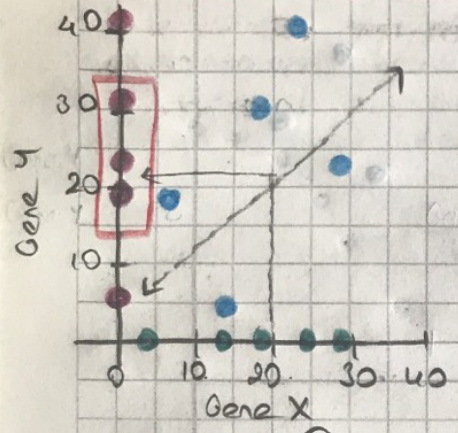
Youtube - Statquest with Josh Starmer

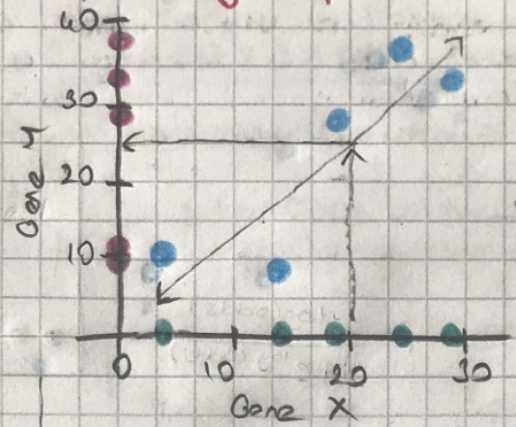
Pearson's Correlation, Clearly Explained!!!



(3)

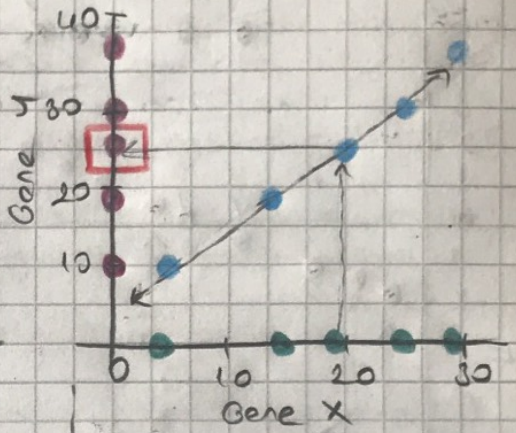
Eğer veri trend line'ından daha uzaktaysa bu durumda Gene Y daha geniş aralığa düşer.

Bu durumda, Gene X'in değerleri bize Gene Y'nin değerleri hakkında daha az bilgi verir.
Alternatif olarak, Gene X ve Gene Y arasındaki ilişkinin zayıf olduğunu söyleyebiliriz.



(1)

Gene X = 20 olduğunda, bu line ile Gene Y'nin 27 civarında olacağını tahmin edebiliriz. Tahminlerimizi veri de gözlemlediğimiz trende göre yapıyoruz.



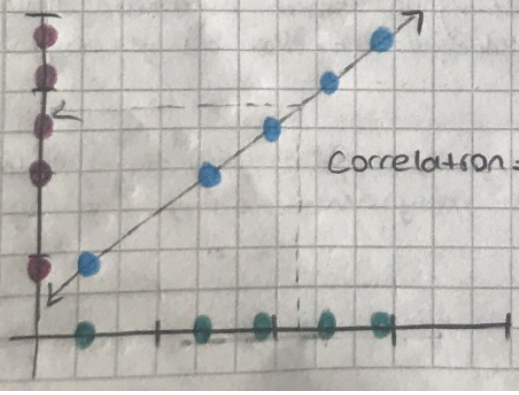
(2)

Eğer veri trend line'ına daha yakınsa verilen gene X değerine göre Y, daha küçük aralığa düşer. Bu durumda, veriler line'a ne kadar yakınsa X geri bize Y hakkında daha fazla bilgi verebilir. Alternatif olarak, Gene X ve Gene Y arasındaki ilişkinin nispeten güçlü olduğunu söyleyebiliriz.

Veriler arasında zayıf bir ilişki varsa	Veriler arasında orta bir ilişki varsa	Veriler arasında güçlü bir ilişki varsa
=	=	=
Küçük korelasyon değeri	Orta korelasyon değeri	Büyük korelasyon değeri

Korelasyonun Değeri

Pozitif eğimli düz bir çizgi, her veri noktasının merkezinden geçebildiğinde korelasyon = 1'dir.

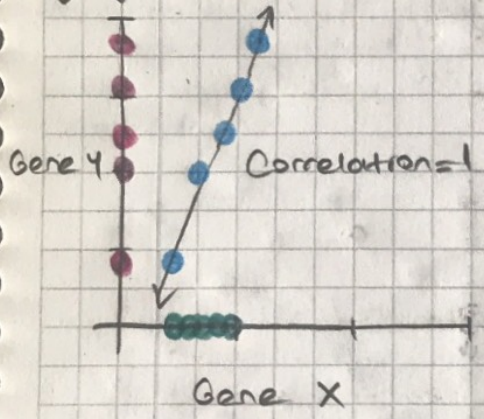


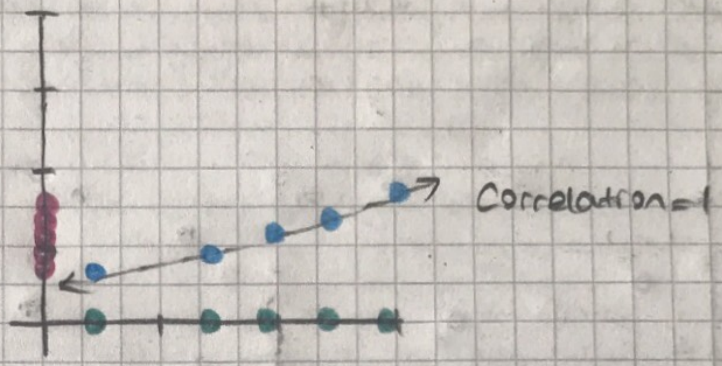
Bu da bize Gene X değeri verildiğinde Gene Y değerini çok dar bir alanda tahmin edebileceğimiz anlamına gelir.

(scale)

Note: Korelasyon verinin ölçeğine bağlı değildir.

Diğer bir deyişle, verinin ölçeği ne olursa olsun, pozitif eğimli bir doğru tüm verilerden geçebildiğinde korelasyon = 1 olur. Bu, eğim büyük olsa da küçük olsa da korelasyonun 1'e eşit olabileceği anlamına geliyor.

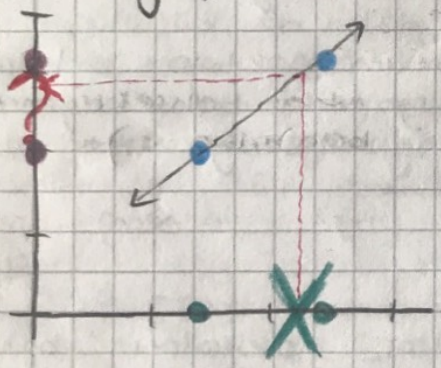




Korelasyonun 1'e eşit olmasının ne kadar data olduğuyla bir ilgisi yok.

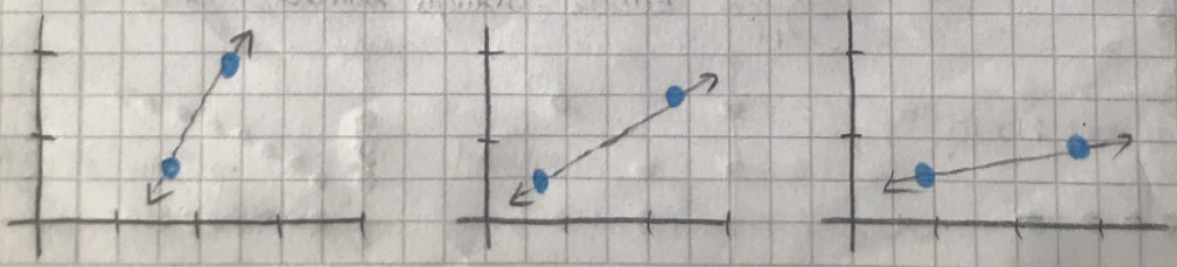
Korelasyon değerine ne zaman güvenebiliriz?

Örneğin yalnızca 2 veri noktası varsa, iki noktayı birleştirerek pozitif eğimli düz bir çizgi çizebiliriz ve korelasyon = 1 olur ve bu da ilişkinin güçlü görünmesini sağlar. Ancak bu çizgi ile yapılan tahminlere güvenmemeliyiz, çünkü elimizde çok az data var.

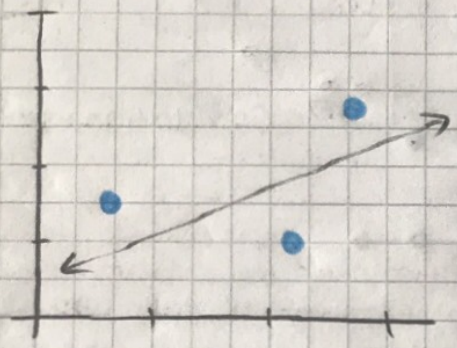


Küçük veri kümeleriyle yapılan korelasyonlara neden düşük güven duymalıyız?

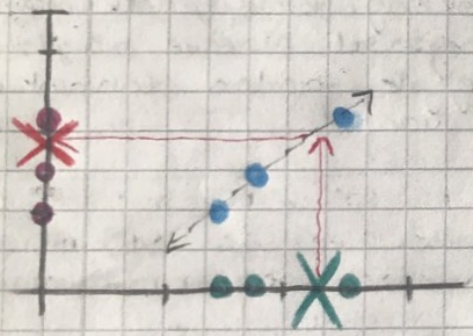
2 data pointimiz olduğunu düşünelim. Bu noktalar nerede olursa olsun 2 random nokta arasında her zaman noktaların merkezinden geçecek şekilde düz çizgi çizebiliriz.



2 data pointimiz olduğunu düşünelim. 2 nokta rain yaptığımız gibi 3 noktadan da düz bir çizgi çizebildiğimiz için korelasyon 1'dir, ancak şimdi bu çizgiyle yaptığımız tahminlere daha fazla güvenebiliriz. Bunun nedeni, boş bir grafikle başlayıp üzerine rastgele 3 nokta çizersek, herhangi 2 noktayı birleştirmek için düz bir çizgi çizmek kolay olsa da 3 noktanın hepsinden düz bir çizgi çizme şansımızın düşük olmasıdır.



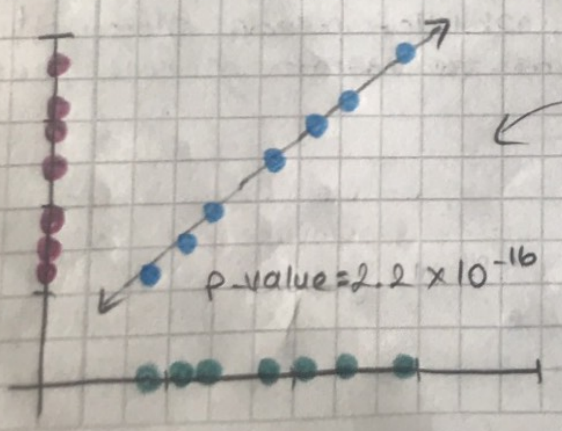
Dolayısıyla, rastgele çizilmiş 3 noktayı düz bir çizgiyle birleştirme olasılığımız çok küçüktür ve bu nedenle, gözlenen korelasyonun tesadüf sonucu gözlenmediğine daha fazla güvenebiliriz.



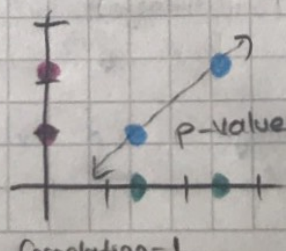
Genel olarak, ne kadar çok veriye sahipsek doğru ile yaptığımız tahminlere o kadar güveniriz çünkü aynı sayıda rastgele yerleştirilmiş noktadan düz bir çizgi çizme olasılığımız her ek nokta ile daha da küçülür. (Dalgalı bir eğriyle tüm noktaları birleştirebiliriz ama korelasyondan bahsettiğimiz zaman düz çizgiden bahsediyoruz.)

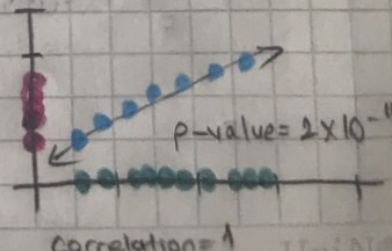
Verinin büyüklüğüyle p-value'nun ilişkisi:

Korelasyon ve p-value bize rastgele çizilmiş noktaların benzer şekilde güçlü bir ilişki veya daha güçlü bir ilişki ile sonuçlanma olasılığını söyler. Böylece, p-value ne kadar küçük olursa, doğru ile yaptığımız tahminlere o kadar güveniriz.



Bu case'te, p-value çok küçük: 2.2×10^{-16}
Bu, rastgele verilerin benzer şekilde güçlü veya daha güçlü bir ilişki yaratma olasılığının oldukça küçük olduğu anlamına gelir.



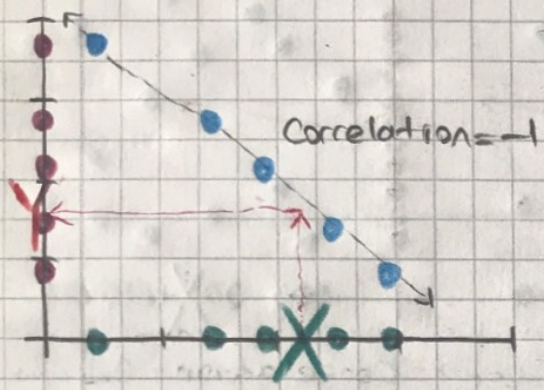


Veri arttıkça p-value azalıyor ve bu, bizim korelasyon değerine olan güvenimiz artıyor.

Negatif eğimli düz bir çizgi her veri noktasının merkezinden geçebildiğinde korelasyon = -1'dir.

---

* Düz bir çizgi tüm veri noktalarından geçebildiğinden, korelasyon verilerde güçlü bir ilişki olduğu anlamına gelir ve eğer birisi bize Ox'in bir değer verirse, Ox'in çok dar bir aralıktan bir değer tahmin edebiliriz.



Aynı şekilde, p-value ile nitelleştirdiğimiz bu tahmine olan güvenimiz, elimizde ne kadar veri olduğuna bağlıdır.

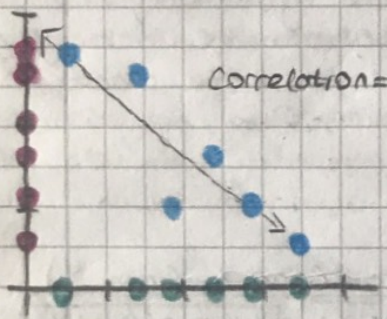
Çok fazla verimiz olsaydı, tahminimize çok güvenebilirdik çünkü p-value çok küçük olurdu. ($p\text{-value} = 2 \times 10^{-16}$)

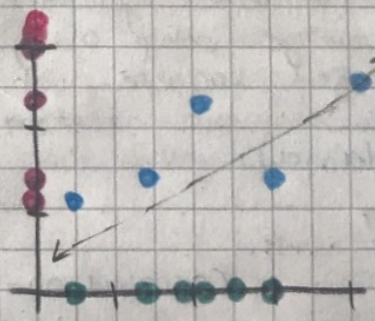
Ne kadar az veriye sahipsek, tahminimize o kadar az güvenimiz olur çünkü p-value büyür. (p-value = 1)

Aynı şekilde negatif eğimli bir doğru tüm verilerden geçebildiğinde korelasyon = -1 olur. Bu, eğim büyük olsa da küçük olsa da korelasyonun -1'e eşit olabileceği anlamına geliyor.

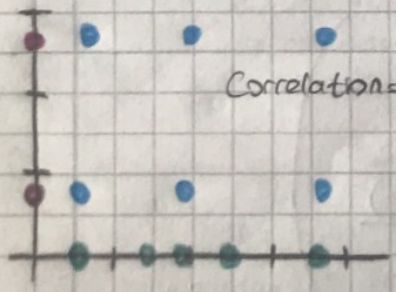
---

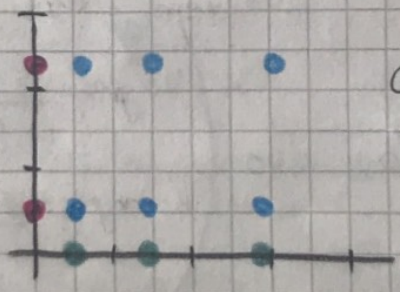
* Şimdiye kadar, eğer line'ın eğrisi pozitifse en güçlü ilişkinin correlation = 1 olacağını, eğer line'ın eğrisi negatifse en güçlü ilişkinin correlation = -1 olacağını gördük. Her iki durumda da eğer doğru bütün datalarının üzerinden geçmiyorsa korelasyon değerleri 0'a yaklaşır.





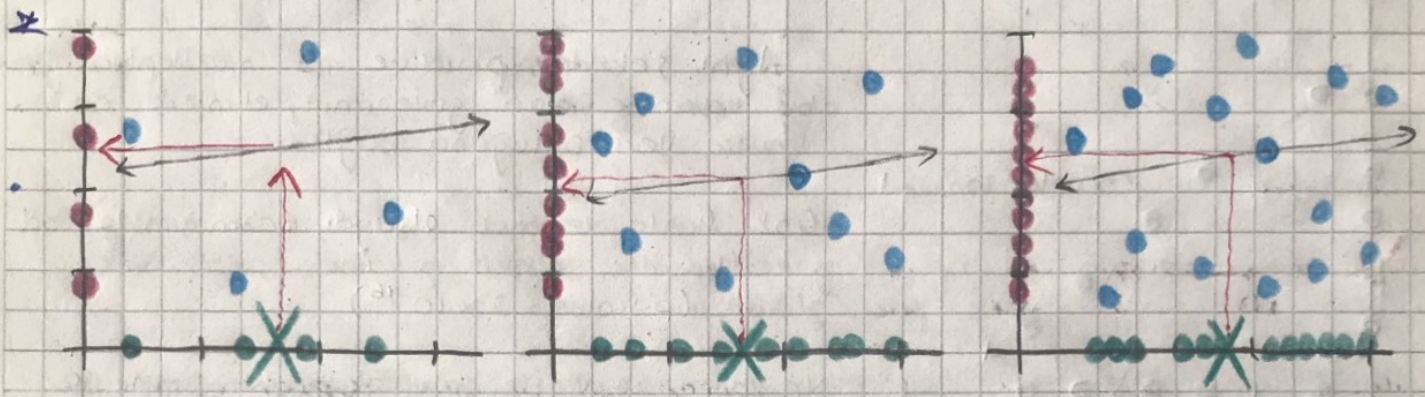
Veriler arasında bir ilişki olmadığında correlation = 0 oluyor.





Correlation = 0 olduğunda x eksenindeki bir değer, y ekseninde hangi değeri beklememiz gerektiği hakkında bize hiçbir şey söylemez. Çünkü x eksenindeki bir değere karşılık y ekseninde birçok yerine diğerini seçmememiz için bir sebep yok.

Korelasyon değeri 0 olmadığı sürece, hala çıkarım yapmak için line'ı kullanabiliriz ancak korelasyon değerleri -1 veya 1'e yaklaştıkça tahminlerimiz daha rafine (hatadan uzak) hale gelir. Daha önce olduğu gibi, çıkarımlarımıza olan güvenimiz veri sayısına ve p-value'ya bağlıdır.



Correlation = 0.3
p-value = 0.8

Verimiz az olduğu için bu trende çok az güveniyoruz.

Correlation = 0.3
p-value = 0.08

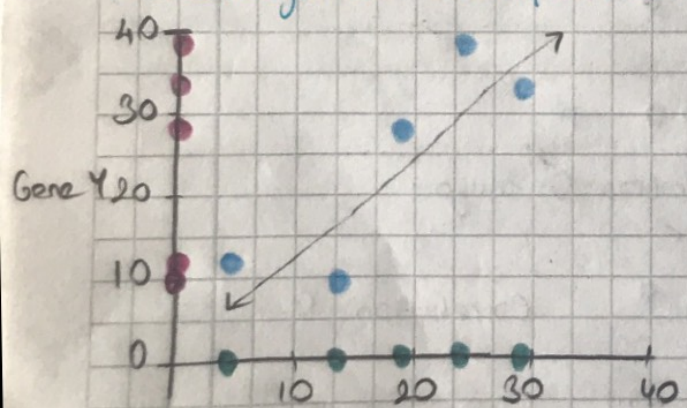
Daha fazla verimiz olduğu için bu trende daha fazla güveniyoruz.

Correlation = 0.3
p-value = 0.008

Bu grafikte çok daha fazla verimiz olduğu için trende çok daha fazla güveniyoruz.

Tüm örneklerde correlation = 0.3 Bu case'te, sample sizeını artırmak korelasyonu artırmadı. Bu da veri eklemenin tahminimizi iyileştirmediği anlamına geliyor. Tek yaptığı tahmine olan güvenimizi artırmaktı. Bu nedenle, tahminlerimiz her üç durumda da muhtemelen kötü olacaktır. Ancak, en çok sonuncu verilerden gelen kötü tahmine güveneceğiz. Başka bir deyişle, çok fazla veriye sahip olduğunuz ve tahminimize çok güvendiğimiz için korelasyon değeri küçükse tahminimiz yine de kötü olacaktır.

Korelasyonun Hesaplaması:

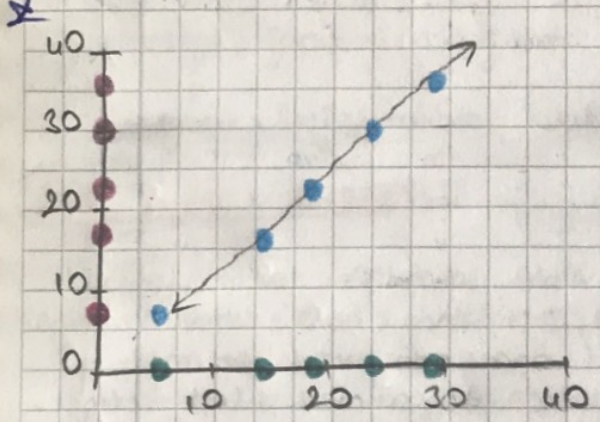


$$\text{Correlation} = \frac{\text{Covariance(Gene X, Gene Y)}}{\sqrt{\text{Variance(Gene X)}}\ \sqrt{\text{Variance(Gene Y)}}}$$

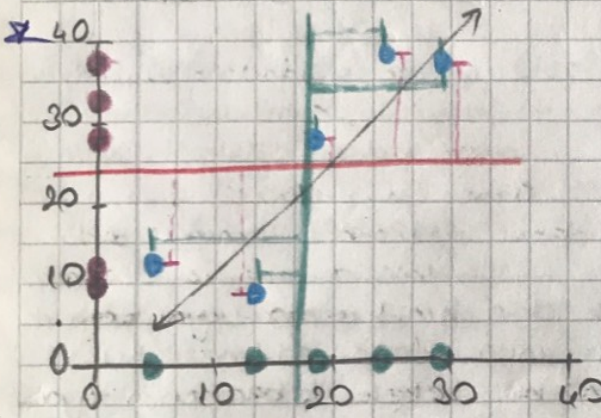
$$\text{Varyans} = \frac{\Sigma (x-\bar{x})^2}{n-1}$$

$$\text{Kovaryans} = \frac{\Sigma (x-\bar{x})(y-\bar{y})}{n-1}$$

Pay'daki covariance değeri $-\infty$ ile ∞ arasında herhangi bir değer olabilir. Ki bu değer ilişkinin pozitif mi negatif mi olduğuna, verinin ortalama etrafında ne kadar uzakta dağıldığına ve verinin scale'ine bağlıdır. Böylece, korelasyonu hesapladığımızda, payda, kovaryansı -1'den 1'e kadar bir sayı olacak şekilde sıkıştırır. Diğer bir deyişle, payda, verilerin ölçeğinin korelasyon değerini etkilemesini söğler ve bu, korelasyonların yorumlanmasını çok daha kolay hale getirir.



Verilerin tümü pozitif veya negatif eğimli düz bir çizgiye düştüğünde, kovaryans ve varyans terimlerinin karekköklerinin çarpımı aynıdır ve bölme eğime bağlı olarak bize 1 veya -1 verir.



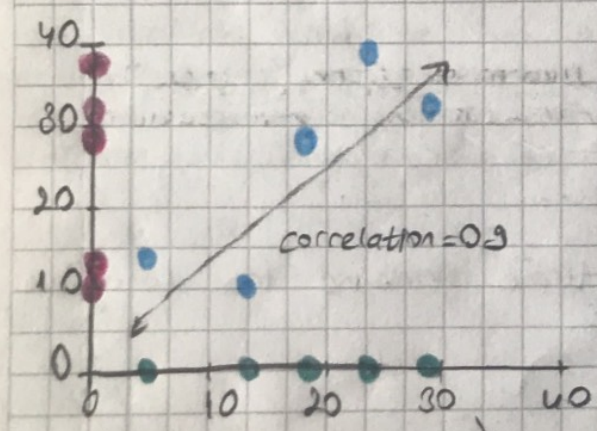
Veriler pozitif veya negatif eğimli düz bir çizgiye düşmediğinde, kovaryans verilerdeki varyansın daha azını açıklar ve korelasyon 0'a daha yakındır.

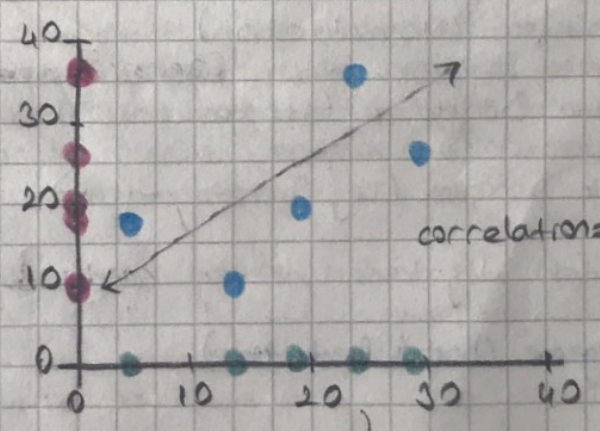
Bu veri ram konuşacak olursak;

$$\text{Correlation} = \frac{116}{\sqrt{101.8}\sqrt{160.3}} = 0.9$$

correlation = 0.9 → bu ilişkiye olan güvenimizi p-value ile ölçebiliriz. p-value ne kadar küçük olursa, yaptığımız tahminlere o kadar güvenebiliriz. Bu case'te p-value = 0.03. Bu, rastgele verilerin benzer şekilde veya daha güçlü bir ilişki üretme olasılığının %3 olduğu anlamına gelir.

Korelasyon değerlerinin yorumlanması, kovaryans değerlerinden çok daha kolay olsa da, yorumlanmaları hala çok kolay değildir.





Örneğin, correlation= 0.9 olduğu bu ilişkinin, correlation=0.64 olduğu bu ilişkiden tahmin yapmakta iki kat daha iyi olduğu açık değildir. R^2 bu sorunu çözüyor.