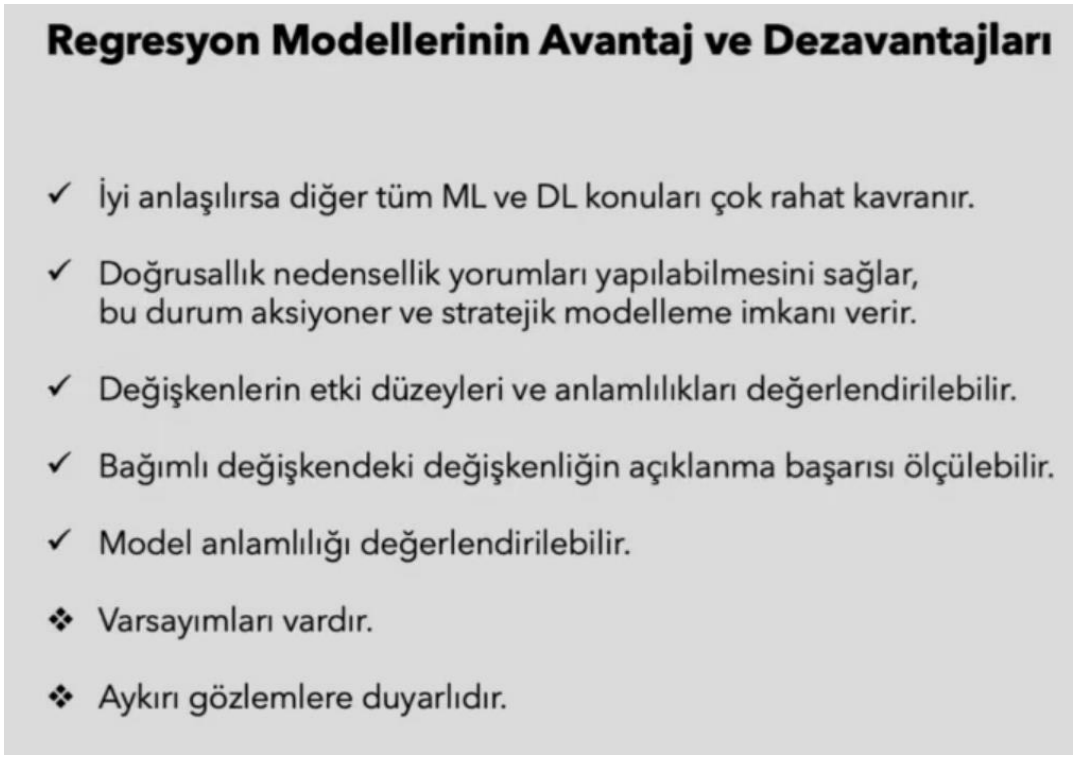
Burada amacımız hata kareler toplamını minimum yapacak kat sayıları bulmak. Bu denklem çözüldüğünde kat sayılar elde edilecektir. Bu denklemin çözülmesinde bayesian bazı metodlar, monte carlo metodu, analitik metodlar dediğimiz normal denklem yöntemleri gibi çeşitli metodlar kullanılmaktadır. Least squares dediğimiz en küçük kareler yöntemi bu konudaki en çok bilinen yöntemlerden birisidir. Buradaki temel amaç buradaki kat sayıları bulmaktır. Bunu isterseniz en küçük karelerle, isterseniz normal denklemler formülüyle, isterseniz bayesçi metodlar yöntemleriyle bulup gerekli denklem kurulabilir.

Çoklu doğrusal regresyon modelini kurduğumuzda öncekine benzer şekilde şöyle bir çıktı elimizde olacak. Bu çıktı üzerinden model yorumlanır. Üst kısımda artıkların (residuals) dağılımı verilmiş. Modelin kat sayıları (coefficients) ve bunların anlamlılıkları verilmiş. Alt kısımda R^2 değeri, p-value da buna benzer değerler verilmiş. Çoklu doğrusal regresyon çıktısı aşağı yukarı bütün programlama dillerinde SAS, SPSS, R vb. dillerde bu model kurulduğunda çıktısı böyle bir şey olacaktır. Ve bu çıktıyı yorumlamak bizim için önemli olacaktır.

```
Call:
lm(formula = Sales ~ TV + Radio, data = caseStudyData)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7977 -0.8752  0.2422  1.1708  2.8328

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.92110    0.29449   9.919   <2e-16 ***
TV           0.04575    0.00139  32.909   <2e-16 ***
Radio        0.18799    0.00804  23.382   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.681 on 197 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8962
F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

Doğrusal Regresyonun Varsayımları

- Hatalar normal dağılır.
- Hatalar birbirinden bağımsızdır ve aralarında otokorelasyon yoktur.
- Her bir gözlem için hata terimleri varyansları sabittir.
- Değişkenler ile hata terimi arasında ilişki yoktur.
- Bağımsız değişkenler arasında çoklu doğrusal ilişki problemi yoktur.



- Bağımlı değişkendeki değişkenliğin açıklanma başarısı R^2 değerine denk geliyor.
- Varsayımları içerisinden özellikle çoklu doğrusal bağlantı problemi ve otokorelasyon problemi bizim sevmediğimiz varsayımlardandır. Çünkü çoklu doğrusal bağlantı problemi olduğunda bu, bağımsız değişkenlerin birbirleri arasında çok yüksek korelasyon olması anlamına gelir ve bunlar bazı problemlere sebep olur. Bu problemleri giderebilmek adına PCR, PLS, Ridge, Lasso vb. bazı yöntemler önerilmiştir.