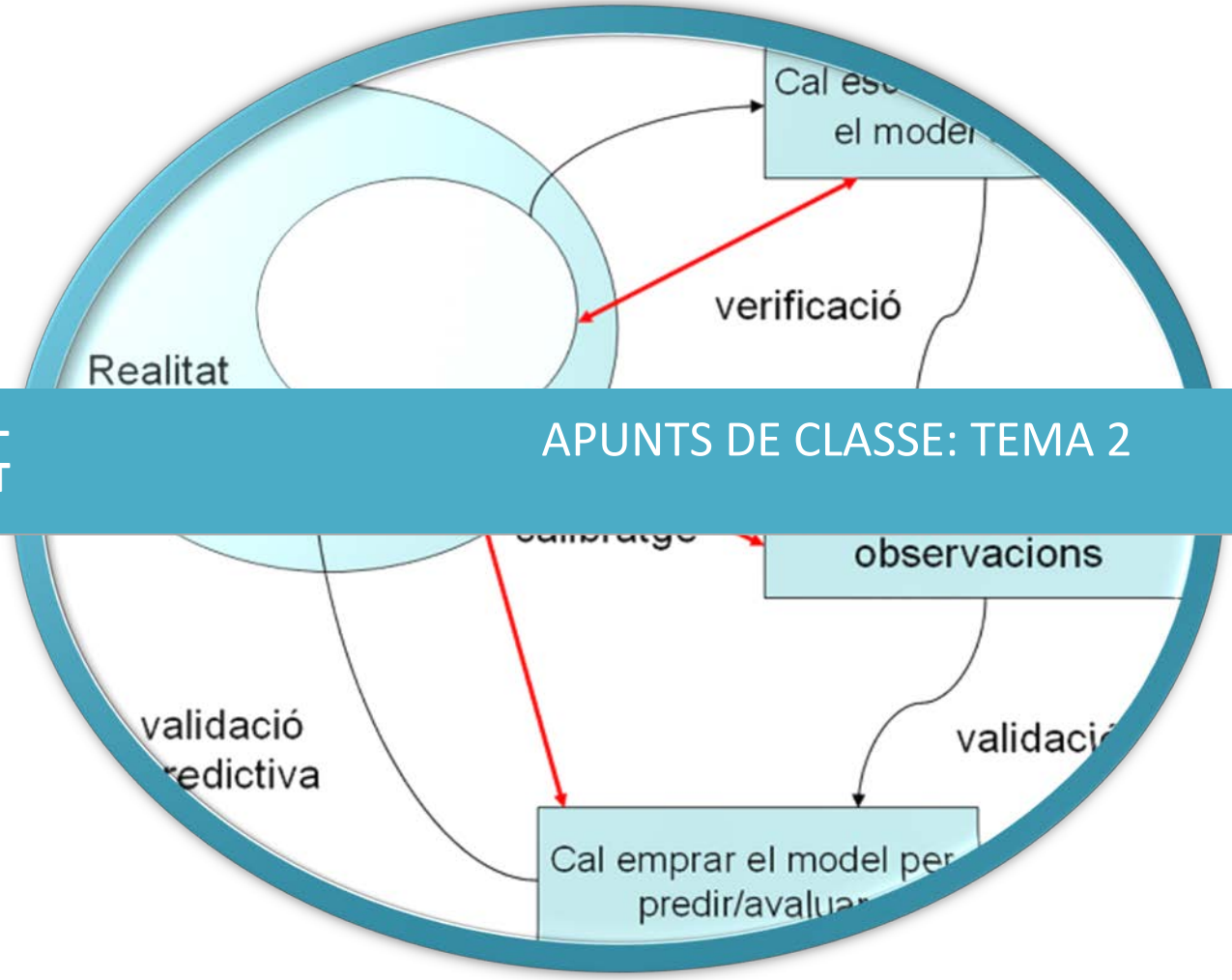
Departament d'Estadística
i Investigació Operativa
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Curs 2017-2018



MODEL LINEAL GENERALITZAT

APUNTS DE CLASSE: TEMA 2

Grau d'Estadística |UB - UPC | Lídia Montero i Josep Anton Sánchez

TABLA DE CONTENIDOS

TEMA 2: RESPUESTA NORMAL. MODELO LINEAL

Objetivo:

Análisis de las relaciones entre variables explicativas (factores y covariables) y la variable respuesta, con distribución condicionada de tipo normal.

Para variables respuesta de tipo continuo, es habitual considerar la distribución Normal, que es un caso particular de Modelo Lineal Generalizado. En los diferentes ámbitos de aplicación de la estadística es fácil encontrar ejemplos donde la variable respuesta de interés representa una cuantificación:

- valores de una análitica en medicina
- dimensiones de una pieza en ingeniería
- indicadores en econometría

EL MODELO LINEAL COMO CASO PARTICULAR DEL MLGZ

Caso particular de MLGz:

$$f_Y(y,\theta,\phi)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)=\exp\left(\frac{y\mu-\mu^2/2}{\sigma^2}-\frac{1}{2}\left(\frac{y^2}{\sigma^2}+\log\left(2\pi\sigma^2\right)\right)\right)$$

donde $a(\phi)=\phi=\sigma^2$, $b(\theta)=\frac{\theta^2}{2}=\frac{\mu^2}{2}$ (es decir, $\theta=\mu$) y $c(y,\phi)=-\frac{1}{2}\left(\frac{y^2}{\phi}+\log(2\pi\phi)\right)$.

$$\ell(\theta,\phi,y)=\log f_Y(y,\theta,\phi)=\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)=\frac{y\theta-\theta^2/2}{\phi}-\frac{1}{2}\left(\frac{y^2}{\phi}+\log(2\pi\phi)\right)$$

EL MODELO LINEAL COMO CASO PARTICULAR DEL MLGZ

Modelo Lineal:

Una muestra formada por n unidades experimentales

Variable respuesta Y medida sobre las diferentes unidades experimentales $Y=(Y_1,\dots Y_n)$

Variables explicativas $X_1,\dots,X_p$ medidas sin error.

- Cada variable da lugar a un vector de medidas realizadas sobre cada unidad experimental: $Xi=(X_{i1},\dots X_{in})$
- Las observaciones de las variables explicativas en el individuo j-ésimo se representan como $(X_{1j},\dots,X_{pj})$

Premisas

- Las observaciones de la variable respuesta son mutuamente independientes.

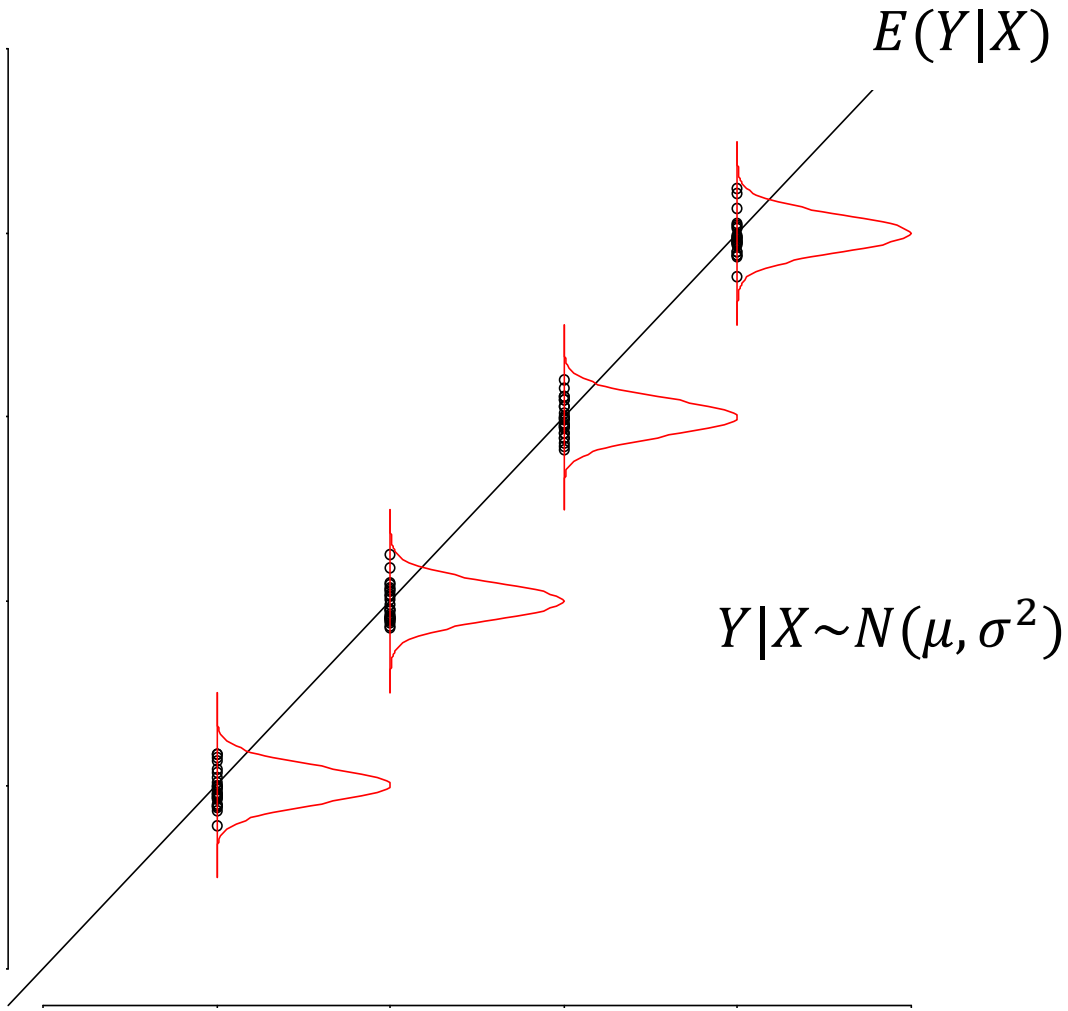
- La distribución condicionada de la variable respuesta a las variables explicativas sigue una distribución Normal:

$$Y|X \sim N(\mu, \sigma^2)$$

Esto último implica que la esperanza condicionada es el parámetro natural y la varianza condicionada no depende del valor esperado:

$$E(Y|X) = \mu \qquad V(Y|X) = \sigma^2$$

EL MODELO LINEAL COMO CASO PARTICULAR DEL MLGZ



EL MODELO LINEAL COMO CASO PARTICULAR DEL MLGZ

En el caso de la distribución normal es posible obtener la distribución de los errores:

$$\varepsilon = Y - X\beta \sim N(0, \sigma^2)$$

Así pues, la ecuación del modelo teórico para el individuo j-ésimo será:

$$y_j = \beta_0 + \beta_1 x_{1j} + \cdots + \beta_p x_{pj} + \varepsilon_j \qquad \varepsilon_j \sim N(0, \sigma^2)$$

ESTIMACIÓN DE PARÁMETROS

El modelo ajustado se obtiene resolviendo las Ecuaciones Normales, asociadas a la estimación basada en Mínimos Cuadrados Ordinarios (OLS: Ordinary Least Squares) :

$$\hat{\beta} = min_{\beta} S(\beta) = min_{\beta} (Y - X\beta)'(Y - X\beta)$$

$$\hat{\mu} = X\hat{\beta}$$

Condiciones de primer orden:

$$\nabla_{\beta} S(\beta) = 0 \leftrightarrow \frac{\partial S(\beta)}{\partial \beta} = 2X'X\beta - 2X'Y = 0$$

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Si la matriz de diseño tiene rango máximo (rangX=p) la solución es única.

Si X no tiene rango máximo, existen infinitas soluciones que dan el mismo vector de predicciones.

ESTIMACIÓN DE PARÁMETROS

Las condiciones de segundo orden implican que:

$$\nabla^2_\beta S(\beta) > 0 \leftrightarrow \frac{\partial^2 S(\beta)}{\partial \beta_i \partial \beta_j} = 2X'X > 0 \qquad i,j = 1,\dots,p$$

Puesto que el rango de X'X es el mismo de X, las condiciones para que las ecuaciones normales tengan solución única son:

- Al menos hay tantas observaciones como variables (n≥p)
- Las columnas de X deben ser linealmente independientes

Si existe multicolinealidad perfecta (las columnas de X son linealmente dependientes) la matriz X'X es singular y no hay solución única de las ecuaciones normales.

Así pues, el modelo ajustado es:

$$\hat{y}_j = \hat{\mu}_j = \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \cdots + \hat{\beta}_p x_{pj}$$

Y los residuos del modelo ajustado son:

$$e_j = y_j - \hat{y}_j$$

ESTIMACIÓN DE PARÁMETROS

Una vez ajustado el modelo, los estadísticos Deviancia y Deviancia Escalada son:

$$D(y,\mu) = (2l(y,\phi,y) - 2l(\hat{\mu},\phi,y))\phi = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$$

$$D'(y,\mu) = \frac{D(y,\mu)}{\phi} = \sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\phi}$$

Por tanto, en el modelo Normal la Deviancia corresponde a la Suma de Cuadrados Residual (RSS)

ESTIMACIÓN DE PARÁMETROS

Propiedades de los estimadores:

- El estimador OLS $\hat{\beta} = (X'X)^{-1}X'Y$ es el estimador insesgado de mínima varianza y coincide con el estimador de máxima verosimilitud
- El estimador eficiente de σ^2 es $S^2 = \sum_{i=1}^{n} \frac{e_i^2}{n-p} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{n-p} = \frac{RSS}{n-p}$ donde RSS denota la suma de cuadrados residuales
- Los estadísticos $\hat{\beta}$ y S^2 son independientes
- Las distribuciones en el muestreo son:
  - $\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$,
  - $\frac{RSS}{\sigma^2} = \frac{(n-p)S^2}{\sigma^2} \sim \chi^2_{n-p}$
- La distribución del estimador se puede expresar mediante una ley que no dependa de los parámetros de la siguiente manera:

$$(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)/\sigma^2 \sim \chi^2_p$$

VARIABLES EXPLICATIVAS NUMÉRICAS. REGRESIÓN MÚLTIPLE

Los modelos de regresión simple (un predictor) y múltiple (varios predictores) corresponden a la situación en que todas las variables predictoras son numéricas:

Modelo de regresión lineal simple:

$$y_j = \beta_0 + \beta_1 x_{1j} + \varepsilon_j \qquad \varepsilon_j \sim N(0, \sigma^2)$$

Modelo de regresión lineal múltiple:

$$y_j = \beta_0 + \beta_1 x_{1j} + \cdots + \beta_p x_{pj} + \varepsilon_j \qquad \varepsilon_j \sim N(0, \sigma^2)$$

VARIABLES EXPLICATIVAS NUMÉRICAS. REGRESIÓN MÚLTIPLE

Interpretación de los resultados obtenidos con el ajuste:

La cuantificación del efecto de una variable en la respuesta viene dada por su coeficiente en el modelo.

β_i corresponde al cambio promedio en la variable respuesta al incrementar la variable X_i en una unidad, mientras el resto se mantienen fijas.

Es posible obtener intervalos de confianza para los coeficientes del modelo usando la distribución del estimador:

$$\frac{\hat{\beta}_i - \beta_i}{S_{\hat{\beta}_i}} \sim t_{n-p} \rightarrow \hat{\beta}_i \pm t_{n-p,\alpha/2} S_{\hat{\beta}_i}$$

donde $S_{\hat{\beta}_i} = s\sqrt{diag(X'X)_i^{-1}}$

El test $H_0 : \beta_i = 0$ permite medir la significación estadística de la variable X_i

Test de Wald: $\hat{\beta}_i / S_{\hat{\beta}_i} \sim t_{n-p}$

El coeficiente y la significación tienen en cuenta la presencia o no de otras variables explicativas en el modelo. Se dice que su efecto está ajustado por otras covariables.

VARIABLES EXPLICATIVAS NUMÉRICAS. REGRESIÓN MÚLTIPLE

Medidas de bondad de ajuste:

Coeficiente de Determinación:

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

Expresado en tanto por ciento, corresponde al porcentaje de variabilidad explicado por el modelo lineal.

También se calcula como el cuadrado de la correlación entre el vector de observaciones y el vector de predicciones. Esta correlación se denomina coeficiente de correlación múltiple.

$$R = cor(Y, \hat{Y})$$

Propiedades:

- $|R| \leq 1$ y $|R|=1 \Leftrightarrow$ existe relación lineal perfecta entre la respuesta y los regresores
- $100(1-R^2)$ representa el porcentaje de variabilidad no explicado por el modelo
- Puesto que añadiendo nuevos regresores, el valor de R^2 disminuye, se utiliza el coeficiente de determinación ajustado (R^2-ajustado) para penalizar por la inclusión de regresores no significativos en el modelo:
  - $R^2_{adj} = 1 - \frac{\frac{RSS}{n-p}}{\frac{TSS}{n-1}} = 1 - (1 - R^2)(\frac{n-1}{n-p})$

VARIABLES EXPLICATIVAS NUMÉRICAS. REGRESIÓN MÚLTIPLE

Test de Hipótesis en Regresión Múltiple:

Sea el modelo: $Y = X\beta + \varepsilon \quad \varepsilon \sim N_n(0, \sigma^2 I_n)$

Se desea realizar una prueba de hipótesis sobre una restricción lineal sobre los coeficientes del modelo:

$$H_0: A\beta = c$$

$$H_1: A\beta \neq c$$

donde A es una matriz qxp con q<p

Si H_0 es cierta, entonces:

$$F = \frac{(RSS_H - RSS)/q}{RSS/(n-p)} = \frac{\left(A\hat{\beta} - c\right)'(A(X'X)^{-1}A')(A\hat{\beta} - c)}{qS^2} \approx F_{q,n-q}$$

Donde RSS y RSS_H son las sumas de cuadrados residuales para el modelo completo y sujeto a restricción respectivamente.

VARIABLES EXPLICATIVAS NUMÉRICAS. REGRESIÓN MÚLTIPLE

Tabla del ANOVA de la regresión:

Para contrastar la hipótesis de no significación simultánea de los coeficientes, se construye la tabla del ANOVA de la regresión (test global de regresión)

$$H_0: \beta_1 = 0, \dots, \beta_p = 0$$
$$H_1: \exists i = 1, \dots, p \mid \beta_i \neq 0$$

$$F = \frac{(RSS_H - RSS)/q}{RSS/(n-p)} = \frac{(TSS - RSS)/(p-1)}{RSS/(n-p)} \approx F_{p-1,n-p}$$

VARIABLES EXPLICATIVAS NUMÉRICAS. REGRESIÓN MÚLTIPLE

Tabla ANOVA	Descomposición	Grados de libertad	Varianza	Contraste
Model SS	$MSS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	p-1	$S^2_{exp} = \frac{ESS}{p-1}$	$F = \frac{S^2_{exp}}{S^2}$
Residual SS	$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	n-p	$S^2 = \frac{RSS}{n-p}$	
Total SS	$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$	n-1		

VARIABLES EXPLICATIVAS CATEGÓRICAS. ANOVA

Los modelos ANOVA corresponden a la situación en que todas las variables predictoras son categóricas (factores).

El objetivo consiste en descomponer la variabilidad total en base a los factores incluidos en el modelo. Para el caso del ANOVA de un factor a k niveles:

$$TSS = MSS + RSS$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{i.}-\bar{y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-y_{i.})^2$$

siendo $Y_{i.}$ el promedio para el grupo i-ésimo

Propiedades:

$$\frac{RSS}{\sigma^2} \sim \chi^2_{n-k} \qquad MSE = \frac{RSS}{n-k} \qquad E(MSE) = \sigma^2$$

$$\frac{MSS}{\sigma^2} \sim_{H_0} \chi^2_{k-1} \qquad MSM = \frac{ESS}{k-1} \qquad E_{H_0}(MSM) = \sigma^2$$

Donde la distribución de MSM está condicionada a la hipótesis nula de igualdad entre grupos

VARIABLES EXPLICATIVAS CATEGÓRICAS. ANOVA

Así pues, se puede contrastar la hipótesis de igualdad entre grupos mediante el estadístico F de la tabla del ANOVA:

$$H_0: \mu_1 = \cdots = \mu_k$$
$$H_1: \exists i, j \mid \mu_i \neq \mu_j$$

Si H_0 es cierta, entonces:

$$F = \frac{MSM}{MSE} = \frac{MSS/(k-1)}{RSS/(n-k)} \approx F_{k-1,n-k}$$

Expresión del modelo para el ANOVA de 1 factor a k niveles:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$$

La observación y_{ij} se refiere a la j-ésima observación correspondiente al nivel i-ésimo del factor. La hipótesis nula se puede reformular de la siguiente manera:

$$H_0: \mu_1 = \cdots = \mu_k \quad \Leftrightarrow \quad H_0: \alpha_1 = \cdots = \alpha_k = 0$$

VARIABLES EXPLICATIVAS CATEGÓRICAS. ANOVA

Así, el test de igualdad de medias se puede reformular como un test de significación de coeficientes del modelo lineal.

Para tratar las variables categóricas como variables explicativas, es preciso crear variable artificiales (o dummies) que serán asociadas a contrastes sobre los niveles del factor.

Codificación de una variable categórica a 4 niveles con 4 variables dummies:

Variable Original	VA	VB	VC	VD
A	1	0	0	0
B	0	1	0	0
C	0	0	1	0
D	0	0	0	1

Como en el modelo se incluye el término independiente (Intercept), si se añaden las cuatro variables dummies, la matriz de diseño tiene las columnas linealmente dependientes y, por lo tanto, no existe solución única de las ecuaciones normales.

Solución: puesto que se incluye el término independiente, se considera la inclusión de k-1 variables dummies

VARIABLES EXPLICATIVAS CATEGÓRICAS. ANOVA

En realidad, lo que hacemos es incluir una restricción sobre los parámetros del modelo de forma que la solución de las ecuaciones de estimación sea única.

Restricciones y contrastes habituales:

- Restricción de tipo baseline o categoría de referencia: $\alpha_i = 0$

  El término independiente se interpreta como el valor esperado de la respuesta para la categoría de referencia. Los parámetros del modelo se interpretan como el cambio en la respuesta de pasar de la categoría de referencia a cada uno de los otros niveles.

  en R: "contr.treatment" supone i=1 (primera categoría de referencia)

$\mu = \mu_1$ $\quad$ $\alpha_i = \mu_i - \mu_1$

Variable Original	VB	VC	VD
A	0	0	0
B	1	0	0
C	0	1	0
D	0	0	1

VARIABLES EXPLICATIVAS CATEGÓRICAS. ANOVA

"contr.SAS" supone i=k (última categoría de referencia)

$$\mu = \mu_k \qquad \alpha_i = \mu_i - \mu_k$$

Variable Original	VA	VB	VC
A	1	0	0
B	0	1	0
C	0	0	1
D	0	0	0

VARIABLES EXPLICATIVAS CATEGÓRICAS. ANOVA

- Restricción de tipo suma: $\sum_{i=1}^{k} \alpha_i = 0$

  El término independiente se interpreta como el valor esperado de la respuesta como promedio de los valores obtenidos para cada categoría. Los parámetros del modelo se interpretan como el cambio en la respuesta si pasamos del valor promedio a cada una de las categorías restantes. Para la última categoría, el cambio corresponde a la suma de los coeficientes cambiados de signo, ya que la restricción impone que la suma total sea cero.

en R: "contr.sum"

$$\mu = \text{promedio global} \qquad \alpha_i = \mu_i - \mu$$

Variable Original	V1	V2	V3
A	1	0	0
B	0	1	0
C	0	0	1
D	-1	-1	-1

VARIABLES EXPLICATIVAS CATEGÓRICAS. ANOVA

Una vez definidos los contrastes (codificación de variables dummies) la estimación de los coeficientes del modelo se realiza de la misma manera que en el caso de la regresión lineal, es decir, mediante las ecuaciones normales. Los coeficientes estimados (α_i) corresponden a variables explicativas del modelo lineal, ya que las variables dummies son variables numéricas.

Los estimadores de los parámetros del modelo del ANOVA continúan siendo eficientes (lineales sin sesgo y de mínima varianza), consistentes y asintóticamente normales. La expresión de su varianza es equivalente a la obtenida en el modelo de regresión lineal.

Inferencia sobre el modelo asociado al ANOVA

- Test de Wald: basado en la normalidad asintótica de los coeficientes. Hay que interpretar el resultado en base al significado del coeficiente según el contraste activo.

$$H_0: \alpha_i = 0 \qquad \hat{\alpha}_i / S_{\hat{\alpha}_i} \sim t_{n-p}$$

- Test ANOVA: basado en la descomposición de la suma de cuadrados de la respuesta. Equivale a un test de significación simultáneo para todos los coeficientes de las variables dummies.

$$H_0: \alpha_1 = \cdots = \alpha_k = 0 \qquad F = \frac{MSS/(k-1)}{RSS/(n-k)} \approx F_{k-1,n-k}$$

VARIABLES EXPLICATIVAS CATEGÓRICAS. ANOVA

Expresión del modelo para el ANOVA de 2 factor cruzdos a k y q niveles sin interacción:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$$

La estimación de estos modelos se obtiene generando la matriz de diseño con variables dummies para cada factor:

k=2, q=3 sin réplicas y contraste tipo "treatment"

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix}$$

VARIABLES EXPLICATIVAS CATEGÓRICAS. ANOVA

Si existen réplicas por condición experimental, es posible estimar el modelo con interacción:

$$y_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$$

La interacción tiene kxq niveles. En la matriz de diseño las columnas correspondientes a los niveles de la interacción se calcula en base al producto de las columnas de los niveles de los factores principales que la conforman. Nuevamente es necesario incluir restricciones:

Tipo Baseline ("treatment"):

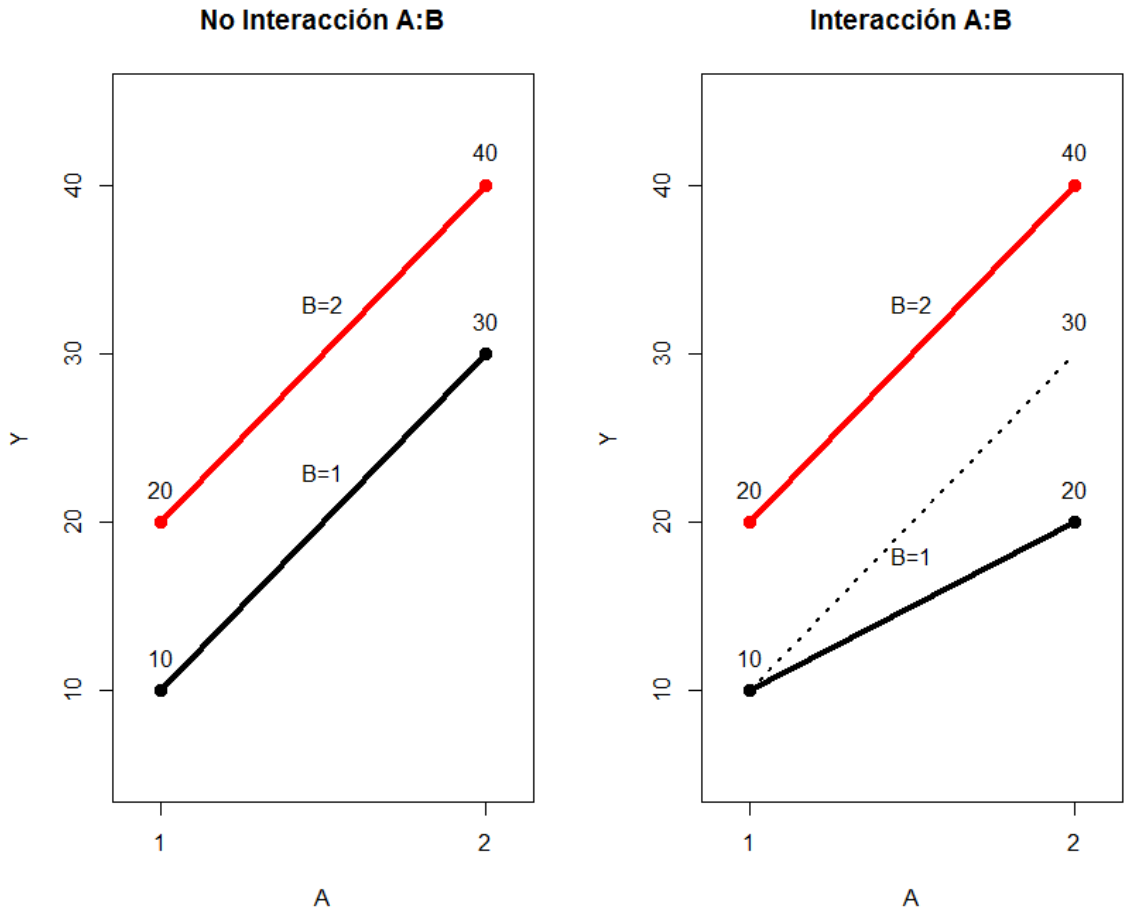
$$\alpha_1 = 0 \qquad \beta_1 = 0 \qquad \alpha\beta_{1j} = 0 \;\; \forall j = 1 \dots q \qquad \alpha\beta_{i1} = 0 \;\; \forall i = 1 \dots k$$

La tabla ANOVA incluye estadísticos de contraste para la significación simultánea de cada efecto principal y de la interacción:

$$H_0: \alpha_1 = \cdots = \alpha_k = 0$$
$$H_0: \beta_1 = \cdots = \beta_q = 0$$
$$H_0: \alpha\beta_{11} = \cdots = \alpha\beta_{kq} = 0$$

VARIABLES EXPLICATIVAS CATEGÓRICAS. ANOVA

Interpretación de la significación de interacciones entre dos factores Ay B:



La combinación A=2 y B=1 da lugar a una respuesta inferior a lo esperado de acuerdo al modelo aditivo.

VARIABLES EXPLICATIVAS CATEGÓRICAS. ANOVA

Los parámetros del modelo permiten calcular el valor de la respuesta en base a las categorías de los factores:

Si k=2, q=2 con interacción y contraste tipo "treatment"

Restricciones:

$$\alpha_1 = 0 \qquad \beta_1 = 0 \qquad \alpha\beta_{1j} = 0 \;\; \forall j = 1 \dots q \qquad \alpha\beta_{i1} = 0 \;\; \forall i = 1 \dots k$$

Niveles de los factores	Valor esperado de la respuesta
i=1 y j=1	$\mu_{ij} = \mu$
i=2 y j=1	$\mu_{ij} = \mu + \alpha_2$
i=1 y j=2	$\mu_{ij} = \mu + \beta_2$
i=2 y j=2	$\mu_{ij} = \mu + \alpha_2 + \beta_2 + \alpha\beta_{22}$

VARIABLES EXPLICATIVAS NUMÉRICAS Y CATEGÓRICAS. ANCOVA

En ocasiones, la comparación de niveles de un factor ha de tener en cuenta la existencia de una covariable numérica que afecta a la respuesta (usualmente de forma lineal). La covariable numérica se suele denominar como variable concomitante y el modelo ha de tener en cuenta que el factor puede implicar diferencias de nivel y/o diferencias en la relación lineal con la covariable.

La técnica del ANCOVA (Análisis de la Covarianza) tiene en cuenta el ajuste por la variable numérica y la categórica, incluyendo la posible interacción entre ambas.

El modelo asociado a la técnica del ANCOVA es:

$$y_{ij} = \mu + \alpha_i + \beta x_{ij} + \alpha\beta_i x_{ij} + \varepsilon_{ij} \quad \varepsilon_{ij} \sim N(0, \sigma^2) \quad i = 1 \dots k, j = 1 \dots n_k$$

VARIABLES EXPLICATIVAS NUMÉRICAS Y CATEGÓRICAS. ANCOVA

La matriz de diseño incluye:

- las columnas de las variables dummies asociadas al factor de acuerdo a los contrastes anteriormente explicados
- una columna con los valores de la covariable numérica, igual que en el caso de la regresión simple
- las columnas correspondientes a la interación entre el factor y las covariables obtenidas como producto de las anteriores

k=2 y contraste tipo "treatment"

$$X = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 3 & 0 \\ 1 & 0 & 4 & 0 \\ 1 & 1 & 3 & 3 \\ 1 & 1 & 2 & 2 \\ 1 & 1 & 5 & 5 \end{pmatrix}$$

Restricciones con contrastes Tipo Baseline ("treatment"):

$$\alpha_1 = 0 \qquad\qquad \alpha\beta_1 = 0$$

VARIABLES EXPLICATIVAS NUMÉRICAS Y CATEGÓRICAS. ANCOVA

La tabla ANOVA incluye estadísticos de contraste para la significación de la covariable, la simultánea del factor y la de la interacción. En general al tratar en el modelo un factor, una covariable y su interacción, la interpretación de la inferencia permite determinar diferentes configuraciones para la variable respuesta.

Si queremos estudiar cómo es la relación lineal entre la covariable y la respuesta en cada subgrupo definido por los niveles del factor, los resultados de los test se pueden intrepretar de la siguiente manera:

- Test de significación del factor:

  Si no se rechaza $H_0: \alpha_1 = \cdots = \alpha_k = 0$ ➔

  "Los términos independientes de las rectas de cada grupo no difieren significativamente"

- Test de significación de la covariable:

  Si no se rechaza $H_0: \beta = 0$ ➔

  "La pendiente de la recta del grupo de referencia no es significativa y por lo tanto no hay relación entre respuesta y covariable en este grupo"

- Test de significación de la interacción:

  Si no se rechaza $H_0: \alpha\beta_1 = \cdots = \alpha\beta_k = 0$ ➔

  "Las pendientes de las rectas de cada grupo no difieren significativamente. Es decir, son rectas paralelas"

VARIABLES EXPLICATIVAS NUMÉRICAS Y CATEGÓRICAS. ANCOVA

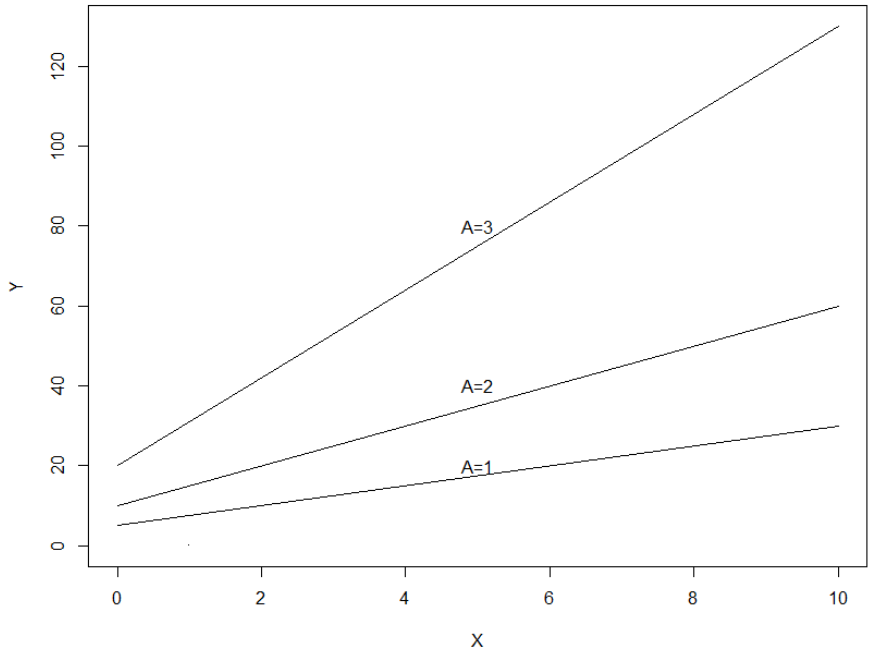
Cálculo de las rectas de regresión por grupos en el caso general:

Niveles de los factores	Recta de regresión
i=1	$y_{1j} = \mu + \beta x_{ij} + \varepsilon_{1j}$
i=2	$y_{2j} = (\mu + \alpha_2) + (\beta + \alpha\beta_2)x_{ij} + \varepsilon_{2j}$
i=3	$y_{3j} = (\mu + \alpha_3) + (\beta + \alpha\beta_3)x_{ij} + \varepsilon_{3j}$
:	:
i=k	$y_{kj} = (\mu + \alpha_k) + (\beta + \alpha\beta_k)x_{kj} + \varepsilon_{kj}$

VARIABLES EXPLICATIVAS NUMÉRICAS Y CATEGÓRICAS. ANCOVA

1) En el ANCOVA, se ajusta el modelo completo y se comprueba si la interacción es significativa:

   Si es significativa ➔ la relación entre la respuesta y la covariable difiere según el grupo

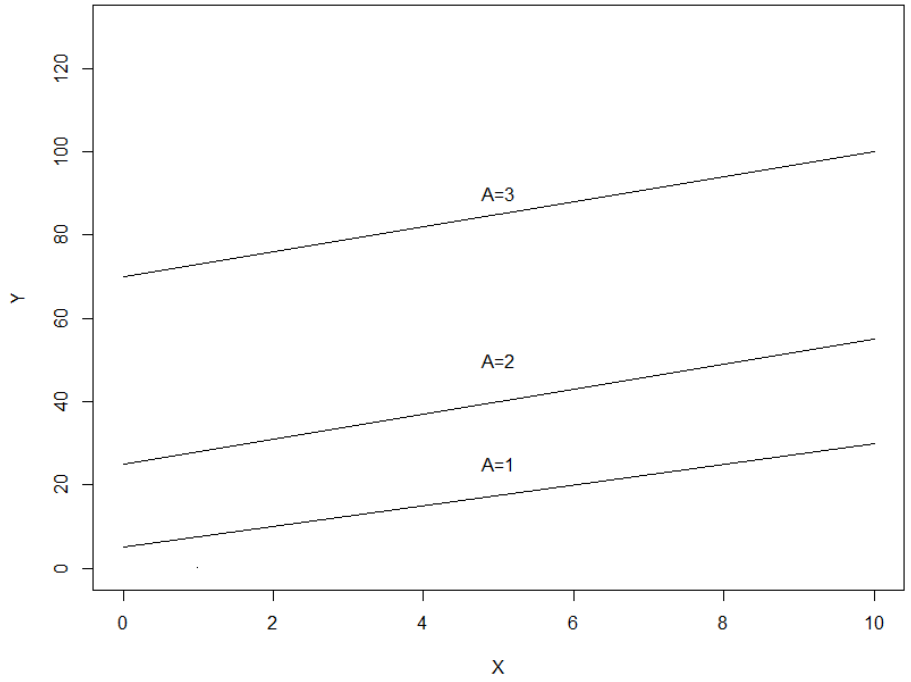


   Si no lo es ➔ las rectas tienen la misma pendiente, ajustamos el modelo sin la interacción y pasamos a verificar la significación de la covariable

VARIABLES EXPLICATIVAS NUMÉRICAS Y CATEGÓRICAS. ANCOVA

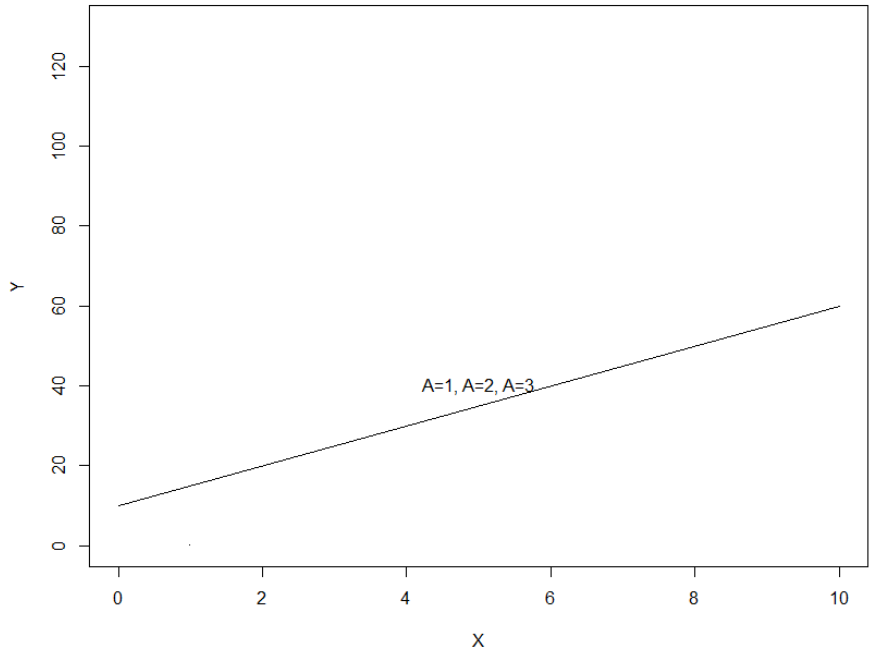
2) Si la covariable es significativa ➔ hay relación lineal entre la respuesta y la covariable y pasamos a verificar la significación del factor:

Si el factor es significativo ➔ las rectas son paralelas y las diferencias entre grupos son solo de nivel, independientemente del valor de la covariable



VARIABLES EXPLICATIVAS NUMÉRICAS Y CATEGÓRICAS. ANCOVA

Si el factor no es significativo ➔ las rectas son iguales y no hay diferencias respecto a los grupos, ni en nivel ni en la relación lineal



3) <u>Si la covariable no es significativa</u> ➔ no hay relación lineal entre la respuesta y la covariable, eliminamos ésta del modelo y pasamos a estudiar el ANOVA de 1 factor resultante.

MODELO LINEAL GENERAL

Analiza la relación entre una variable respuesta cuantitativa y variables explicativas numéricas y categóricas y sus interacciones.

La regresión múltiple, los ANOVA y ANCOVA son casos particulares de Modelo Lineal General

Formulación Matricial

$$Y = X\beta + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2 I_n)$$

$$Y|X \sim N(X\beta, \sigma^2 I_n)$$

n unidades experimentales independientes

p variables explicativas

Y vector columna (nx1)

X matriz (nxp)

β vector de coeficientes (px1)

ε vector de errores (nx1)

σ^2 varianza del error

MODELO LINEAL GENERAL

En el Modelo Lineal General…

- Las variables explicativas de tipo numérico aparecen como columnas de la matriz de diseño
- Las variables explicativas de tipo categórico se incluyen en el modelo mediante variables dummies
- Las interacciones se incluyen a partir del producto de las columnas asociadas

Regresión Lineal Múltiple (2 variables predictoras numéricas)

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

$$\begin{pmatrix} 1 \\ 3 \\ 4 \\ 6 \\ 7 \\ 8 \\ 5 \\ 7 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 1 & 5 & 1 \\ 1 & 8 & 4 \\ 1 & 3 & 5 \\ 1 & 5 & 6 \\ 1 & 1 & 9 \\ 1 & 2 & 3 \\ 1 & 1 & 7 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \end{pmatrix}$$

MODELO LINEAL GENERAL

Modelo Lineal General (1 variables predictora numérica y 1 categórica a 3 niveles, contraste baseline)

$$y_i = \beta_0 + \beta_1 x_{1i} + \alpha_i + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 d_{2i} + \beta_3 d_{3i} + \varepsilon_i$$

$$\begin{pmatrix} 2 \\ 3 \\ 4 \\ 2 \\ 5 \\ 6 \\ 4 \\ 8 \\ 9 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 1 & 5 & 0 & 0 \\ 1 & 8 & 1 & 0 \\ 1 & 3 & 1 & 0 \\ 1 & 5 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 2 & 0 & 1 \\ 1 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \end{pmatrix}$$

MODELO LINEAL GENERAL

ANCOVA (1 variables predictora numérica y 1 categórica a 3 niveles, contraste baseline y la interacción entre ambas)

$$y_i = (\alpha_0 + \alpha_i) + (\beta + \beta_i)x_{1i} + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 d_{2i} + \beta_3 d_{3i} + \beta_4 d_{2i} x_{1i} + \beta_5 d_{3i} x_{1i} + \varepsilon_i$$

$$\begin{pmatrix} 2 \\ 3 \\ 4 \\ 2 \\ 5 \\ 6 \\ 4 \\ 8 \\ 9 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 & 0 \\ 1 & 5 & 0 & 0 & 0 & 0 \\ 1 & 8 & 1 & 0 & 8 & 0 \\ 1 & 3 & 1 & 0 & 3 & 0 \\ 1 & 5 & 1 & 0 & 5 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 2 & 0 & 1 & 0 & 2 \\ 1 & 1 & 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \end{pmatrix}$$

MODELO LINEAL GENERAL

En R, la sintaxis para especificar los modelos permite incluir diferentes configuraciones. Considerando las variables X, Y y Z como variables numéricas y A, B y C como variables categóricas (factores), podemos especificar los siguientes modelos como casos particulares del Modelo Lineal General

Y~ X	Regresión simple
Y ~ X + Z	Regresión Múltiple
Y ~ X+Z+ X:Z	Regresión Múltiple con interacciones
Y ~ X+ X:X	Regresión Polinomial
Y ~ A	ANOVA 1 Factor
Y ~ A+ B+ C	ANOVA Efectos principales
Y ~ A+ B+ A:B	ANOVA factorial (con interacciones)
Y ~ A+ A%in%B	ANOVA anidado
Y ~ A+ X	ANCOVA
Y ~ A+ X+ X:A	Regresión con pendientes por grupo (ANCOVA con interacción)
Y ~ X*A+Z+B	Modelo Lineal General

MODELO LINEAL GENERAL. ESTIMACIÓN, INFERENCIA Y PREDICCIÓN

Estimación del Modelo $Y|X \sim N(X\beta, \sigma^2 I_n)$

Función de log-verosimilitud

$$-\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)$$

Estimación de los parámetros del modelo

$$\hat{\beta} = (X'X)^{-1}XY$$

Estimación de la variabilidad

$$\hat{\sigma}^2 = MSE = \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{n - p} = \frac{e'e}{n - p}$$

MODELO LINEAL GENERAL. ESTIMACIÓN, INFERENCIA Y PREDICCIÓN

Inferencia

$$\widehat{Var(\hat{\beta})} = \hat{\sigma}^2 (X'X)^{-1}$$

Test de Hipótesis:

$$H_0: L\beta = 0$$
$$H_1: L\beta \neq 0$$

donde L es un vector o matriz de coeficientes de funciones lineales estimables de los parámetros

$$SS(L\beta = 0) = \left(L\hat{\beta}\right)'(L(X'X)^{-1}L')^{-1}\left(L\hat{\beta}\right)$$

$$MS(L\beta = 0) = \frac{SS(L\beta = 0)}{Rang(L)}$$

$$F_L = \frac{MS(L\beta = 0)}{MSE} \sim F_{Rang(L),n-p}$$

MODELO LINEAL GENERAL. ESTIMACIÓN, INFERENCIA Y PREDICCIÓN

Predicción

1) Predicción Puntual para el valor esperado: $E(Y|X_0) = X_0\beta$ ➔ $\widehat{E(Y|X_0)} = X_0\hat{\beta}$

Varianza de la predicción: $V(E(\widehat{Y|X_0})) = \sigma^2 X_0{}'(X'X)^{-1}X_0$

Intervalo de Confianza (para el valor esperado):

$$X_0\hat{\beta} \pm t_{n-p,\alpha/2}\hat{\sigma}\sqrt{X_0{}'(X'X)^{-1}X_0}$$

2) Predicción Puntual para el valor observado: $Y_0 = X_0\beta + \varepsilon_0$ ➔ $\hat{Y}_0 = X_0\hat{\beta}$

Varianza de la predicción: $V(\hat{Y}_0) = \sigma^2 X_0{}'(X'X)^{-1}X_0 + \sigma^2$

Intervalo de Predicción (para un individuo concreto):

$$X_0\hat{\beta} \pm t_{n-p,\alpha/2}\hat{\sigma}\sqrt{1 + X_0{}'(X'X)^{-1}X_0}$$

MODELO LINEAL GENERAL. VALIDACIÓN

Validación

Premisas del modelo $Y|X \sim N(X\beta, \sigma^2 I_n)$:

Linealidad: Dependencia funcional de la respuesta en función de las covariables de tipo lineal

$$E(Y|X) = X\beta$$

Errores con distribución Normal, incorrelacionados y con varianza común

$$\varepsilon \sim N(0, \sigma^2 I_n)$$

Para validar el modelo es necesario verificar las premisas en base al análisis de los residuos

$$e = Y - \hat{Y} = Y - X\hat{\beta} = Y - X(X'X)^{-1}X'Y = (I_n - H)Y$$

La expresión de la varianza de los residuos es la siguiente

$$V(e_i) = \sigma^2(1 - h_{ii})$$

donde h_{ii} es el elemento i-ésimo de la diagonal de la matriz de proyección H

MODELO LINEAL GENERAL. VALIDACIÓN

Tipos de residuos en el Modelo Lineal General

Residuos "crudos":

$$e_i = y_i - \hat{y}_i$$

Residuos escalados:

$$c_i = \frac{e_i}{\hat{\sigma}}$$

Residuos estandarizados:

$$p_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

Residuos estudentizados:

$$r_i = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}}$$

donde $\hat{\sigma}_{(i)}$ es la estimación de la desviación estándar residual obtenida sin incluir la observación i-ésima

MODELO LINEAL GENERAL. VALIDACIÓN

El escalamiento de los residuos permite comparar su distribución empírica con la distribución normal (de forma aproximada). Los distintos tipos de escalamiento intentan que la aproximación a la distribución de referencia sea lo más fiable posible.

Los residuos estandarizados tienen en cuenta el efecto leverage (apalancamiento) de la observación. Básicamente ajusta la variabilidad del error de cada observación en función de la posición relativa del vector de variables explicativas respecto al centro de gravedad.

Los residuos estudentizados, además utilizan una estimación de la desviación estándar residual que es independiente del error, y por lo tanto es una mejor estimación si el error considerado es un valor atípico.

MODELO LINEAL GENERAL. VALIDACIÓN

Hipótesis sobre los residuos (estudentizados):

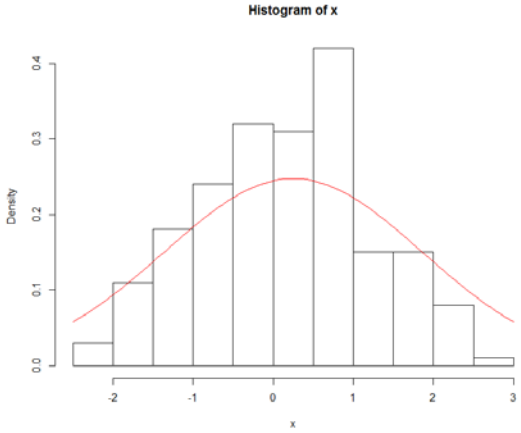
1) Normalidad: Se verifica mediante representaciones gráficas i pruebas de hipótesis adecuadas
    a. Histograma (hist):
        i. Es una aproximación gráfica a la densidad.
        ii. Se suele superponer la curva normal (curve).
        iii. Permite detectar fácilmente asimetría y/o valores extremos
        iv. Con tamaño de muestra pequeño no es adecuado
        v. Es muy dependiente de cómo se han definido los intervalos

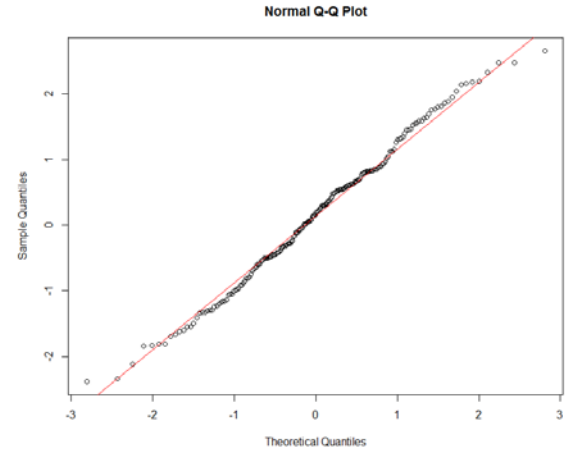
    b. Normal Probability plot (qqplot):
        i. Su construcción no depende de ningún parámetro.
        ii. Es muy útil para determinar desviaciones de normalidad (asimetría, colas pesadas, atípicos)
        iii. Se suele incluir una línea de referencia que pasa por los cuartiles 1 y 3 (qqline)
        iv. Se pueden incluir bandas de confianza (método qqPlot del package car)

    c. Test de bondad de ajuste a la distribución Normal
        i. Hay diferentes opciones: Shapiro-Wilks, Kolmogorov-Smirnoff,…

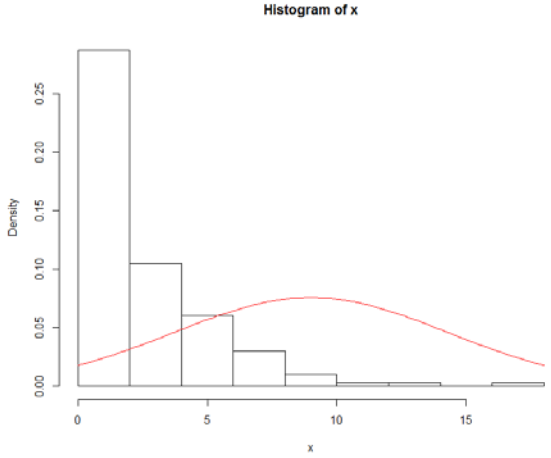
MODELO LINEAL GENERAL. VALIDACIÓN

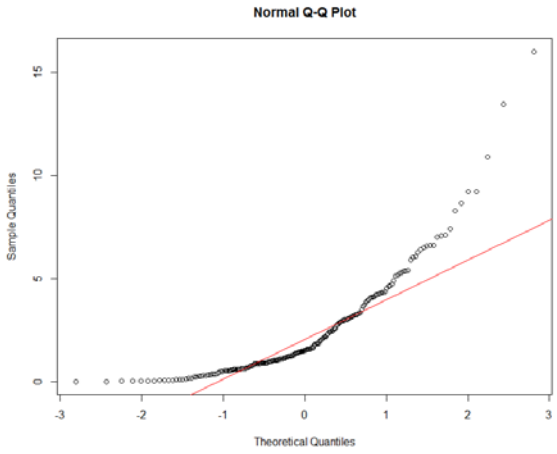
Normalidad



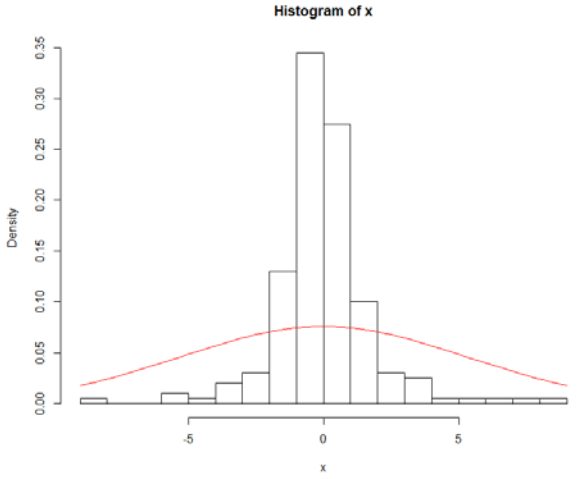


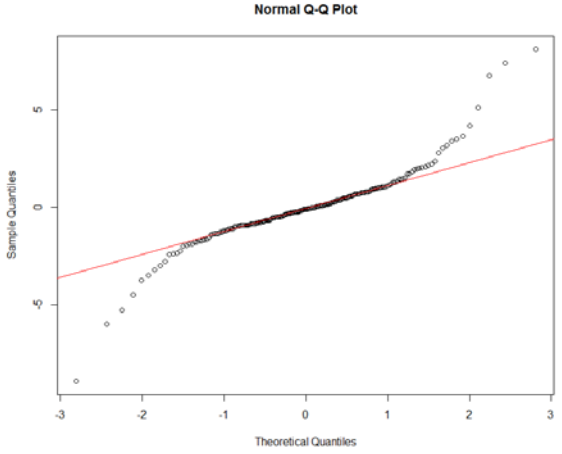
Asimetria





Exceso de Kurtosis





MODELO LINEAL GENERAL. VALIDACIÓN

2) Linealidad y Homogeneidad de varianzas: Se comprueban estas hipótesis sobre los residuos también usando gráficos. En modelos ANOVA existen test para comparar las varianzas en cada grupo (Fisher, Levene, Barlett,…)

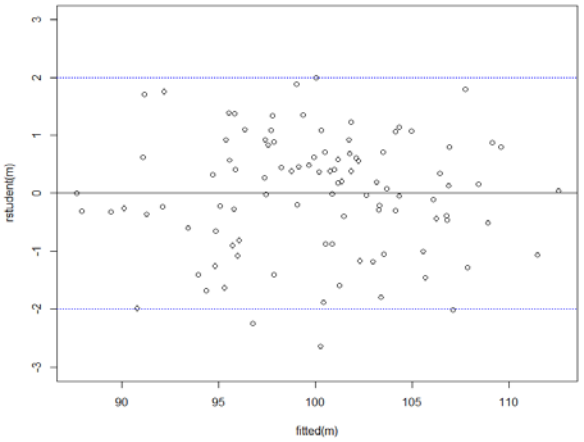
   a. Residuos estandarizados/estudentizados vs. Predicciones: Permite verificar visualmente las hipótesis de linealidad y homogeneidad de varianzas. Para validar el modelo es necesario una disposición aleatoria de las observaciones alrededor del cero, sin que la amplitud varíe.

   b. Residuos vs. Variables explicativas en el modelo: Si se detectan patrones no aleatorios es indicación de no linealidad respecto a la variable explicativa (puede ser necesaria una transformación o añadir términos no lineales)

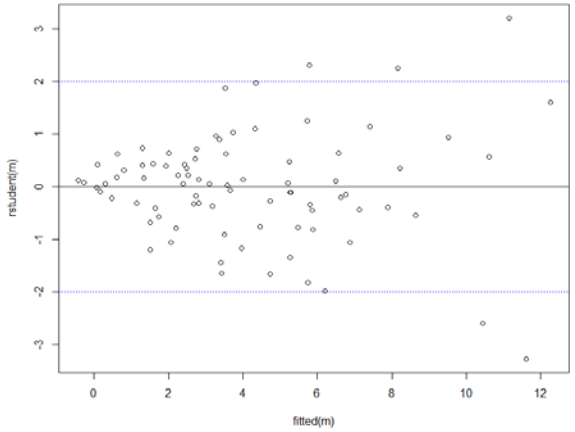
   c. Residuos vs. Variables explicativas no del modelo: Si se está construyendo el Modelo Lineal General y hay variables excluidas, puede ser útil verificar que los residuos son aleatorios respecto a otras variables descartadas.

MODELO LINEAL GENERAL. VALIDACIÓN

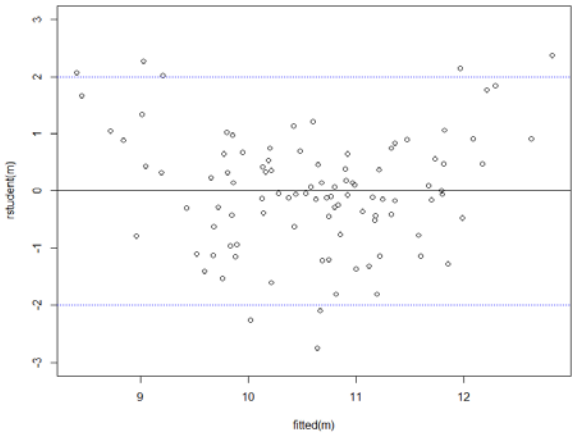
Linealidad y Homocedasticidad



Heterocedasticidad



No linealidad

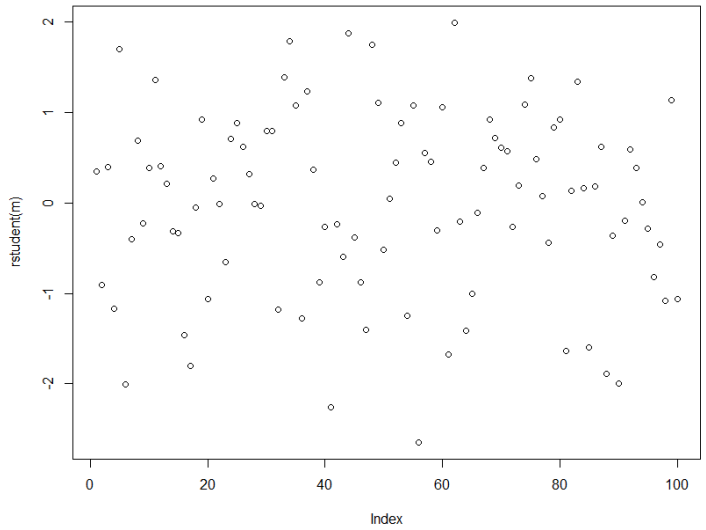


MODELO LINEAL GENERAL. VALIDACIÓN

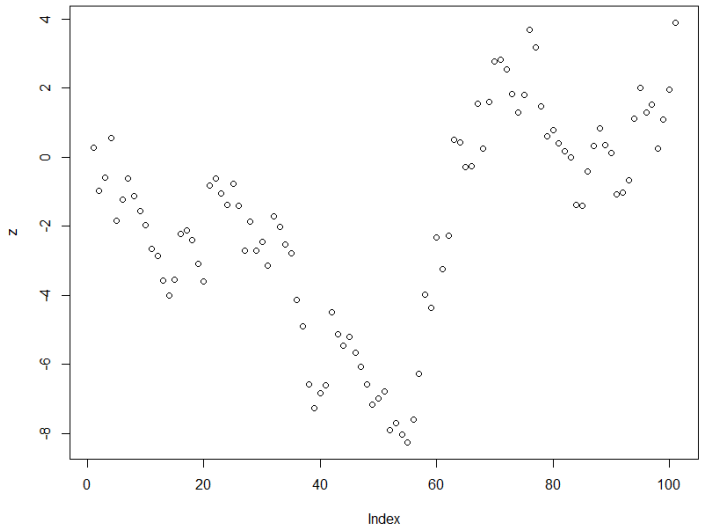
3) Independencia: Pese a que se razona en base al método de obtención de los datos, en ocasiones es posible detectar patrones de dependencia en los residuos. En diseños experimentales se incluyen protocolos para garantizar la independencia de las observaciones (diseños aleatorizados,…)

a. Plot de residuos vs. Orden: Si los datos se han obtenido en un orden natural y existe dependencia temporal, es posible detectarlo mediante este gráfico

Independencia temporal



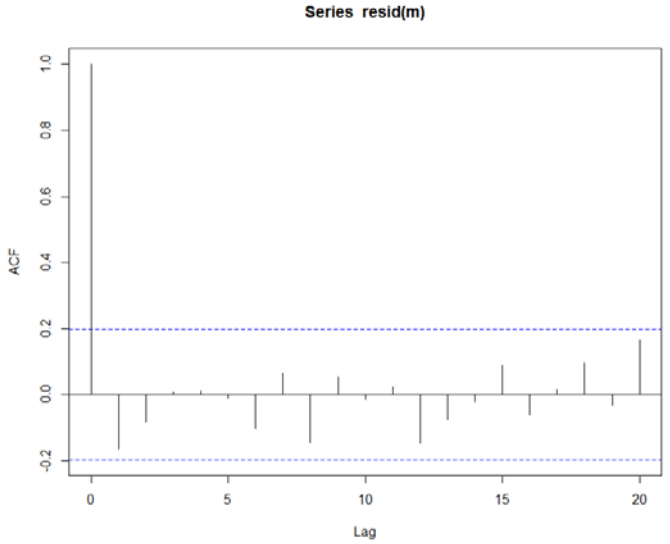
Dependencia temporal



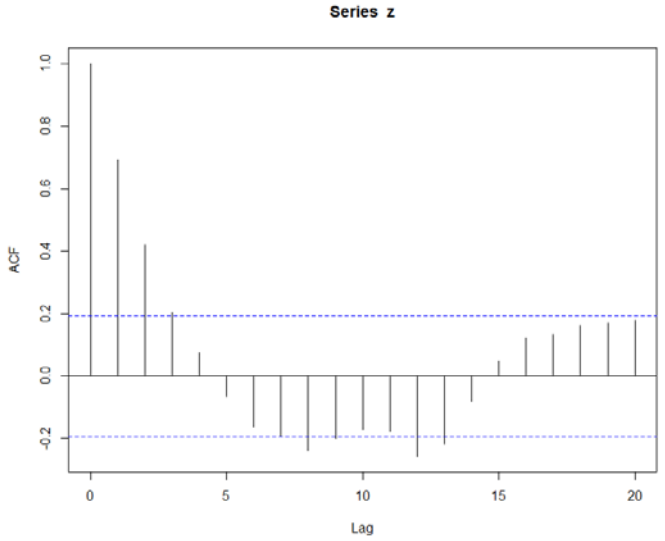
MODELO LINEAL GENERAL. VALIDACIÓN

b. Función de autocorrelación: También permite detectar correlación temporal en los residuos

Independencia temporal



Dependencia temporal



c. Test de Durbin-Watson: Contraste de hipótesis para determinar la independencia temporal de los residuos

MODELO LINEAL GENERAL. VALIDACIÓN

4) Datos Atípicos y/o Influyentes:

Un valor atípico (outlier) es una observación con un valor inusual de la variable respuesta condicionada a los valores de sus variables explicativas. Si el residuo atípico es positivo se interpreta como que el valor observado es muy superior a lo esperado por el modelo. En el caso de ser negativo, la observación está muy por debajo de la predicción realizada por el modelo.

Un valor atípico no necesariamente afecta a la estimación de los parámetros del modelo, pero incrementa la estimación de la varianza residual, y en consecuencia los errores estándar de los estimadores aumentan.

ATENCIÓN: La presencia de datos atípicos puede afectar a la interpretación de los gráficos!

Un valor influyente es una observación que afecta al ajuste del modelo. Si se suprime del conjunto de datos, los parámetros estimados y/o la validación del modelo pueden variar sustancialmente.

MODELO LINEAL GENERAL. VALIDACIÓN

Para cada observación es posible cuantificar el efecto de eliminarla sobre cada parámetro del modelo:

$$\mathrm{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{i(-i)}}{\hat{\sigma}_{(-i)}\sqrt{h_{ii}}}$$

Mide el efecto de suprimir el caso sobre la predicción. Para tamaños de muestra pequeños, valores por encima de 1 son "sospechosos"

$$\mathrm{DFBETAS}_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{\hat{\sigma}_{(-i)}\sqrt{(X'X)^{-1}}}$$

Mide el efecto de suprimir el caso sobre los coeficientes. Para tamaños de muestra pequeños, valores por encima de 1 son "sospechosos"

$$\mathrm{DCOOK}_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(-i)})^2}{\hat{\sigma}(p+1)}$$

Distancia de Cook: Mide el efecto de suprimir el caso sobre el conjunto de la regresión.

MODELO LINEAL GENERAL. SELECCIÓN DEL MEJOR MODELO

Para determinar el conjunto de regresores que determinan el mejor modelo, los criterios básicos se basan en obtener un modelo en el que:

- Los términos incluidos en el modelo sean significativos, maximizando en lo posible la bondad del ajuste del modelo
- Los términos excluidos no incluyen una mejora significativa en el ajuste del modelo, ya que el criterio de parsimonia determina que no existan términos no significativos en el modelo

Diferentes criterios para comparar globalmente dos modelos anidados:

R2: No es adecuado porque siempre aumenta al incrementar el número de variables en el modelo

R2-ajustada: Si el término incluido es significativo, tiende a aumentar, pero al incluir términos no significativos va reduciendo su valor

- $C_p = \frac{RSS_p}{S^2} - (n - 2p)$ (Cp de Mallows) RSSp es la suma de cuadrados residual con p regresores
- $AIC = -2l(\hat{\beta}, y) + 2p$ (Akaike Information Criteria)
- $BIC = -2l(\hat{\beta}, y) + \log(n) * p$ (Bayesian Information Criteria)

Estos indicadores son diferentes versions que constituyen un compromise entre lo bien que ajusta el modelo y el número de parámetros (complejidad) Modelos con valores inferiores son preferidos.