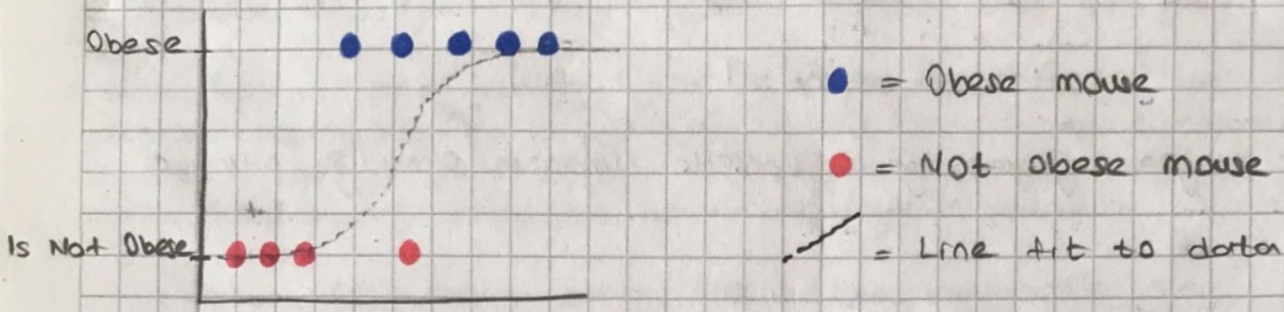
Youtube - Statquest with Josh Starmer

Logistic Regression Details Pt 1: Coefficients

* Bu video lojistik regresyonun nasıl çalıştığına dair büyük resmi göstermeyi amaçlıyor.



* Özellikle her lojistik regresyonun sonucu olan "coefficients"ları nasıl belirlediğimizi ve nasıl yorumlayacağımızı öğreneceğiz.

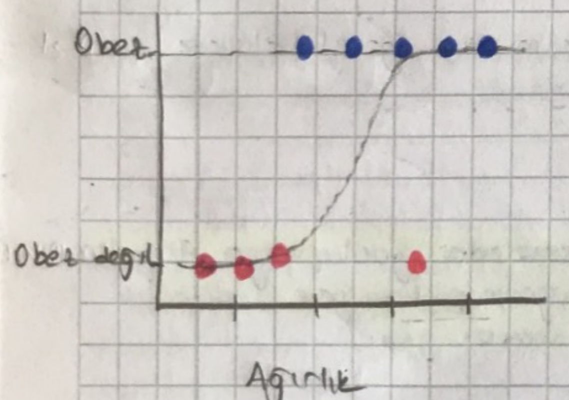
Coefficients;

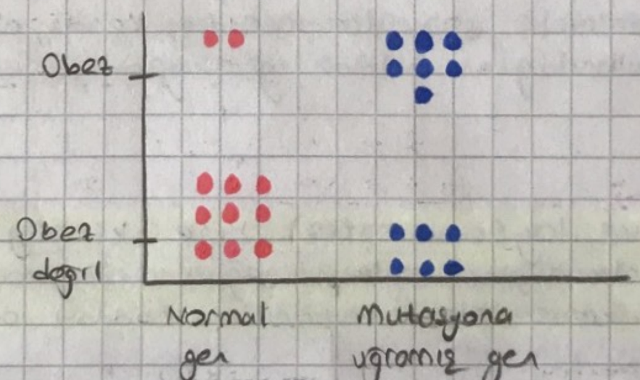
	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	-3.476	2.364	-1.471	0.1414
weight	1.825	1.088	1.678	0.0934

→ weight

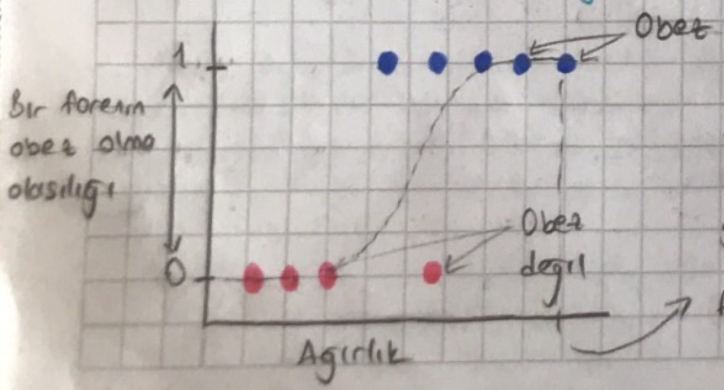
* Kat sayıları hem continuous variable contexti içerisinde hem de discrete variable contexti içerisinde öğreneceğiz.

mutasyona uğramış gen veya değil





Main Ideas About Logistic Regression

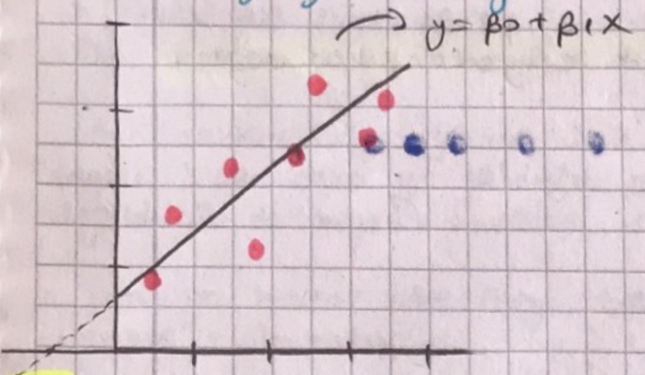


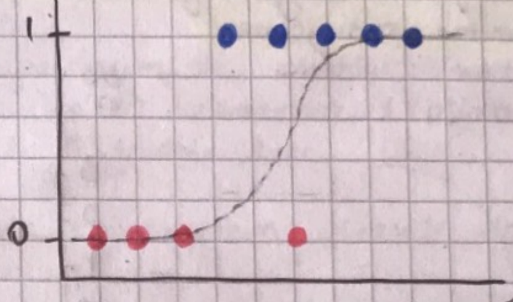
Y ekseni bir farenin obez olma olasılığını veriyor. Farenin obez olmaması 0'dan, olması 1'e uzanıyor. Çizilen eğri verilen ağırlığa göre farenin obez olma olasılığını veriyor.

→ Ağırlığı fazlaysa obez olma ihtimali daha yüksek.

* Lojistik regresyon belirli bir genelleştirilmiş lineer model türüdür. Genelleştirilmiş (generalized) lineer modeller, düzenli lineer modellerin kavram ve yeteneklerinin bir genellemesidir.

Lineer Regresyon ve Lojistik Regresyon





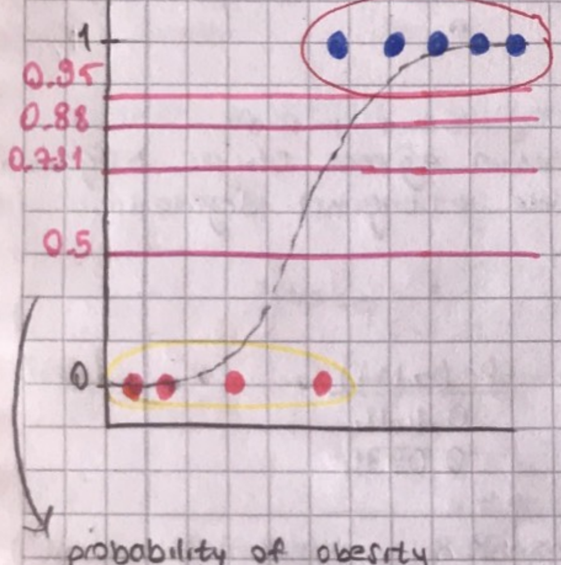
Kat sayıları ifade edebilmek için

Lineer regresyonda y ekseninde her türlü sayı yatabilirken

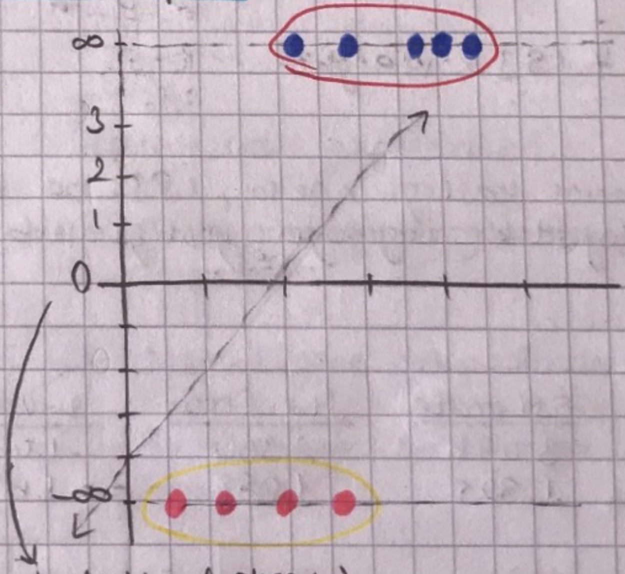
Lojistik regresyonda y ekseninde 0 ile 1 arasında olasılık değerleri yatıyor.

* Bu sorunu aşmak için lojistik regresyondaki y ekseni "probability of obesity (obezite olasılığından)", "log (odds of obesity)"'e dönüştürülür. Böylece, lineer regresyondaki gibi -sonsuzdan +sonsuza gidebilir.

$\boxed{\log(\text{odds of obesity}) = \log\left(\frac{p}{1-p}\right)}$ olarak ifade edebiliyorduk.



probability of obesity



log (odds of obesity)

$\log\left(\frac{p}{1-p}\right) \rightarrow$ p = 0.5 için $\log\left(\frac{0.5}{0.5}\right) = \log 1 = 0$

p = 0.95 için $\log\left(\frac{0.95}{1-0.95}\right) = 3$

p = 0.7 için $\log\left(\frac{0.731}{1-0.731}\right) = \log(2.717) = 1$

$p = 1 \rightarrow \log\left(\frac{1}{0}\right) = \log 1 - \underbrace{\log 0}_{-\infty} = \infty$

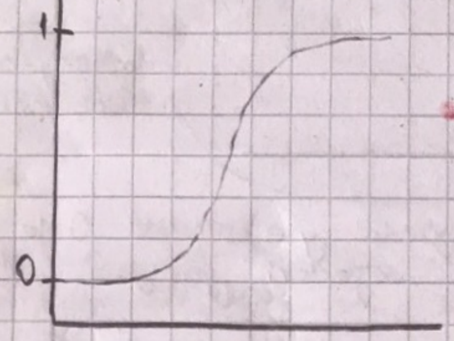
p = 0.88 için $\log\left(\frac{0.88}{1-0.88}\right) = \log(7.33) = 2$

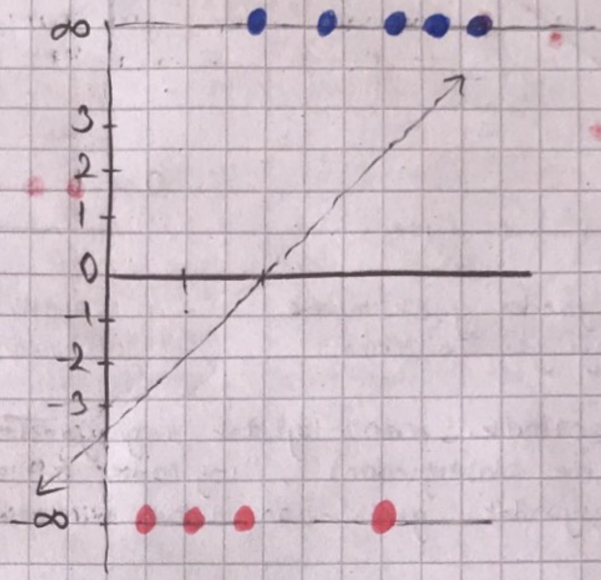
$p = 0 \rightarrow \log\left(\frac{0}{1}\right) = \log 0 - \log 1 = -\infty$

Olasılığı 0.5 ile 1 arasında olan değerler yeni y ekseninde 0 ile ∞ arasına yerleşirken, olasılığı 0 ile 0.5 arasında olan değerler yeni y ekseninde 0 ile -∞ arasına yerleşiyor.

Yeni y eksenimizle dalgalı olan line düz line'a dönüşmüş oldu.

* Lojistik regresyonu dalgalı grafikle ifade ediyorken, kat sayıları log(odds) cinsinden sunuyoruz.





Kat sayıları ifade etmek için çizdiğimiz bu grafiği maximum likelihood ile çiziyoruz. Bu line'da da linear regresyonda olduğu gibi;

$y = -3.48 + 1.83 \times \text{weight}$

-3.48 y eksenini kestiği nokta, 1.83 ise doğrunun eğimi oluyor. Doğrunun kat sayıları, lojistik regresyon yaptığımızda elde ettiğimiz değerlerdir.

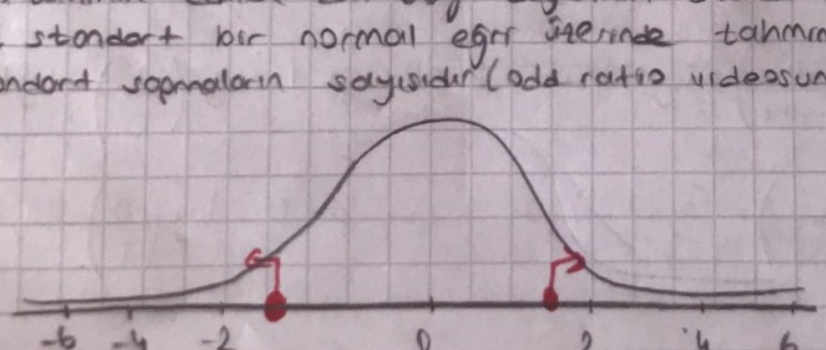
Coefficients: → p-value

	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	-3.476	2.364	-1.471	0.1414
weight	1.825	1.088	1.678	0.0934

→ İlk kat sayı weight=0 olduğundaki y eksenini kesen değerdir. Yani weight=0 olduğunda log(odds of obesity) = -3.476.

$-3.476 / 2.364 = -1.471$

→ z-value: tahmin edilen kat sayı değerinin standart hataya oranıdır. Başka bir deyişle, standart bir normal eğri üzerinde tahmini kestirimlerin 0'dan uzak olduğu standart sapmaların sayısıdır (odd ratio videosundaki Wald Test)



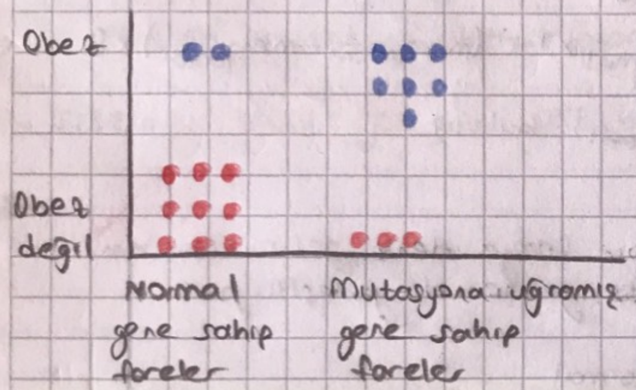
0'dan 2 standart hata uzaklığından daha az bir uzaklıkta (-1.471) olduğu için, istatistiksel olarak bunun anlamlı olmadığını söyleyebiliriz. Bu da p-value'nun (0'dan 1.471 standart sapma daha uzak olan standart eğrinin altında kalan alan) büyük olacağını garanti ediyor.

→ İkinci kat sayı eğim. Her bir ağırlık artışında log(odds of obesity) 1.825 artar.

Aynı şekilde weight'in katsayısı için de, standart hata sapmalarının sayısı 2'den daha az olduğu için (1.678) istatistiksel olarak anlamlı değildir. Bu da büyük p-value değerini garanti eder.

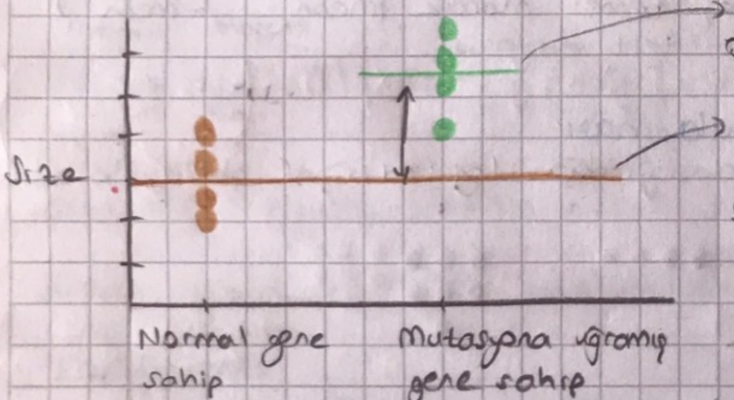
Şimdiye kadar obeziteyi tahmin etmek için sürekli değişken olarak weight'i kullanmıştık.

Discrete Variable Kullanarak Obeziteyi Tahmin Etmek



Bu tür bir lojistik regresyon, lineer modeller kullanılarak bir t testinin yapılmasına çok benzer.

Lineer modelleri kullanarak t-testini nasıl yapıyorduk?



→ Mutasyona uğrayan gene sahip farelerin boy ortalamaları

→ Normal gene sahip farelerin boy ortalamaları

Bu iki çizgi bu eşitliğin kat sayılarını oluşturuyor.

$$size = mean_{normal} \times \beta_1 + (mean_{mutant} - mean_{normal}) \times \beta_2$$

Daha sonra, genin normal veya mutasyona uğramış versiyonuna sahip olan bir farenin boyutunu tahmin etmek için bu denklemi bir design matrix ile eşleştiriyoruz.

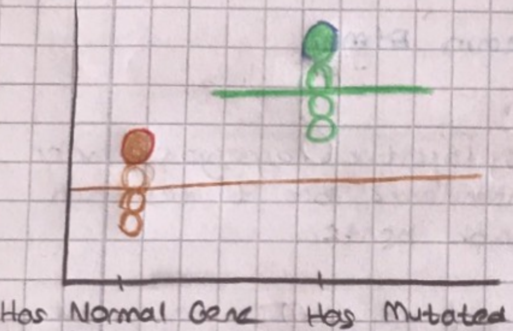
$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \rightarrow \text{Design matrix}$$

İlk column   ikinci column

$$size = mean_{normal} \times \beta_1 + (mean_{mutant} - mean_{normal}) \times \beta_2$$

İlk column β_1'in değerlerine karşılık geliyor ve ilk katsayı $mean_{normal}$'i aktive ediyor.

İkinci column β_2'nin değerlerine karşılık geliyor ve ikinci katsayı $(mean_{mutant} - mean_{normal})$'ı aktive ediyor. Aktive etmekten kastımız 0 veya 1 olması β_1/β_2'nin.



İlk row için (1-0) normal gene sahip bir fareye denk gelir. Size'ını tahmin etmek için β_1 yerine 1, β_2 yerine 0 yazıyoruz.

$$size = mean_{normal} \times 1 + (mean_{mutant} - mean_{normal}) \times 0$$

$size = mean_{normal}$ buluruz.

(1-1) rowu mutasyona uğramış gene sahip bir fareye denk gelir. Farenin boyunu tahmin etmek için β_1 yerine 1, β_2 yerine 1 yerleştiriyoruz.

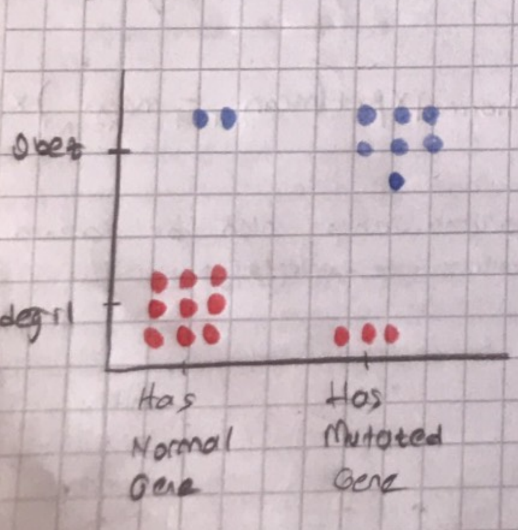
$$size = mean_{normal} \times 1 + difference_{mutant-normal}$$

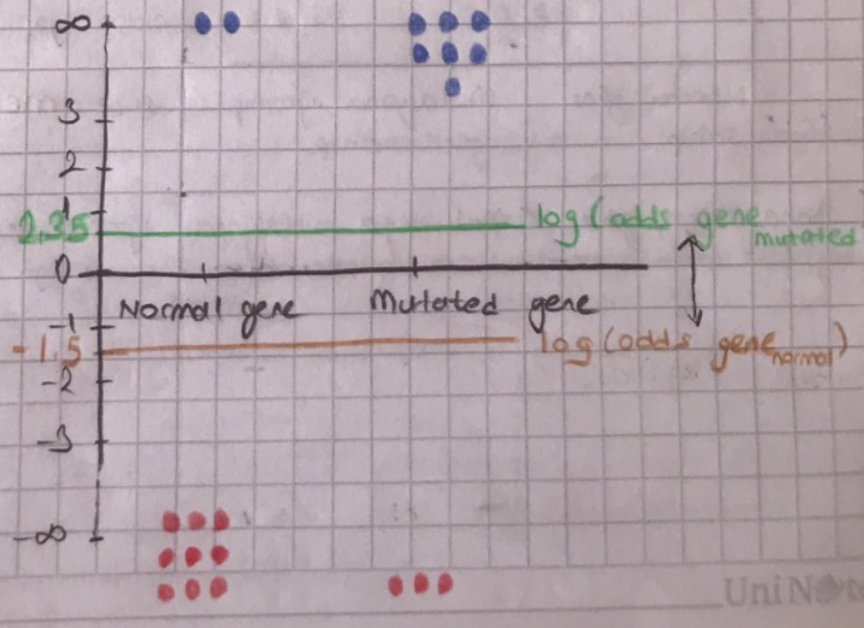
$$size = mean_{normal} + (mean_{mutant} - mean_{normal})$$

Bu şekilde bir t testi yaptığımızda, temel olarak $(mean_{mutant} - mean_{normal})$ bu katsayının 0'a eşit olup olmadığını test ediyoruz.

T testinin lojistik regresyona uygulanması

İlk olarak y eksenini obez olma olasılığından log(odds of obesity)'e dönüştürmek





Şimdi verilere iki doğruyu fit ediyoruz. İlk doğru için "Normal Gen" verilerini alıyoruz ve normal gene sahip fareler için log (odds of obesity) hesaplamak için kullanıyoruz. Buna log (odds $gene_{normal}$) adını verdik.

mavi → farelerin sayısını kullandık, olasılığını değil.

$$\log\left(\frac{2}{9}\right) = \log(0{,}22) = -1{,}5$$

kırmızı

Daha sonra mutasyona uğramış gene sahip fareler için log (odds of obesity) hesaplıyoruz. Bunu da log (odds $gene_{mutated}$) olarak isimlendirdik.

mavi

$$\log\left(\frac{7}{3}\right) = \log(2{,}33) = 0{,}85$$

kırmızı

Bu iki çizgi; size = log (odds $gene_{normal}$) $\times \beta_1$ + (log (odds $gene_{mutated}$) − log (odds $gene_{normal}$)) $\times \beta_2$

bu eşitlikteki kat sayıları şekillendirmek için bir araya geliyor.

Bu eşitliği şu şekilde tekrar yazabiliriz:

$$size = \log(\text{odds } gene_{normal}) \times \beta_1 + \underbrace{\log\left(\frac{\text{odds } gene_{mutated}}{\text{odds } gene_{normal}}\right)}_{\log(\text{odds ratio})} \times \beta_2$$

log (odds ratio) bize mutasyona uğramış gene sahip olmanın bir farenin obez olma olasılığını ne kadar arttırdığını (veya azalttığını) log ölçeğinde bize söyler.

$$size = \log(2/9) \times \beta_1 + \log\left(\frac{7/3}{2/9}\right) \times \beta_2$$

$$size = -1{,}5 \times \beta_1 + 2{,}35 \times \beta_2$$

Burada, lojistik regresyon yaptığımızda elde ettiğimiz kat sayıları bulduk.

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1,5041	0,7817	-1,924	0,0544
geneMutant	2,3514	1,0427	2,255	0,0241

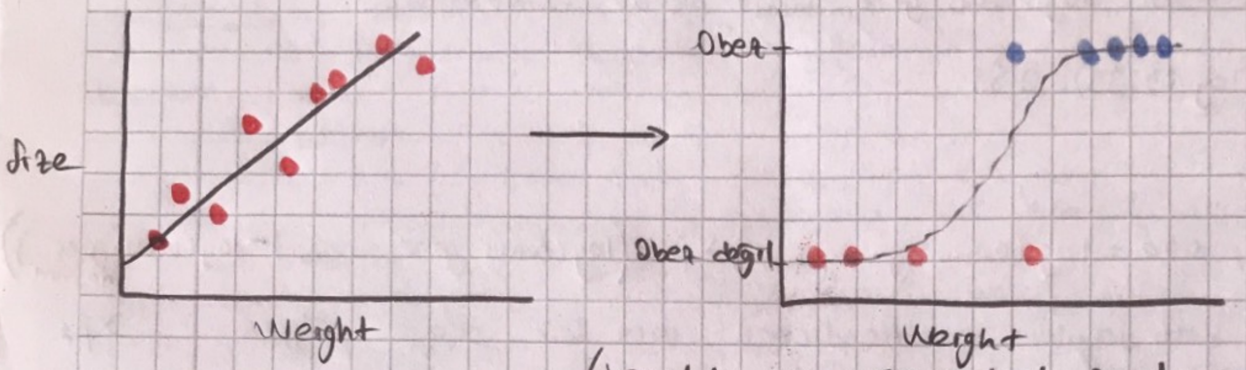
Intercept: log (odds $gene_{normal}$), geneMutant: log (odds ratio)

−1,5 intercept değerinin −1,9 z value değeri (interceptin 0'dan −1,9 standart hata kadar uzakta olduğunu ifade ediyor) 2'den küçük olduğu için istatistiksel olarak anlamlı değil.

gene Mutant'ın z-value su 2'den büyük olduğu için istatistiksel olarak anlamlı

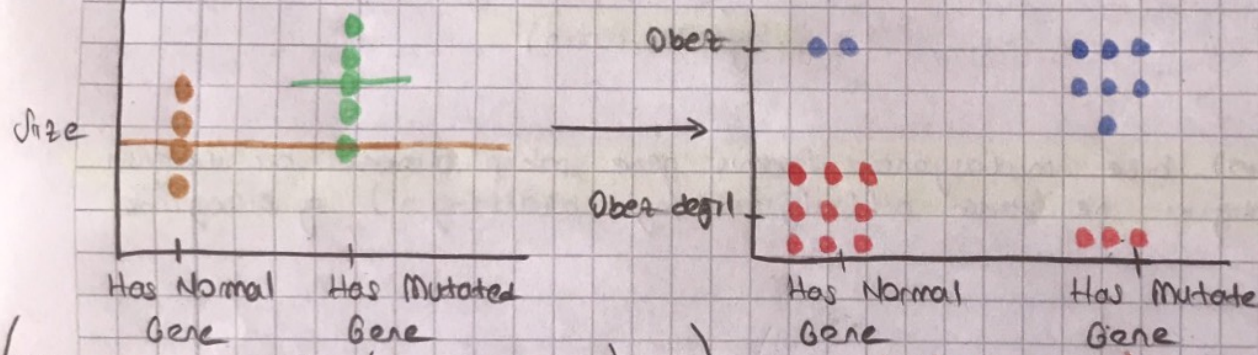
Özetlemek gerekirse;

1. Regresyon için bazı lineer model kavramlarının lojistik regresyona nasıl uygulandığını gördük



(Lojistik regresyonun kat sayılarını log(odds) eğrisi ile gösterdik)

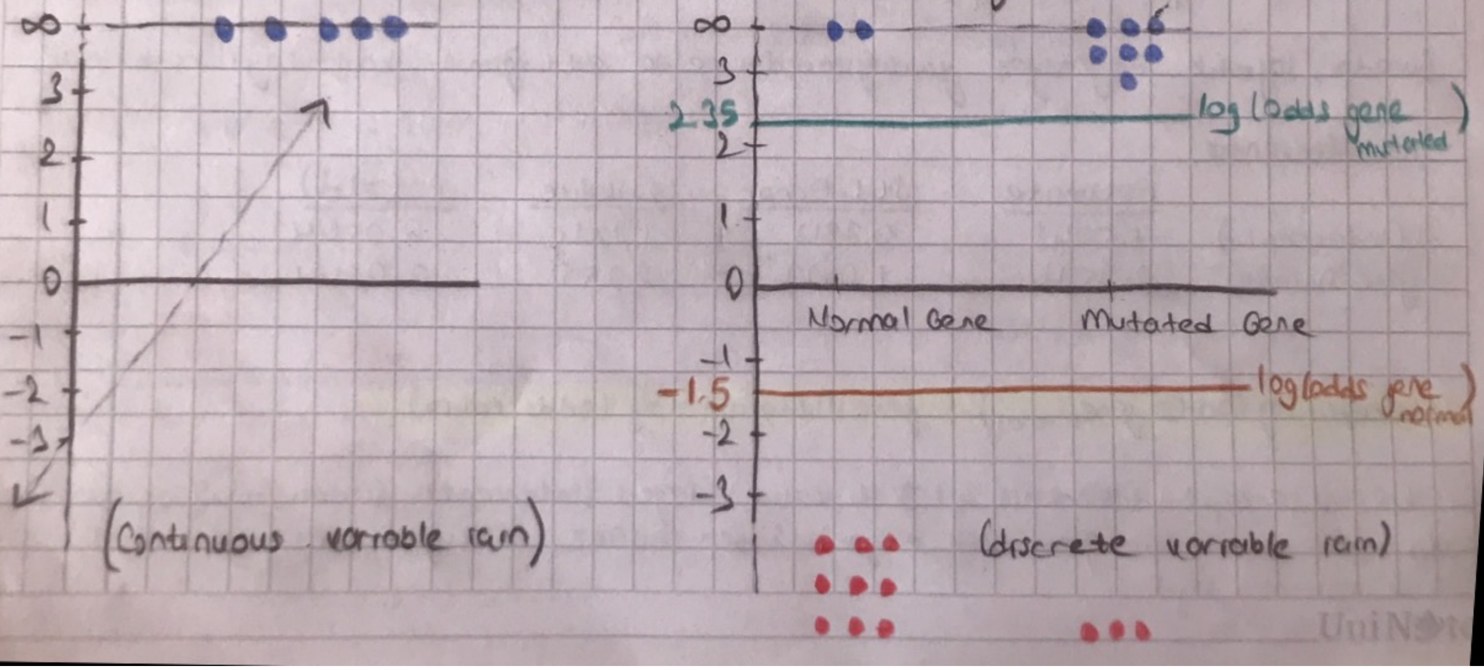
2. T testleri için bazı lineer model konseptlerinin lojistik regresyona nasıl uygulandığını gördük



$$size = mean_{normal} \times \beta_1 + (mean_{mutant} - mean_{normal}) \times \beta_2$$

$$size = \log(odds\ gene_{normal}) \times \beta_1 + \log\left(\frac{odds\ gene_{mutated}}{odds\ gene_{normal}}\right) \times \beta_2$$

Kısacası katsayılar açısından lojistik regresyon, katsayıların log(odds) cinsinden olması dışında, lineer modellerle tamamen aynıdır.



Bu, lineer modellerle yaptığımız şeyleri (multiple regression gibi) lojistik regresyon için de yapabileceğimiz anlamına geliyor. Burada sadece kat sayıların ölçeğinin log (odds) olduğunu hatırlamamız gerekiyor.