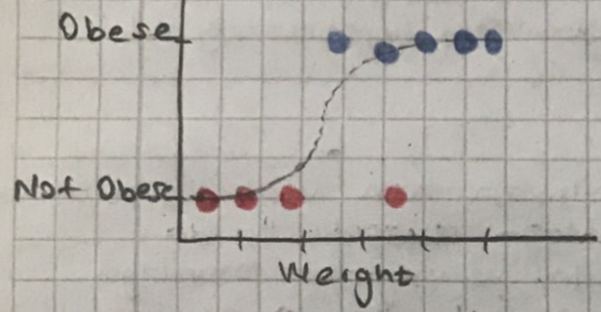
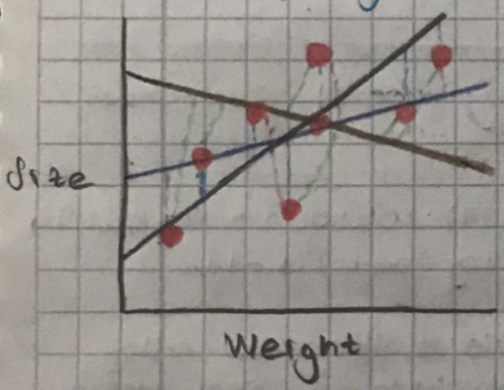
Youtube - Statquest with Josh Starmer
Logistic Regression Details Pt 2: Maximum Likelihood

* Bu dalgalı çizginin verilere en iyi şekilde uyması için nasıl optimize edildiğinden bahsedeceğiz.



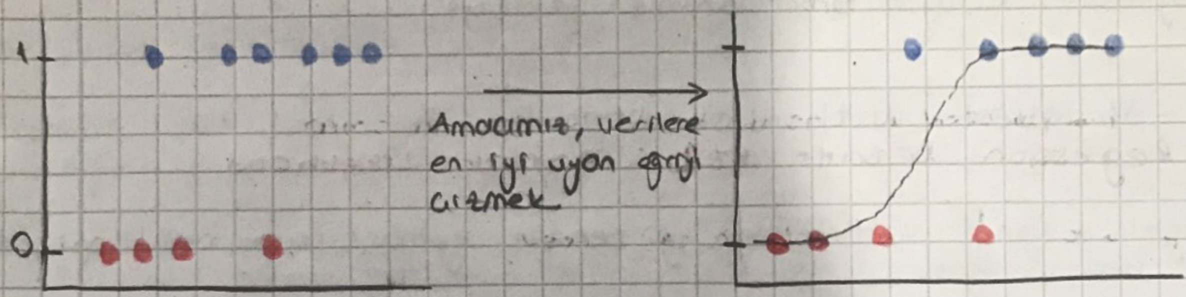
* Logistic regresyonun lineer regresyonla benzer mantıkta olduğundan fakat logistic regresyonun olasılıklarla ifade edilmesinden bahsetmiştik. Lineer regresyondaki gibi katsayıları grafik üzerinde görebileceğimiz doğrusal bir eğri elde edebilmek için olasılık değerlerini log(odds) cinsine çeviriyoruz. Bu eğriyi maximum likelihood ile çiziyoruz.

Lineer regresyonda eğriyi nasıl çiziyorduk?

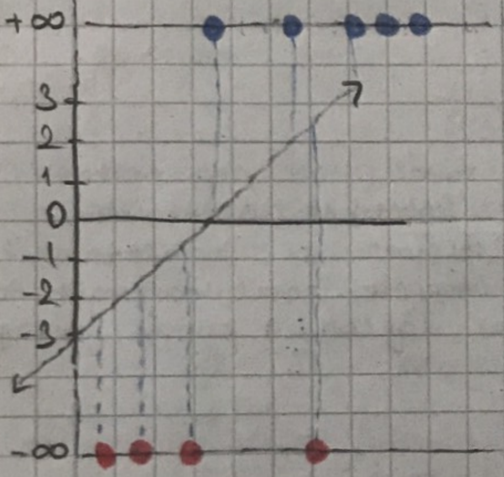


- Verilerimiz var en başta elimizde.
- Bu verilere least-squares (en küçük kareler) yaklaşımıyla bir line fit ediyoruz.
- Diğer bir deyişle veriyle line arasındaki uzaklıkları (residuals) ölçüyoruz. Bunların karesini alıyoruz (ki negatif değerler pozitif değerleri götürmesin) ve hepsini topluyoruz.
- Ve line'ı bir ölçüde döndürüyoruz. Tekrar residualları ölçüyoruz, karelerini alıyoruz ve onları topluyoruz.
- Bize en küçük sum of squared residualları veren (ki buna the least squares diyoruz) line'ı seçiyoruz.

Lojistik Regresyonda Nasıl Yapıyoruz?

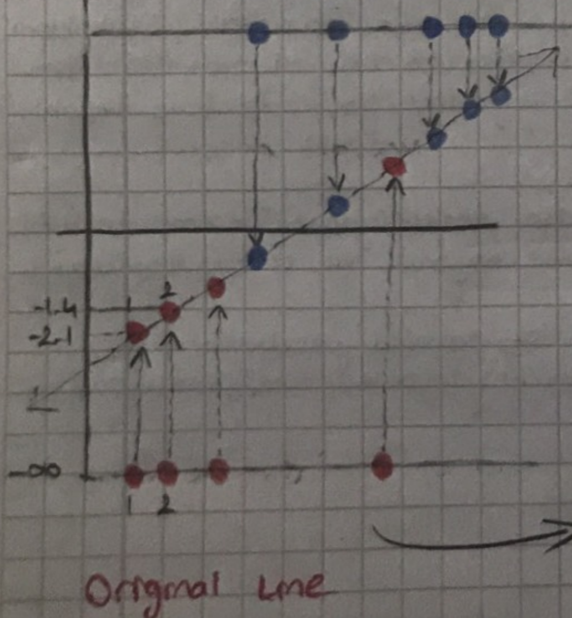


* Lojistik regresyonda y eksenindeki obezite olasılığını (probability of obesity) log(odds of obesity)'ye çeviriyorduk.



Tek sorun, dönüşümün ham verileri pozitif ve negatif sonsuzluğa itmesidir. Bu da, residualların pozitif ve negatif sonsuzluğa eşit olduğu anlamına gelir. Bu da en iyi line'ı bulmak için least-squares'i kullanamayacağımız anlamına gelir. Bunun yerine maximum likelihood'u kullanıyoruz.

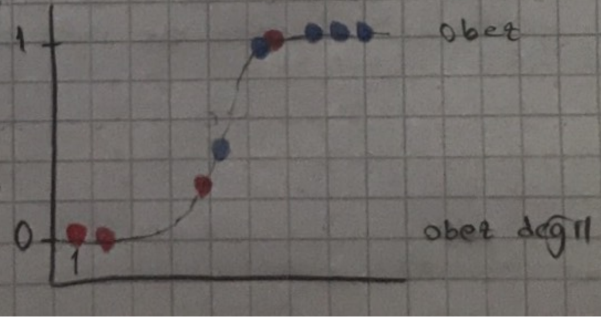
Yapacağımız ilk şey, orijinal veri noktalarını aday line'a yansıtmak. Bu, her örneğe bir aday log(odds) değeri verir.



Diğer bir deyişle (1) noktasının log(odds)'u -2.1, (2) noktasının log(odds)'u -1.4.

Ardından bu formülü kullanarak aday log(odds)'u aday olasılıklara (probabilities) dönüştürüyoruz.

$$p = \frac{e^{\log(odds)}}{1 + e^{\log(odds)}}$$



$p = \frac{e^{\log(odds)}}{1+e^{\log(odds)}}$ eşitliği $\log\left(\frac{p}{1-p}\right) = \log(odds)$ eşitliğinin yeniden düzenlenmesiyle elde edilen bir eşitlik.

$\log\left(\frac{p}{1-p}\right) = \log(odds)$ → iki tarafın da üssünü alıyoruz.

$\frac{p}{1-p} = e^{\log(odds)}$ → her iki tarafı $1-p$ ile çarpıyoruz.

$p = (1-p) \cdot e^{\log(odds)}$ → $(1-p)$'yi $e^{\log(odds)}$ ile çarpıyoruz.

$p = e^{\log(odds)} - pe^{\log(odds)}$ → $pe^{\log(odds)}$'u karşı tarafa geçiriyoruz.

$p + pe^{\log(odds)} = e^{\log(odds)}$ → p parantezine alıyoruz.

$p\left(1+e^{\log(odds)}\right) = e^{\log(odds)}$ → her iki tarafı $1+e^{\log(odds)}$'e bölüyoruz.

$$\boxed{p = \frac{e^{\log(odds)}}{1+e^{\log(odds)}}}$$

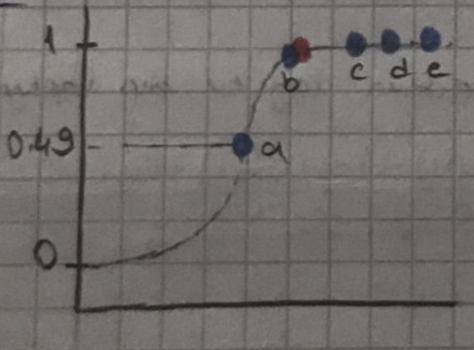
(1) noktası için bu formülü uygularsak;

$$p = \frac{e^{-2.1}}{1+e^{-2.1}} = 0.1$$

0.1 değeri bize dalgalı eğride noktanın nereye konacağı bilgisini verir. Tüm noktalar için aynı işlemi tekrarlarız.

Şimdi verilen line'da likelihoodu hesaplamak için gözlemlenen durumu (obez veya değil) kullanıyoruz.

1



Obez olan farelerden başlıyoruz. (a) faresinin likelihoodu (y eksenini kesen değer), 0.49. Yani, bu farenin likelihoodu, predicted probability ile aynı.

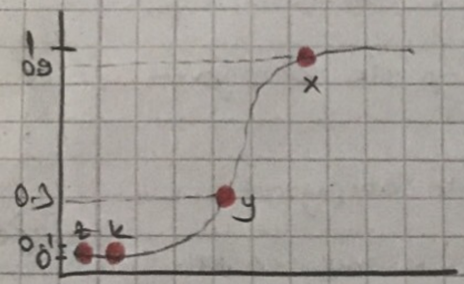
Bu case'de probability eğrinin altında kalan alan ile değil de y eksenindeki değerle hesaplanıyor, bu yüzden likelihood ile aynı.

(b), (c), (d), (e) noktalarının likelihoodu sırasıyla 0.9, 0.91, 0.91 ve 0.92

Obez olan tüm farelerin likelihoodu her bir likelihoodun çarpımı olarak hesaplanıyor.

Verilerin likelihoodu = $0.49 \times 0.9 \times 0.91 \times 0.91 \times 0.92 \times \ldots$

(2) * Şimdi obez olmayan farelerin likelihoodunu hesaplayalım.



Not: Obez olma olasılığı ne kadar düşükse, obez olmama olasılığı da o kadar yüksektir.

Bu yüzden, bu fareler için;

likelihood = (1 - farenin obez olma olasılığı)

(x) noktası için obez olma olasılığı 0.9, bu yüzden obez olmamanın probabilitysi ve likelihoodu 1-0.9

(y) noktası için obez olma olasılığı 0.3, bu yüzden obez olmamanın probabilitysi ve likelihoodu 1-0.3

(z) ve (k) noktaları için obez olma olasılığı 0.01 bu yüzden obez olmamanın probabilitysi ve likelihoodu 1-0.01

Bu değerleri de tüm verilerin likelihoodunu hesapladığımız eşitliğe dahil ediyoruz.

Verilerin likelihoodu = $\underbrace{0.49 \times 0.9 \times 0.91 \times 0.91 \times 0.92}$ $\times \underbrace{(1-0.9) \times (1-0.3) \times (1-0.01) \times (1-0.01)}$

Verilerin likelihoodunu her bir pointin likelihoodunu çarparak hesaplamamız mümkün olsa da istatistikçiler likelihoodun logunu hesaplamayı tercih ederler.

Her ikisi de mümkün, çünkü likelihoodu maximize etmekle, likelihoodun logunu maximize etmek aynı şeydir.

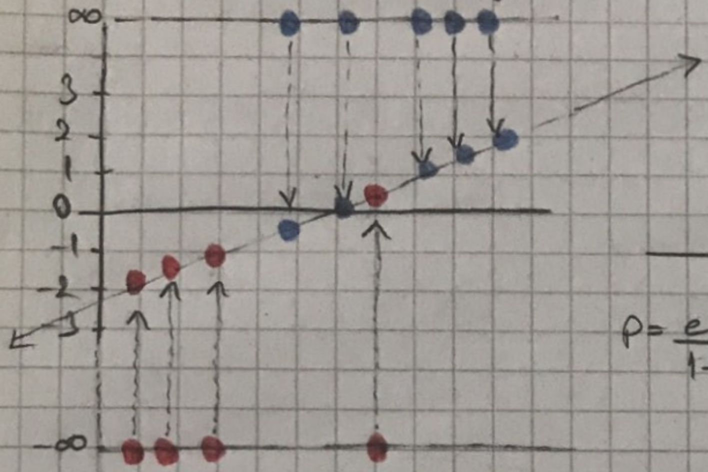
* Likelihoodun logunu aldığımızda her bir likelihoodun logunu alıp topluyoruz.

log (verilerin likelihoodu) = $\log(0.49) + \log(0.9) + \log(0.91) + \log(0.91) + \log(0.92) +$
$\log(1-0.9) + \log(1-0.3) + \log(1-0.01) + \log(1-0.01)$

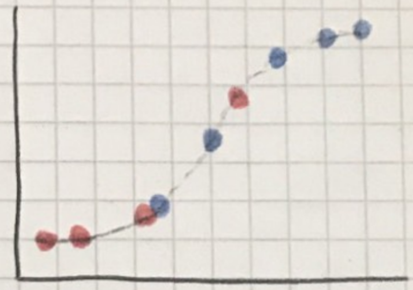
$= -3.77$

Dalgalı eğriye yerleşen noktaların likelihoodu -3.77 çıktı. Bu da orijinal line da (log(odds)) log-likelihoodun -3.77 olduğu anlamına geliyor. (2.nin öncesi)

Orginal line'ı rotate ediyoruz ve verileri üzerine yansıtıp log(odds)'ları olasılıklara dönüştürerek log-likelihoodu hesaplıyoruz.



$$p = \frac{e^{\log(odds)}}{1+e^{\log(odds)}}$$

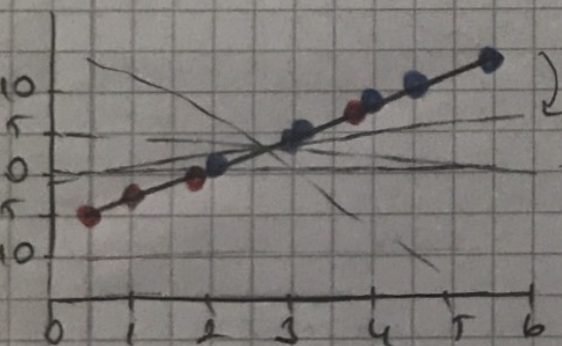


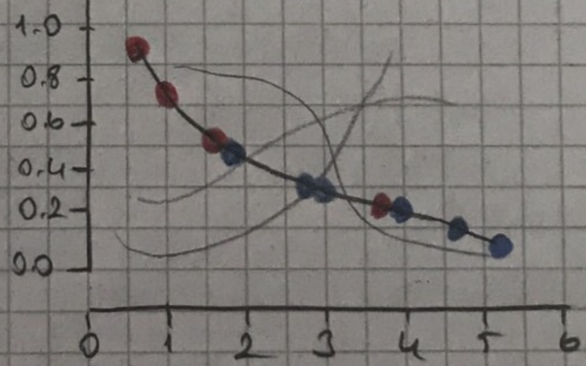
$$\log(\text{verilerin likelihoodu}) = \log(0.22)+\log(0.4)+\log(0.8)+\log(0.89)+\log(0.92)+$$
$$\log(1-0.6)+\log(1-0.2)+\log(1-0.1)+\log(1-0.05)$$

$$= -4.15$$

Bu line ilki kadar iyi değil.

Bu şekilde log(odds) line'ını döndürmeye ve verileri üzerine yansıtmaya log(odds)'ları olasılıklara dönüştürüp log-likelihoodu hesaplamaya devam ediyoruz.





Maksimum olasılığa (maximum likelihooda) sahip line'ı bulan algoritma oldukça akıllıdır. Line'ı her döndürdüğünde, bunu log olasılığını arttıracak şekilde yapar. Böylece algoritma birkaç döndürmeden sonra en uygun uyumu bulabilir. Sonunda, likelihoodu en üst düzeye çıkaran bir line elde ederiz ve bu en uygun olan olarak seçilir.

Bunun sonrasında bu line'ın faydalı bir modeli temsil edip etmediğini bilmek istiyoruz ve bu, bir R^2 değerine ve p-value'ya ihtiyacımız olduğu anlamına geliyor.