REGRESYON

DOGRUSAL REGRESYON VE KUZENLERİ

✓ **Basit Doğrusal Regresyon:** Temel amaç bağımlı ve bağımsız değişken arasındaki ilişkiyi ifade eden doğrusal ilişkiyi bulmak. Bağımlı ve bağımsız değişken arasındaki bu ilişki hata karelerinin toplamını minimize edecek katsayı tahminleri bulmaya çalışarak bulunacak.

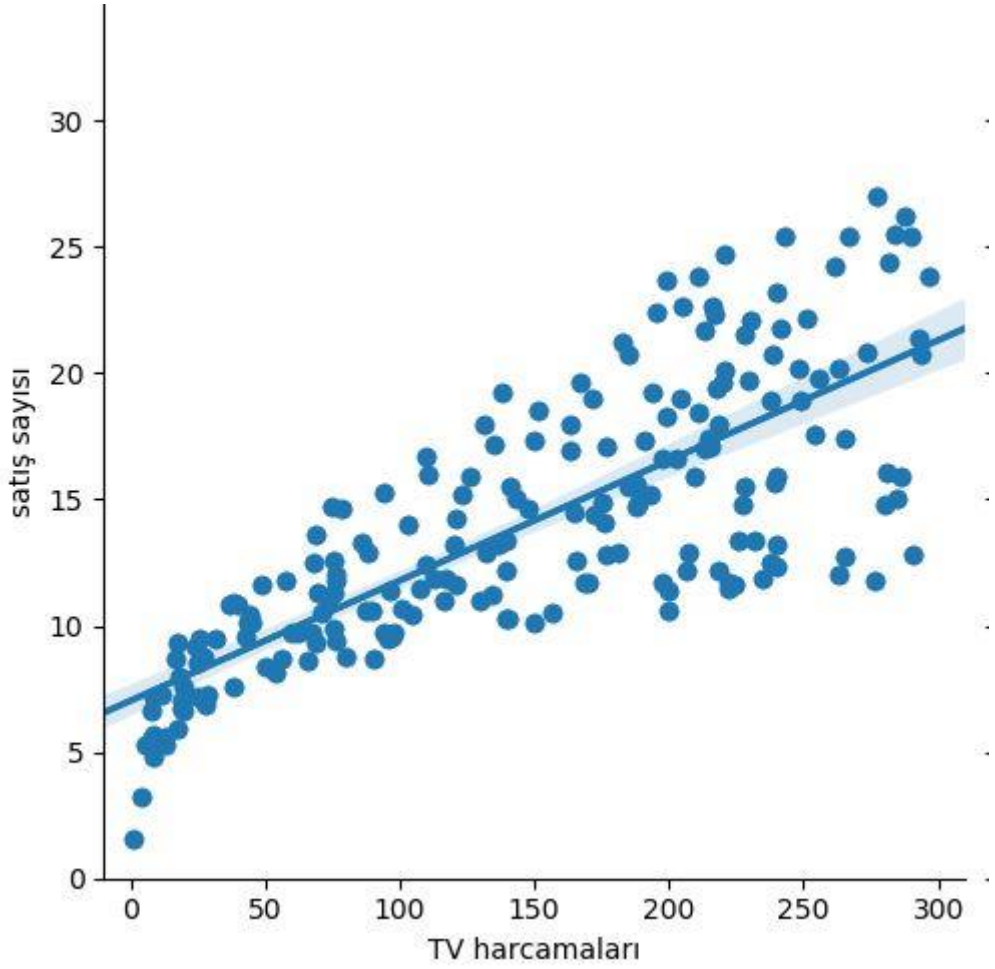
$SSE=\sum_{i=0}^{n}(y_i-\acute{y}_i)^2$

$y_i=b_0+b_1x_i+e_i \qquad \acute{y}= b_0+b_1x_i$

b_0 : doğrunun y eksenin kestiği nokta

b_1 : doğrunun eğimi

e : hata terimi



- R-Squared(R^2 degeri): Açıklanabilirlik oranıdır. Bağımsız değişkenin bağımlı değişkendeki değişkenliği açıklama başarısıdır. Modele ne kadar değişken eklersek ekleyelim şişmeye meyillidir. İlgili ilgisiz, anlamlı anlamsız, değişken olması durumunda artacaktır. Bu durum iyi bir şey değildir.
- Adj. R-Squared(Düzeltilmiş R^2 degeri): R^2'nin her parametre eklenmesi karşı duyarlılıpını törpüleyen, düzenleyen duyarlılığı daha az olan metriktir.
- F-statistic: Modelin anlamlığının anlaşılması için kurulan test istatistiğidir.

✓ **Çoklu Doğrusal Regresyon:** Bağımlı ve bağımsız değişkenler arasındaki ilişkiyi ifade eden doğrusal fonsiyonu bulmak.

$$\sum_{i=0}^{n}(e_i^2)=\sum_{i=0}^{n}(y_i-\acute{y}_i)^2$$

❖ **Varsayımları**

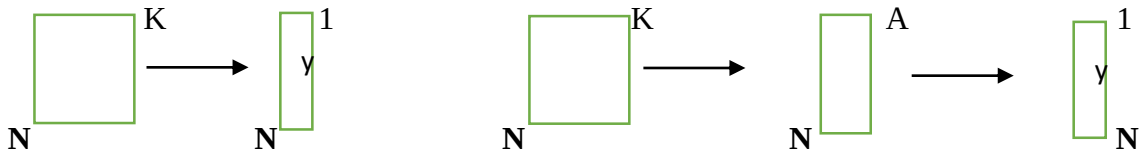
Hatalar normal dağılır, hatalar biririnden bağımsızdır ve aralarında otokolerasyon yoktur. Herbir gözlem için hata terimi rasında ilişki yoktur. Bağımsız değişkenler arasında çoklu doğrusal ilişki problemi yoktur.

Dezavantajları, varsayımları vardır ve aykırı gözlemlere duyarlıdır.

Avantajları ise, iyi amlaşılırsa ML ve DL konuları çok rahat kavranır. Doğrusallık ve nedensellik yorumları yapılabilmesini sağlar. Bu durum aksiyoner ve stratejik modelleme imkanı verir. Değişkenlerin etki düzeyleri ve anlamlılıkları değerlendirilir. Bağımlı değişkenler değişkenliğin açıklanma başarısı ölçğlebilir. Model anlamlılığı değerlendirilebilir.

<u>Not:</u> Basit doğrusal regresyon analizinde bir bağımlı ve bir bağımsız değişken söz konusu iken çoklu doğrusal regresyon analizinde ise bir bağımlı değşken varken iki yada daha fazla bağımsız değişken vardır.

✓ **Temel Bileşen Regresyonu(PCR):** Değişkenlere boyut indirgeme uyguladıktan sonra çıkan bileşenlerle regresyon modeli kurulması fikrine dayanır.



Çoklu Linear Regresyon Temel Bileşen Regreyonu

Çoklu doğrusal bağlantı, değişkenler arasında yüksek kolerayon olması, benzer tahminsel bilgileri barındırları anlamına gelir. Bu durum probleme sebep olur.

Çoklu dogrusal bağlantı problemi olursa ne olur?

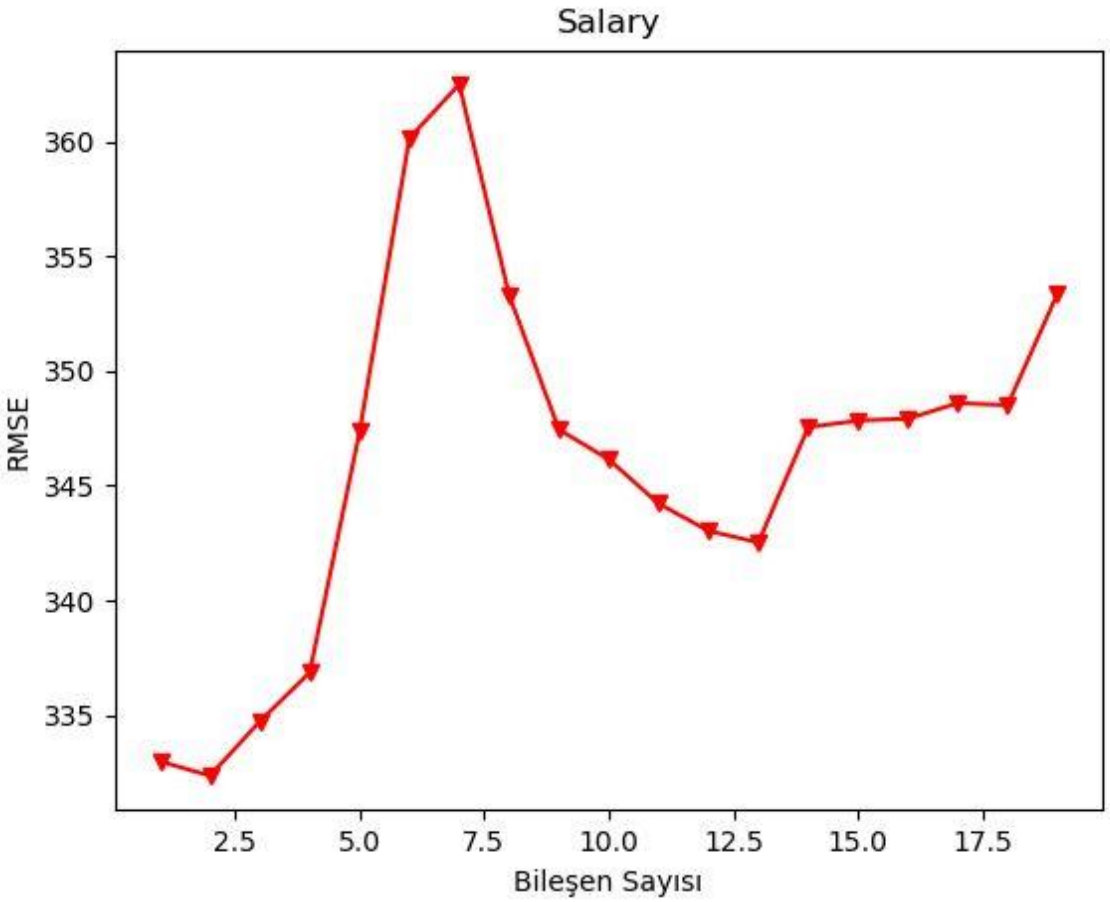
Tahmincilerin, katsayıların varyansını artırmakta ve yanlılık yaratmaktadır.

Gerçek hayatta veri setlerinde bazen değişken sayısının gözlem sayısından fazla olduğu durumlar ile karşılaşılıyor. Bu durum **çok boyutluluk laneti** olarak adlandırılıyor.

Temel bileşen analizi, elimizde p adet değişken olduğunda, bu p adet değişkenin içerdiği bilginin yüksek bir kısmını ondan daha az sayıda bir değişken ile ifade etme fikrine dayanır. Örneğin elimizde 100 tane değişken olsun. Bu 100 tane değişkenin içerdiği bir bilgi var. Bu 100 değişkenin içerdiği bilgiyi ondan daha az sayıda değişken ile bu bilginin maksimumunu temsil etmeye çalışmaktır.Daha az sayıda değişkenle temsil edilen bilgiye regresyon modeli uygulandığında buna temel bileşen regresyonu denir.

Çözüm olarak birincisi çok boyutluluk lanetini ortadan kaldırıyor. İkincisi burada oluşan bileşenler birbirleri ile kolerasyonlu olmuyor. (Bağımsız değişkenler olduğu için benzer bilgileri barındırmıyor.)

✓ **Kısmi En Küçük Kareler Regresyonu(PLS):** Değişkenler arttırılmak istenmiyorsa ve açıklanabilirlik aranıyorsa kullanılır. Değişkenlerin daha az sayıda ve aralarında çoklu doğrusal bağlantı problemi olmayan bileşenlere indirgenip regresyon modeli kurulması fikrine dayanır



PLS→gözetimli boyut indirgeme prosedürü

Bileşenler bağımlı değişken ile olan kovaryansını maksimum şekilde özetleyecek şekilde oluşturulur.

PCR→gözetimsiz boyut indirgeme prosedürü

Bileşenler bağımsız değişken uzayındaki değişkenliği maksimum şekilde özetleyecek şekilde oluşturulur.

Biz boyut indirgeme işlemi yapıyoruz fakat en son kuracak olduğumuz ve tahmin üretmesini beklediğimiz model söz konusu olduğunda ana veri setimizde 10 tane değişken varsa biz bu değişkenlerin aslında hepsinden kurtulmuyoruz. Yeni bir gözlem birimi geldiğini düşünelim. Bu gözlem birimi verisetinde ilk hali ile kaç tane değişken varsa o değişken sayısı kadar gözlem

değeri göndermesi gerekir. Dolayısıyla veriseti bize u gözlem değerlerini gönderdiğinde biz bu katsayıların üzerine bunları çarpıştırıp, burdan çıkan tahmin değerini sonuç olarak döneriz. Bunu dönebilmek adına PLS ve PCA'da değişken sayısı kadar katsayı çıkmasını bekleriz.

Modelleme işleminin yapılması için, bileşen sayısının ayarlanması ise ayrı bir işlemdir.

Biz bileşen sayısını ayarlayarak, model kurma sürecinde bazı problemlerden kurtuluyoruz ve nihayetinde kurmuş olduğumuz modelin tahmin modeli olarak kullanılması süreci söz konusu olduğunda elimizde veri setinde yer alan değişkenler kadar katsayı olması söz konusu olur.

✓ **Ridge Regresyon:** Amaç hata kareler toplamını minimize eden katsayıları, bu katsayılara bir ceza uygulanarak bulmaktır.

$$SSE_{l2}=\sum_{i=0}^{n}(y_i-\acute{y}_i)^2+\lambda\sum_{j=0}^{p}(B_j^2) \rightarrow \text{Ceza terimi}$$

l_2 : Düzenlileştirme normu

λ : Ayar parametresi

- Aşırı öğrenmeye karşı dirençlidir.
- Yanlıdır fakat varyansı düşüktür(bazen yanlı modeller tercih edilir).
- Çok fazla parametre olduğunda EKK'ya göre daha iyidir.
- Çok boyutluluk lanetine karşı çözüm sunar.
- Çoklu doğrusal bağlantı problemi olduğunda etkilidir.
- Tüm değişkenler ile model kurar, ilgisiz değişkenleri modelden çıkarmaz,katsayılarını sıfıra yaklaştırır.
- λ kritik roldedir.İki terimin (formüldeki) göreceli etkilerini kontrol etmeyi sağlar.
- λ için iyi bir değer bulunması önemlidir. Bunun CV kullanılır.
- Bazı katsayılar bazen etkilerini önemlerine göre kaybediyor ama diyoruz ki yinede bunlar modelde kalsın, sıfıra yaklaştıralım ama komple modelden çıkaralım.

✓ **Lasso Regresyon:** Amaç hata kareler toplamını minimize eden katsayıları, bu katsayılara bir ceza uygulayarak bulmaktır.

Ridge regresyonda farkı katsayılara uygulanan ceza işlemini biraz abartarak katsayıların cezalarını, onları sıfır yapacak şekilde uygulamaktır.

$$SSE_{l2}=\sum_{i=0}^{n}(y_i-\acute{y}_i)^2+\lambda\sum_{j=0}^{p}|B_j| \rightarrow \text{Ceza terimi}$$

l_2 : Düzenlileştirme normu

λ : Ayar parametresi

- Ridge regresyonun ilgili ilgisiz tüm değişkenleri modelde bırakma dezavantajını gidermek için önerilmiştir.

- Lasso katsayıları sıfıra yaklaştırır. Fakat l1 normu λ yeteri kadar büyük olduğunda bazı katsayıları sıfır yapar. Böylece değişken seçimi yapmış olur.
- λ' nın doğru seçilmesi çok önemlidir. Burada da CV kullanılır.
- Ridge ve Lasso yöntemleri birbirinden üstün değildir.

Ayar parametresinin belirlenmesi

- λ'nın sıfır olduğu yer EKK'dır. HKT'yi minumum yapan λ'yı arıyoruz.
- λ için belirli değerleri içeren bir küme seçilir ve her birisi için cross validation test hatası hesaplanır.
- En küçük cross validaton'ını veren λ ayar parametresi seçilir.
- Son olrak seçilen bu lambda ile model yeniden tüm gözlemlere fit edilir.

✓ **ElasticNet Regrasyonu:** Amaç hata karreler toplamını minimize eden katsayılara bir ceza uygulayarak bulmaktır.

ElasticNet l_1 ve l_2 yaklaşımlarını birleştir.

$$SSE_{ENet}=\sum_{i=0}^{n}(y_i-\acute{y}_i)^2+ \lambda_1\sum_{j=0}^{p}|B_j| +\lambda_2\sum_{j=0}^{p}(B_j^2)$$