The correlation matrix is defined thus:

Let $X = [x_1, x_2, \ldots, x_n]$ be the $m \times n$ data matrix: m observations, n variables.

Define $X_b = [\frac{(x_1 - \mu_1 e)}{s_1}, \frac{(x_2 - \mu_2 e)}{s_2}, \frac{(x_3 - \mu_3 e)}{s_3}, \ldots]$ as the matrix of normalized data, with μ_1 being mean for the variable 1, μ_2 the mean for variable 2, etc., and s_1 the standard deviation of variable 1, etc., and e is a vector of all 1s.

The correlation matrix is then

$$C = X_b' X_b$$

A matrix A is positive semi-definite if there is no vector z such that $z' A z < 0$.

→ A Symmetric $n \times n$ real matrix (M) is said to be +ve definite if $z^T M z > 0$ (i.e strictly +ve) for non-zero vector z

Suppose C is not positive definite. Then there exists a vector w such that $w' C w < 0$.

However $(w' C w) = (w' X_b' X_b w) = (X_b w)'(X_b w) = z_1^2 + z_2^2 \ldots$, where $z = X_b w$, and thus $w' C w$ is a sum of squares and therefore cannot be less than zero.

$z^T M z \geq 0$ (i.e non-zero) ⇒ +ve semi-definite

So not only the correlation matrix but any matrix U which can be written in the form $V'V$ is positive semi-definite.

Covariance = $E[(x - E(x))(y - E(y))] = \sigma(x, y)$

→ for a sample of vectors $x_i = (x_1, x_2 \ldots x_k)^T$

sample mean = $\frac{1}{n} \sum_{i=1}^{n} x_i$

sample covariance $Q_y = \frac{1}{n} \sum (x_i - \bar{x})(x_i - \bar{x})^T$

$z^T M z = \frac{1}{n} \sum_{i=1}^{n} [(x_i - \bar{x})^T z]^2 \geq 0$

∴ M is always +ve semi-definite

→ Q:- Why do we need covariance matrix to be positive semi-definite?

→ Symmetric

A - square matrix

$A = A^T$ ⇒ symmetric (by default it is a square matrix)

Consider three variables, X, Y and $Z = X + Y$. Their covariance matrix, M, is not positive definite, since there's a vector z $(= (1, 1, -1)')$ for which $z' M z$ is not positive.

Population covariance matrices are positive semi-definite.

↳ data pulled from a specific distribution

→ The ... matrix can be singular.

+ve definite ⇒ Symmetric & non-singular.

Single Value Decompositions

SVD of matrix A is factorization of A into the product of 3 matrices $A = U \Sigma V^T$, where columns of U & V are orthonormal & Σ is positive diagonal matrix.

→ can be done for both rectangular & square matrices.

$$A = \begin{bmatrix} u_1 & u_2 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \end{bmatrix}$$

u_1, u_2 → Single vectors

$$A = U \Sigma V^T$$

$$A^T A = (V \Sigma^T U^T)(U \Sigma V^T) = V(\Sigma^T \Sigma) V^T$$

V → Eigen vectors

$(\Sigma^T \Sigma)$ → has Eigen values (diagonal matrix) and +ve.

A → was rectangle / square

$A^T A$ → positive semi-definite (Symmetric)

Eigen values (λ) for $A^T A$ ⇒ is σ^2 for A.

→ U & V are orthonormal
i.e row & column vectors are unit vectors.
|orthogonal matrix| = ±1
→ prop: $A^T A = A A^T = I$, $A^T = A^{-1}$

Contribution: $AA^T = U \Sigma V^T V \Sigma^T U^T = U(\Sigma \Sigma^T) U^T$, where $V^T V = I$

Ex:

$$A = \begin{bmatrix} 2 & 2 \\ -1 & -1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} \sqrt{10} & \\ & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

with $\frac{1}{\sqrt{5}}$, Σ, $\frac{1}{\sqrt{2}}$

$(2, -1)$ → 1st left vector; $(1, 1)$ → 1st Right vector

$u_1 \sigma_1 v_1^T$ → most information

↳ principle Component (PCA)

multi-collinearity, or highly

vif

https://stats.stackexchange.com/questions/86269/what-is-the-effect-of-having-correlated-predictors-in-a-multiple-regression-model?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa

Variance Inflation factor

→ Inflation in the variance of regression coefficients due to correlation.

→ collinearity → is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from others with substantial degree of accuracy.

correlation — two variables vary together, if one changes the other changes, but it does not mean collinearity or that one can Explain the other.

→ It is possible to have very low correlation among the variables but perfect collinearity.

Ex - 11 predictors — 10 Independent
— 1 is sum of other 10

Here correlations will be about 0.1 but collinearity is perfect.

→ Singular Correlation / Covariance matrices situations

① Number of variables is Equal or greater than the number of cases.

② two or more variable sum up to constant.

③ 2 variables are identical or differ merely in mean (level) or variance (Scale).

→ Variance Inflation Factor = is ratio of variance in a model with multiple predictors divided by variance of a model with one predictor alone.

→ quantifies severity of multicollinearity

→ provides an index that measures how much the variance of Estimated coefficients is increased by collinearity.

→ High variance in weight causes model to be unstable.

→ VIF = 1 ⇒ no multicollinearity

VIF > 1 ⇒ slight "

5 < VIF < 10 ⇒ high correlation maybe problematic

VIF > 10 ⇒ regression coefficients are poorly estimated

→ Multicollinearity (MC) ⇒ Standard Error ↑ Coefficients ↑

due to overinflated Std. Errors, MC makes some variables statistically insignificant, when they should be significant.

Soln: ① Remove highly correlated predictors

② PLS (partial least squares) & PCA.

→ [low p-value ⇒ statistically significant, but if VIF is high don't trust p-value.]

{Note: p-value = 0 (low) ⇒ Reject Null hypothesis (i.e coefficient = 0)}

MAP – Maximum a Posteriori

→ Seeks the most probable value for parameter w (ith parameter in Regression) under the posterior.

→ SVD + QR factorization can also help in identifying Collinear predictors

To Study

Standard Error

Std. Dev of mean of samples.

p-value

Relation between Std. Error Vs P-value.

→ Correlation between X and Residual

https://s... /questio... regress... orrelate... ariab?u... m_source=google_rich_qa&utm_campaign=google_rich_qa

→ $Corr(X, \hat{\varepsilon}) = 0$ always

→ $Corr(X, \varepsilon) = 0$ if $\hat{\beta} = \beta$ (ε: true Error; β: true β)

NULL SPACE

→ Let A be an m×n matrix. The Null space is the set of solutions to the homogenous system $A\bar{x} = \bar{0}$

Nullspace of A ← $N(A) = \{ \bar{x} \in \mathbb{R}^n : A\bar{x} = \bar{0} \}$

Linear Vs Non-Linear Regression

→ Linear regression can produce curved lines and nonlinear regression is not named for its curved lines.

Non-Linear ⇒ Non linear combination of model parameters and depends on one or more independent variables.

$y \sim f(x, \beta)$

$f(x, \beta) = \frac{\beta_1 x}{\beta_2 + x}$ → non-linear

$f(x, \beta) = \beta_1 x_1 + \beta_2 x_2^2$ → linear

Non-linear, cause this $f(x,\beta)$ cannot be expressed as a linear combination of β_1 & β_2.

Unbiased Estimator

If X_i are independent Bernoulli r.v with unknown parameter p, then the probability mass function of each X_i is:

$f(x_i; p) = p^{x_i} (1-p)^{1-x_i}$

likelihood ← $L(p) = \prod_{i=1}^{n} f(x_i; p) = p^{\sum x_i} (1-p)^{n - \sum x_i}$

→ apply log & maximize ⇒ $\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$ → ML Estimator of p

Unbiased - If the following holds: $E\left[u(X_1, X_2, \cdots X_n)\right] = \theta$

then the statistic $u(X_1, X_2 \cdots X_n)$ is an unbiased Estimator of the parameter θ. Otherwise, $u(X_1, X_2 \cdots X_n)$ is a biased Estimator of θ.

Q Is the MLE of p an unbiased Estimator of p?

$\bar{X}$ is $\hat{p}$

$$E[\hat{p}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}(np) = p$$

$\Rightarrow$ $\hat{p}$ an unbiased Estimator of p.

- We've shown that, under the Gaussian assumption $y \sim N(Xw, \sigma^2 I)$,

linear regression $\Rightarrow$ $\mathbb{E}[w_{\text{ML}}] = w, \quad \text{Var}[w_{\text{ML}}] = \sigma^2 (X^T X)^{-1}$

- When there are very large values in $\sigma^2 (X^T X)^{-1}$, the values of w_{ML} are very sensitive to the measured data y (more analysis later).

- This is bad if we want to analyze and predict using w_{ML}.

→ Using Ridge Regression we can answer the w_{ML} sensitivity issue.

→ $P(x/\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma)}} \left(\exp\left(-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right) \right)$

$E[x] = \int_{\mathbb{R}^d} x\, p(x/\mu, \Sigma)\, dx = \mu$