LINEAR MODELS

- **P-value:** if p≤0.05 ➔ Reject H_0 (RH). // if p>0.05 ➔ Do not reject H_0 (NRH).
- **R^2**: always between 0 and 1. Percentage of variance explained.
- **F-statistic p-value (< 0.05):** Tests overall model significance.
- **Multicollinearity: vif(model)> 4 ➔** take out the last significant variable, and test it again.
- **Influential observations:** Strange value of Y conditioned to X. Stays far away from the rest.
  - ➔ **Priori:** Leverage (X'X) ➔ Don't take them out of the model. Not necessarily affecting Y. `influencePlot(model)`

    `hatvalues(model)` **threshold** ➔ $R > \frac{2*p}{n}$ p ➔ number of parameters n ➔ number of observations .

    `Z-score > |3|` (`rstudent()`)
  - ➔ **Posteriori:** Cook's distance `cooks.distance(model)` **threshold** ➔ Chatterjee and Hadi $D_i > 4/(n-p)\ D_i > 0.5$
- **Model selection:** The lower AIC / BIC **AIC(model,k=log(nrow(data.frame)))** the better the model is.
  - ➔ **Forward ➔** null and increasing **(specify in R).**
  - ➔ **Backward ➔** complete and remove.

ANCOVA

- **Combinació** de: *Y* dependent (numèrica), *Factors* (categòriques) i *Covariavbles* (numèriques).
- **Objectiu:** veure si el factor segueix sent rellevant controlant per la covariable.
- **Reparametrization:**
  - ➔ **Baseline:** One category is set as a reference level. The model estimates relative to this baseline.

    Restriction $\alpha_1 = 0$, $nº\ of\ categories - 1$
  - ➔ **Zero-sum:** restriction $\sum_{i=1}^{I} \alpha_i = 0$ ($o\ \alpha_I = -\sum_{i=1}^{I-1} \alpha_i$), $\alpha_1 + \alpha_2 + \alpha_3 = 0$

Model	Fórmula	Interpretació
M0 – Null Model	$Y_{ijk} = \mu + \varepsilon_{ijk}$	Només la mitjana global. Sense factors ni covariables.
M1 – Full Model (Two-Way amb interacció)	$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$	ANOVA de dos factors A i B amb **interacció** entre ells.
M2 – Additive Model (Two-Way sense interacció)	$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$	ANOVA de dos factors A i B **sense interacció**, només efectes principals.
M3 – One-Way A only	$Y_{ik} = \mu + \alpha_i + \varepsilon_{ik}$	ANOVA simple (One-Way) pel **factor A.**
M4 – One-Way B only	$Y_{jk} = \mu + \beta_j + \varepsilon_{jk}$	ANOVA simple (One-Way) pel **factor B.**
ANCOVA amb interacció	$Y_{ik} = \mu + \alpha_i + (\eta + \theta_i) \cdot x_{ik} + \varepsilon_{ik}$	ANCOVA amb **pendent diferent per grup** (interacció entre X i el factor).
ANCOVA additiu	$Y_{ik} = \mu + \alpha_i + \eta \cdot x_{ik} + \varepsilon_{ik}$	ANCOVA amb **mateixa pendent per a tots els grups.**
Regressió simple	$Y_{ik} = \mu + \eta \cdot x_{ik} + \varepsilon_{ik}$	Només covariable X, sense tenir en compte cap grup.
ANOVA One-Way (sense X)	$Y_{ik} = \mu + \alpha_i + \varepsilon_{ik}$	Només factor A, com a comparació de mitjanes entre grups.

anova(m0,m4) ➔ Only Nested models. The lower the residual sum of squares (RSS), the better the model is. P-value.

BINARY

- Variables amb **resposta binària** (0/1) ➔ predicció d'una probabilitat [0, 1]
  - ➔ **Distribució Bernoulli ➔ Disgregades:** cada fila un individu. Cada observació 0 o 1.
  - ➔ **Distribució Binomial ➔ Agregades:** Agrupació d'individus que tenen les mateixes variables explicatives. Probabilitat d'obtenir K èxits en M intents.
- **Odds:** quentes vegades és més probable que passi? $odds = \frac{\pi\ (probabilitat\ que\ passi)}{1-\pi\ (probabilitat\ de\ que\ NO\ passi)} = e^{\beta} (per\ un\ coeficient,\ es\ multipliquen\ els\ odds)$ odds = 1 ➔ p = 0.5, odds > 1 ➔ probable, odds < 1 ➔ improbable.
- **Logit Link (Y):** resultat de la combinació lineal dels regressors. β indica efecte sobre el logit. Després transformem a probabilitat [0, 1].

  $\eta = log(\frac{\pi}{1-\pi})$ ➔ (g) ➔ de probabilitat [0, 1] a un valor real $(-\infty, \infty)$ (logaritme dels odds)

  $\pi = \frac{exp(\eta)}{1+exp(\eta)}$ ➔ (g-1) de valor real $(-\infty, \infty)$ a probabilitat [0, 1]
- **Probit link:** Uses the inverse of the standard of normal distribution. Transforma π segons distribució normal estàndard.

  $\Phi(\eta)$ // $\eta = g_2(\pi) = \Phi^{-1}(\pi)$ // $\pi_2(\eta) = g_2^{-1}(n) = \Phi(n)$
- **Values of η and π :**
  - ➔ If the value is 0 on the functions ➔ probability π = 0.50 because both functions are centered.
  - ➔ Higher values than 0 ➔ higher probabilities than 0.50
  - ➔ Lower values than 0 ➔ lower probabilities than 0.50.
- **Unnested models ➔** AIC is used to compare. The lower the AIC, the better the model is.
- **Nested models ➔** We use deviance to compare. The lower the deviance, the better the model is.
  - ➔ **Null deviance** is the deviance associated with the **null model**, with no regressors, only the intercept.
  - ➔ **Residual deviance** is the deviance associated with the **full model** (with all predictors). Same or lower than the simpler model.
  - ➔ $\Delta D = D_{null} - D_{full} > 0$ compare whether a particular regressor affects the output. $\Delta D \sim \chi^2$ ➔ p-value
- **Matriu de confusió:** compara el que el model ha predit $(\widehat{Y})$ amb el que realment ha passat (Y) **threshold** ➔ 0.5. TP (a), FP (b), FN (c), TN (d).
- **Evaluation metrics:**
  - ➔ **Accuracy** = (TP + TN) / Total ➔ Percentatge total d'encerts
  - ➔ **Precision** = TP / (TP + FP) ➔ How many predicted positives are actually positive
  - ➔ **Recall (Sensitivity)** = TP / (TP + FN) ➔ How many actual positives were caught
  - ➔ **Specificity** = TN/(TN + FP) ➔ How many actual negatives were caught
- **ROC:** sensibilitat vs 1–especificitat; **AUC** = àrea sota la corba. AUC = 1 perfecte, AUC = 0.5 aleatori, AUC < 0.5 pitjor que aleatori.
- **Linear model? Or do we search for other model options?**
  - ➔ **resiudalPlots(m) ➔** no linearitat / punts infuents
  - ➔ **marginalModelPlots(m) ➔** importància marginal del predictor
  - ➔ **avPlots(m) ➔** contribució real de cada variable (recta amb pendent millor contribució)
  - ➔ **crPlots(m) ➔** no linearitat més suau i detallada que els residuals

CLUSTERING

- **Agrupa individus** amb alta similitud dins del clúster i gran diferència entre clústers.
- **Mètodes de partició:** *Ex: K-means.* Necessites fixar prèviament el nombre de grups (K) → input. Assigna individus als clústers segons la distància a un centre i actualitza. Es ràpid i senzill però sensible a outliers i assumeix formes esfèriques.
- **Mètodes jeràrquics (Hierarchical):** no cal fixar K, genera dendograma → ascendents (fusió) o descendents (divisió) → tallar a certa alçada = obtenció de grups. Les fulles del dendograma son els individus.
  - **Criteris d'agregació**: **Ward:** agrupa minimitzant la pèrdua d'inèrcia (la variància que expliquen) intra-grup (molt usat). Equilibrats.
  - **Distància per variables mixtes: Gower** → combina distàncies normalitzades segons tipus de variable.

PROFILING

- Tècnica per **descriure i entendre els grups** trobats mitjançant **variables qualitatives i quantitatives**.
- **Passos profiling:**
  - Trobar variables significatives: p-value < 0.05 → variable significativa
    - **Variables qualitatives**: χ^2 independence test / Multiple box plot
      H_0: X,Y are independent
      H_1: X,Y are associated
      LeBart test: Detecta si una modalitat específica és significativa en un grup concret.
      H_0: $\mu_k = \mu$ (the mean of my group (k) is the same of the hole) → pval < 0.05 → presència/absència rellevant
    - **Variables quantitatives**: *ANOVA test* (diferència entre mitjanes entre grups) / Multiple bar plot / Mean plot
  - Descriure les diferències entre els grups
  - Describe the groups:
    - Fer una **taula resum** de les variables significatives per grup
    - Frases clau per interpretar
    - Assignar **etiqueta significativa** a cada grup
- **Aplicacions profiling:**
  - **Màrqueting**: segmentar clients, crear perfils de consumidor
  - **Polítiques públiques**: detectar col·lectius vulnerables
  - **Prevenció de riscos**: identificar grups amb alt risc (ex: impagaments)
- El profiling **NO sempre requereix clustering**. Qualsevol **variable categòrica pot definir grups** i aplicar el profiling a partir d'aquí.

PCA (Principal Component Analysis)

- Redueix la dimensionalitat de dades numèriques mantenint la màxima variància possible.
- Transforma variables originals en noves variables (components principals, PC), ortogonals i no correlacionades. PC1 explica la màxima variància, PC2 la següent, etc.
- Cada component és una **combinació lineal** de les variables originals → **càrregues factorials (pc1$rotation)**. Com més càrrega factorial (en valor absolut) més pes en el component, més explica. Each component has a little portion of all the original variables, but not all the components contain the same contribution of each component. **ONLY NUMERICAL VARIABLES.**
- **Passos previs:**
  - **Centrar dades**: restar la mitjana (x - μ).
  - **Normalitzar** (si unitats diferents): dividir per la desviació estàndard.
- sdev^2 = variància (inèrcia) explicada per component (**eigenvalue**).
- Scree plot → decideix quants components conservar (fins a **80% variància acumulada**).
- **Biplot (gràfic de variables)**: fletxes = variables originals:
  - Direcció → cap on augmenta la variable.
  - Longitud → importància (variància explicada).
  - Fletxes juntes → correlació positiva.
  - Fletxes oposades → correlació negativa.
  - Fletxes amb un angle gran → no correlacionades.
- **Modalitats categòriques**: es projecten com a **centres de gravetat** → relacions entre modalitats i variables.
- Each dimension has different contributions, not all the plots have the same quantity of information.
- Les noves variables (PCs) es calculen combinant les **dades centrades** amb els **vectors propis** (**eigenvectors**). PC1 = Xs(1) · (−0.707) + Xs(2) · 0.707
- **Usos del PCA:**
  - Visualization of multidimensional data
  - Associative method of variables
    - Relations between variables (numeric and categorical)
  - Preprocessing data method
    - Latent variables (Variables that can not be measured: freedom, happiness, richness)
    - Reduccion of dimensionality
- **Conclusions:**
  - Relacions entre variables → fletxes juntes.
  - Associació entre modalitats → pròximes a fletxes.
  - Detecció d'individus extrems multivariants/Multivariate outliers (**Multivariate:** an observation that looks normal in individual variables but is unusual when considering multiple variables together).
  - Descobrir variables latents (ex: llibertat, happiness → combinacions de las variables en una dimensió).
  - Detecció de les variables amb més contribució a cada eix.