MUBD

Màster Universitari en Enginyeria de Dades Massives (Big Data)

Estadística

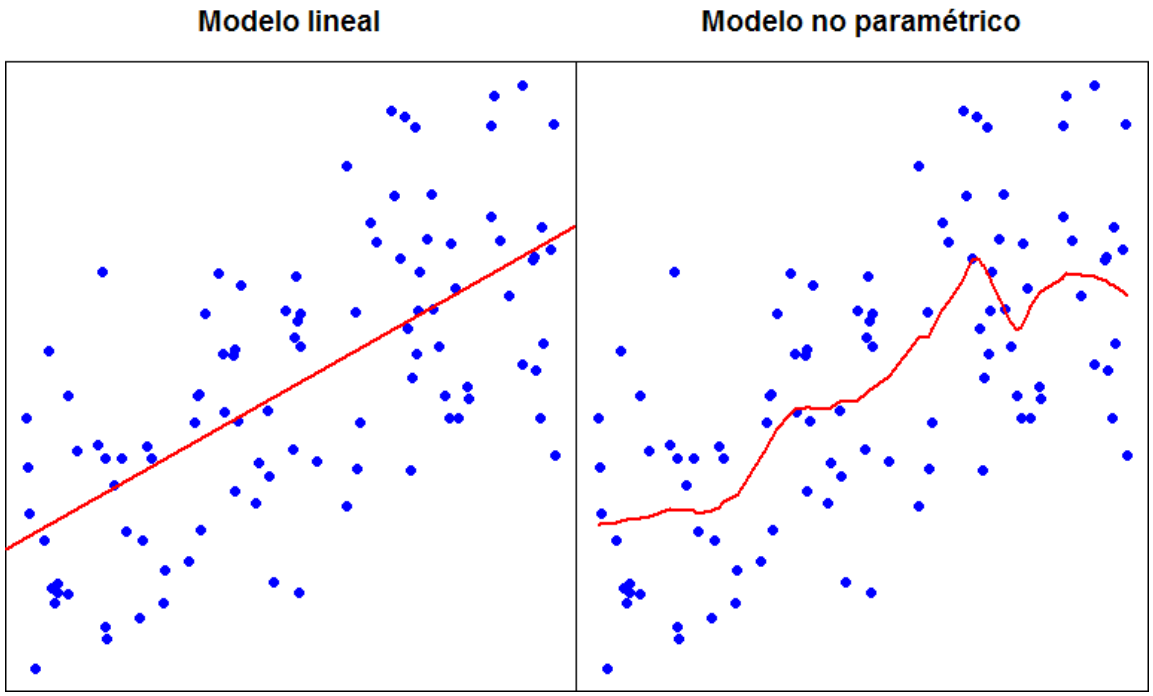
laSalle ENG
Universitat Ramon Llull

Índice

Modelo lineal

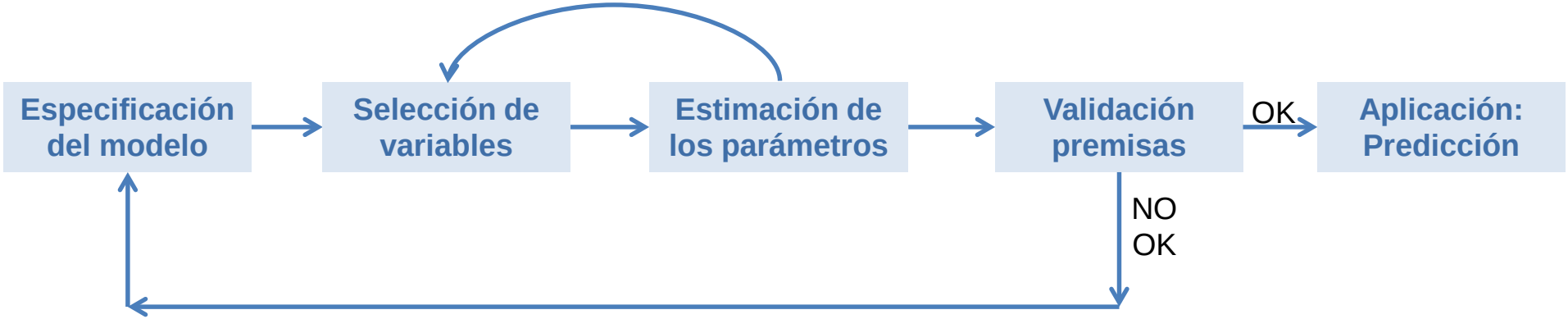
Introducción

- El modelo lineal, a pesar de su simpleza, puede ser una buena aproximación en un gran abanico de situaciones
- Otros modelos tienden a realizar sobrestimaciones que después no se corroboran con nuevos datos



Diseño de un modelo

Fases



- Una vez especificado el modelo y estimado los parámetros, para que sea útil (aplicarlo y hacer predicciones), hace falta estudiar las premisas asumidas.
- Si el modelo no se valida, se puede:
  - Realizar **transformaciones** (log(X), log(Y), 1/Y, Y/X, raíces, potencias, recategorizaciones...)
  - Buscar **otras variables predictoras**
  - **Otros modelos**

Modelo lineal

Formulación

- Modelo lineal
  - $Y_i = \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \cdots + \beta_p \cdot X_{pi} + \epsilon_i \quad \epsilon_i \sim N(0, \sigma)$ $(k=1...p;\ i=1...n)$
  - Y_i: $valor\ de\ la\ variable\ respuesta\ en\ el\ caso\ i-ésimo$
  - X_{ki}: $valor\ de\ la\ variable\ predictora\ k\ en\ el\ caso\ i\text{-}ésimo$
  - β_0: $término\ independiente$
  - β_k: $coeficiente\ (pendiente)\ de\ la\ variable\ X_k$ $(k=1...p)$
  - ϵ_i: $error$
  - σ: $desviación\ típica\ de\ los\ errores\ (desviación\ residual)$
- Los parámetros a estimar del modelo serán β_0, β_k y σ^2
- El modelo lineal **simple** contiene únicamente una variable predictora
- En el caso de las variables categóricas, se tendrá tantos coeficientes a estimar como número de categorías menos uno (se calcula el efecto de cada categoría respecto a una categoría escogida de referencia – se denomina contraste basal)

Modelo lineal

Ejemplo - R

- Estimar la dureza de un material en función de sus componentes

```
Call:
lm(formula = Strength ~ Cement + BlastFurnaceSlag + FlyAsh +
    Water + Superplasticizer + CoarseAggregate + FineAggregate +
    Age, data = datos)
Coefficients:
     (Intercept)            Cement  BlastFurnaceSlag            FlyAsh             Water
       -18.78742           0.11391           0.10193           0.08609         -0.16940
Superplasticizer   CoarseAggregate     FineAggregate               Age
         0.20413           0.01583           0.02437           0.12058
```

- Cada Kg. de cemento de más, se asocia con un incremento en la dureza de 0.11 unidades
- No poner ningún componente a la mezcla supone una dureza de -18.8.
  - En este caso, la interpretación de b_0 (estimador de β_0) no tiene sentido
  - Para darle sentido, p.ej. se podrían centrar todas las variables en su media

Modelo lineal

Estimación de los parámetros

- β_0 , β_k y σ^2 son valores poblacionales *auténticos* y desconocidos que se han de estimar. La estimación de las β's proporciona la recta (o hiperplano) estimada:

$$\hat{y}_i = b_0 + b_1 \cdot X_{i1} + b_1 \cdot X_{i1} + \cdots + b_p \cdot X_{ip}$$

- Esta recta (o hiperplano) permite hacer predicciones para cada observación con su error de predicción:

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

- La estimación mínima cuadrática consiste en calcular los estimadores **b's** de las **β's**, minimizando la suma de los errores de predicción al cuadrado:

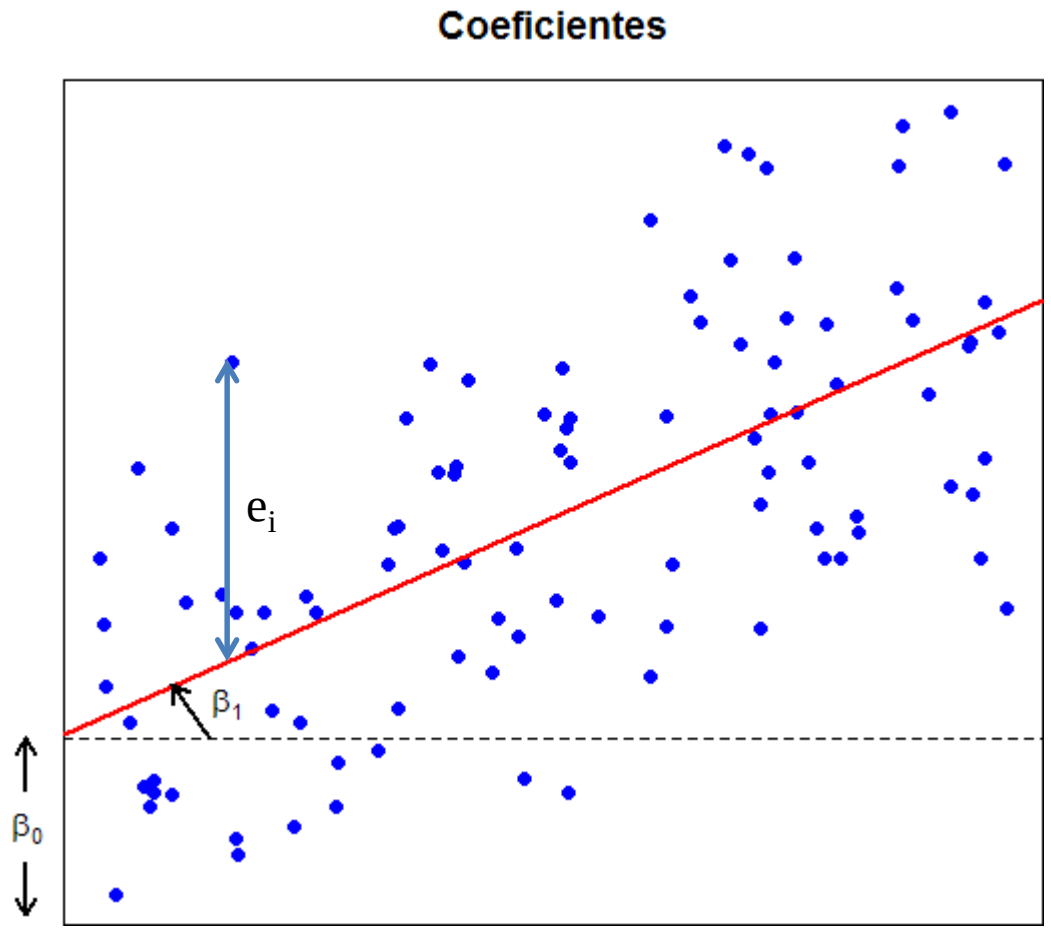
$$\sum (e_i)^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2$$

- En R, la solución al problema de minimización la proporciona la instrucción *lm*

Modelo lineal

Interpretación de los coeficientes

- β_0 representa el valor de la variable respuesta al anularse todas las variables predictoras
- Si X_1 es continua, β_1 representa el incremento de la variable respuesta por cada incremento unitario de X_1
- Si X_1 es categórica, β_1 representa el incremento de la variable respuesta asociado a la categoría i-ésima respecto a la categoría de referencia de X_1
- En el modelo lineal simple, los residuos son las distancias verticales de los puntos observados a la recta.



Selección de variables

Métodos

- Métodos automáticos (basados en AIC o BIC):
  - forward
  - backward
  - forward-backward

- Selección manual:
  - Devianza. ¿Es mejor el modelo con una nueva variable?
  - Significación. ¿Es la variable estadísticamente significativa?
  - Colinealidad. ¿La información aportada por la variable ya es aportada por otras?

Selección de variables

Métodos automáticos

- En R, la instrucción *step* realiza una selección automática de variables basándose en el *AIC* (Akaike Information Criteria) o en el BIC (Bayesian Information Criteria)

$$AIC = 2k - 2\log(L) \qquad BIC = \log(n) \cdot k - 2\log(L)$$

- *k* es el número de parámetros del modelo y *L* es la verosimilitud del modelo (indicador numérico que cuantifica un buen ajuste). Por tanto, el AIC cuanto más pequeño, mejor.
- El AIC y, sobre todo, el BIC favorecen a los modelos parsimoniosos (pocos parámetros)
- El proceso de selección se puede hacer de 3 maneras:
  - Forward. Se parte del modelo nulo (sin ninguna variable) y se van incluyendo progresivamente aquellas variables que más minimizan el AIC o el BIC
  - Backward . Se parte del modelo con todas las variables y se van suprimiendo progresivamente aquellas variables que más minimizan el AIC o el BIC
  - Forward-Backward. Combinación de los 2 anteriores
- LASSO es otra metodología que penaliza aquellas variables poco relevantes (no se verá)

Selección de variables

Metodología manual

- Realizar un ajuste univariante para cada una de las variables. Seleccionar aquellas con una significación de $p < 0.10$. Puede usarse cualquier sistema a nivel univariado.
- Ajustar el modelo multivariante con las variables seleccionadas
- Escoger las variables candidatas a eliminarse una a una según su significación (empezando por las menos significativas). Una variable podrá ser eliminada si:
  - La devianza (igual que el AIC pero sin el término 2k) entre los 2 modelos no es estadísticamente significativa (se puede usar la instrucción anova aplicada a 2 modelos)
  - El coeficiente de las otras variables no se modifica en más de un 10% al eliminarla (en caso contrario es una variable confusora que se debe conservar)
  - El VIF (ver diapositiva siguiente) asociado a esa variable es superior a 5 se eliminará aunque no cumpla ninguna de las condiciones anteriores
- Tanto si se elimina, como no, se continua evaluando la eliminación para el resto de variables no significativas ($p > 0.05$)

Selección de variables

Colinealidad - VIF

- Las variables predictoras no deben tener excesiva correlación entre ellas.
- Correlaciones altas indican que la cantidad de información compartida entre esas variables es alta y que, por tanto, explicaran casi lo mismo de la variable respuesta
- El VIF es un indicador de la colinealidad de la variable *j* con el resto de variables:

$$VIF = \frac{1}{1 - R_j^2}$$

`R: vif (paquete car)`

- El R^2 es una medida de la capacidad predictiva de un modelo. Fluctúa entre 0 (el modelo no predice nada) y 1 (la respuesta queda determinada conociendo las predictoras). El R_j^2 corresponde al modelo que tiene como variable respuesta la variable *j*, y como predictoras el resto de variables (sin incluir la respuesta)
- Un VIF de una variable superior a 5 indica que esa variable debería ser suprimida ya que un 80% de la información contenida en la misma, ya se encuentra presente en las otras.
- El no considerar la colinealidad provoca estimaciones más imprecisas ya que aumenta los errores estándar de las estimaciones

Modelo lineal

Ejemplo - R

- Estimar la dureza de un material en función de sus componentes

```
Call:
lm(formula = Strength ~ Cement + BlastFurnaceSlag +…+ FineAggregate + Age, data = datos)
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -18.78742   45.34568  -0.414 0.678925
Cement             0.11391    0.01466   7.772 1.11e-13 ***
BlastFurnaceSlag   0.10193    0.01793   5.683 3.01e-08 ***
FlyAsh             0.08609    0.02235   3.852 0.000142 ***
Water             -0.16940    0.06732  -2.517 0.012349 *
Superplasticizer   0.20413    0.16432   1.242 0.215076
CoarseAggregate    0.01583    0.01645   0.962 0.336727
FineAggregate      0.02437    0.01895   1.286 0.199477
Age                0.12058    0.01048  11.507  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 11.08 on 315 degrees of freedom
Multiple R-squared:  0.5973,    Adjusted R-squared:  0.5871
F-statistic: 58.41 on 8 and 315 DF,  p-value: < 2.2e-16
```

El cemento aumenta la dureza un 0.11 por cada incremento unitario

El error esperado en la estimación del coeficiente es 0.015

El cemento influye en la dureza ya que el p-valor es muy inferior a 0.05

Error típico. Error esperado en la predicción

El modelo predice un 59% de la variabilidad de la dureza

Validación

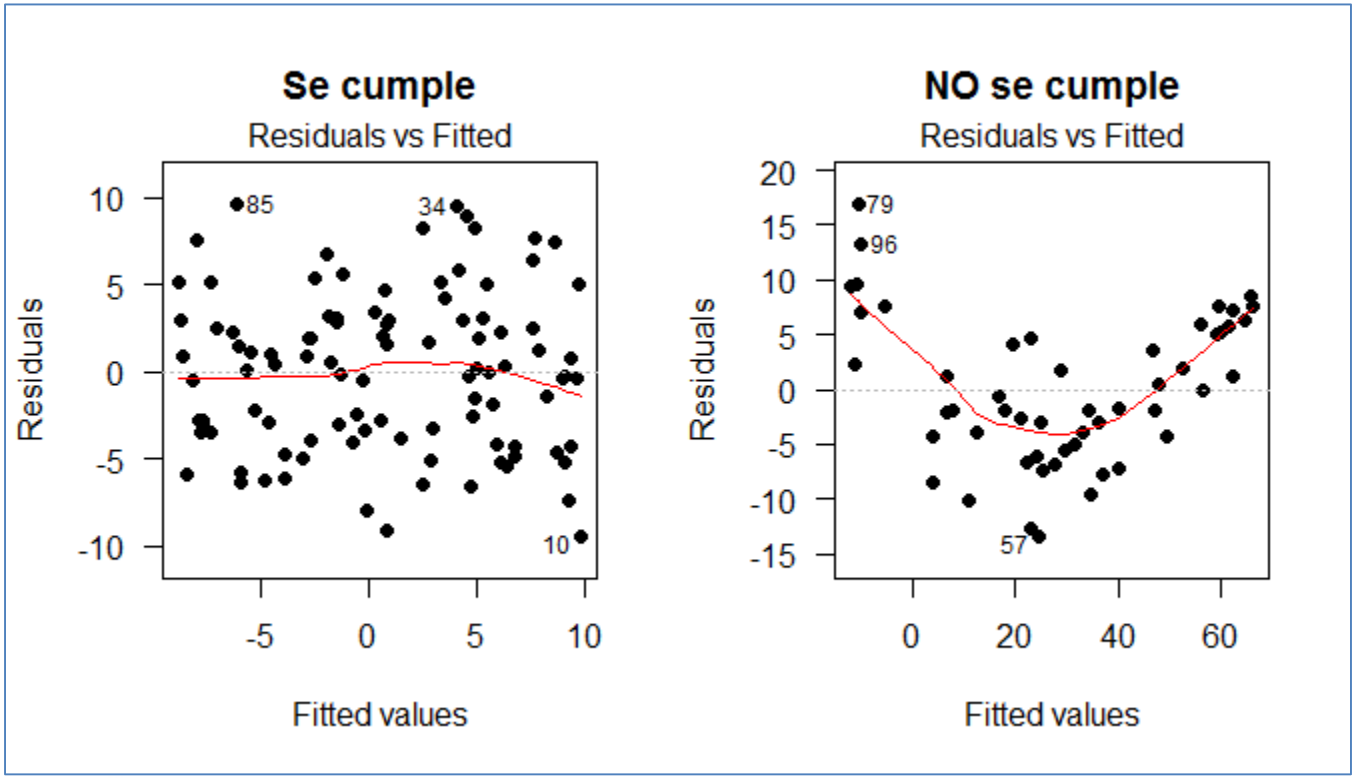
Premisas

- Las premisas del modelo lineal son:
  - Linealidad. Una recta/plano/hiperplano se ajusta bien a los datos
    - $Y_i = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} \cdot \boldsymbol{X_i}$
  - Homoscedasticidad. Variabilidad constante
    - $\epsilon_i \sim N(0, \boldsymbol{\sigma})$
  - Normalidad de los residuos. Los errores son normales
    - $\epsilon_i \sim \boldsymbol{N}(0, \sigma)$
  - Independencia
    - La muestra es aleatoria simple y el resultado de una observación no condiciona el resto.
- Todas estas premisas se verifican mediante el análisis de los residuos

Validación

Premisas - Linealidad

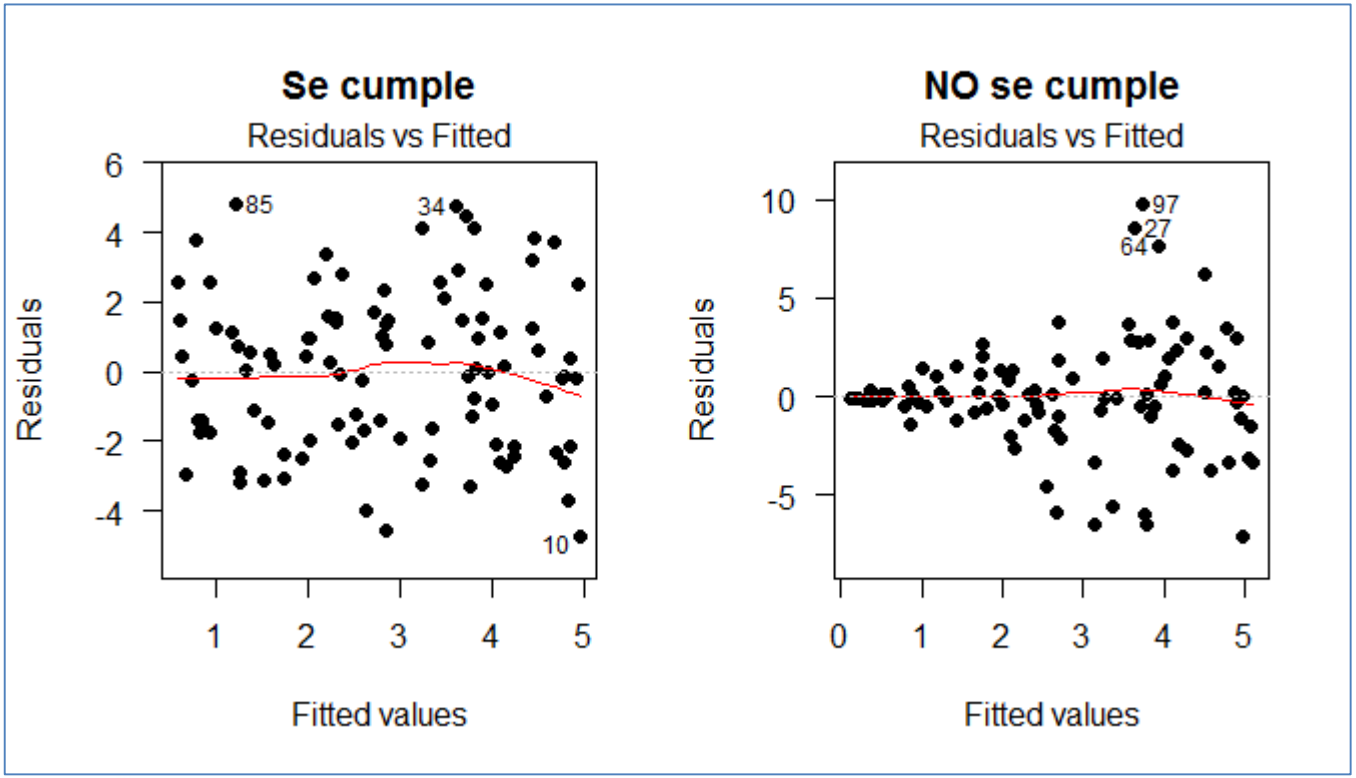
- Se comprueba con el gráfico de residuos versus valores predichos:
  - Los residuos deben distribuirse uniformemente por encima y por debajo del cero a lo largo de los valores predichos.



Validación

Premisas - Homoscedasticidad

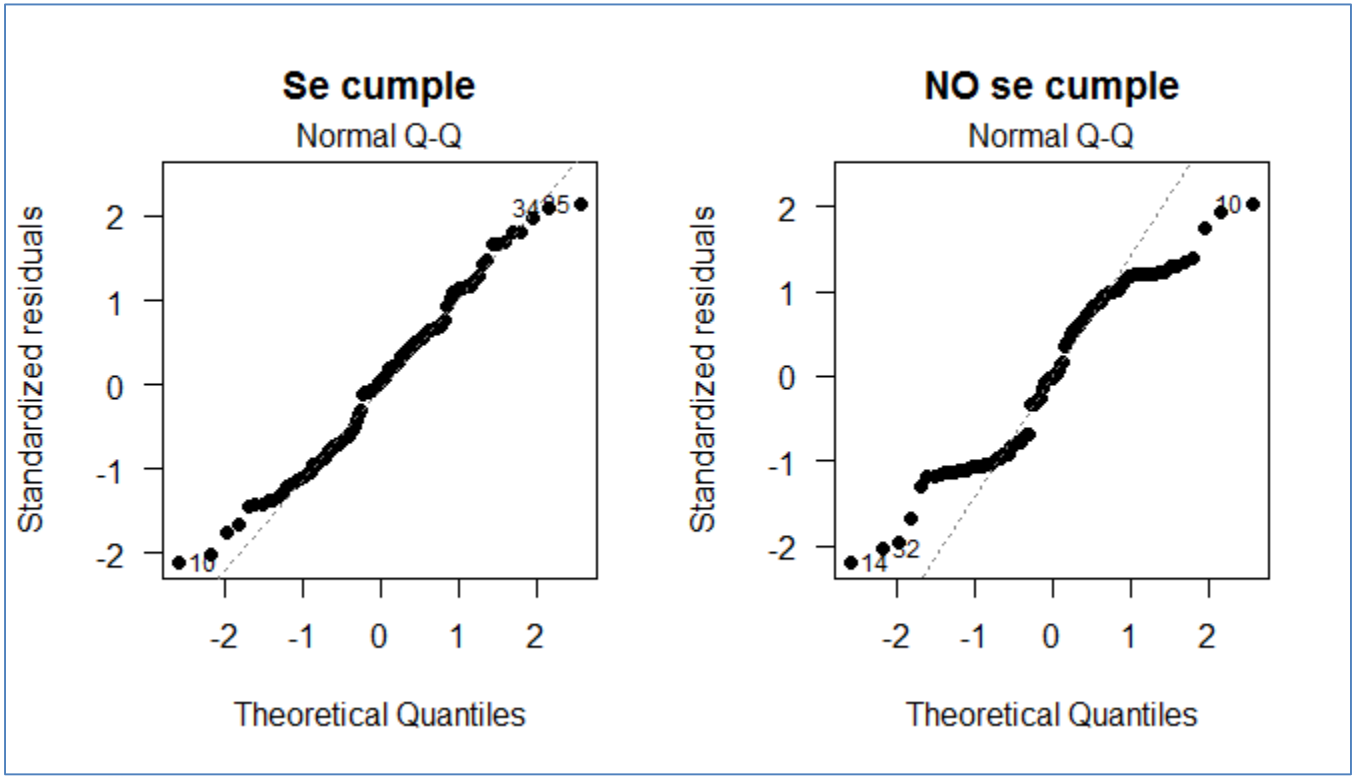
- Se comprueba (también) con el gráfico de residuos vs valores predichos:
  - Los residuos deben distanciarse del cero lo mismo a lo largo de los valores predichos (no tener forma de embudo)



Validación

Premisas – Normalidad de los residuos

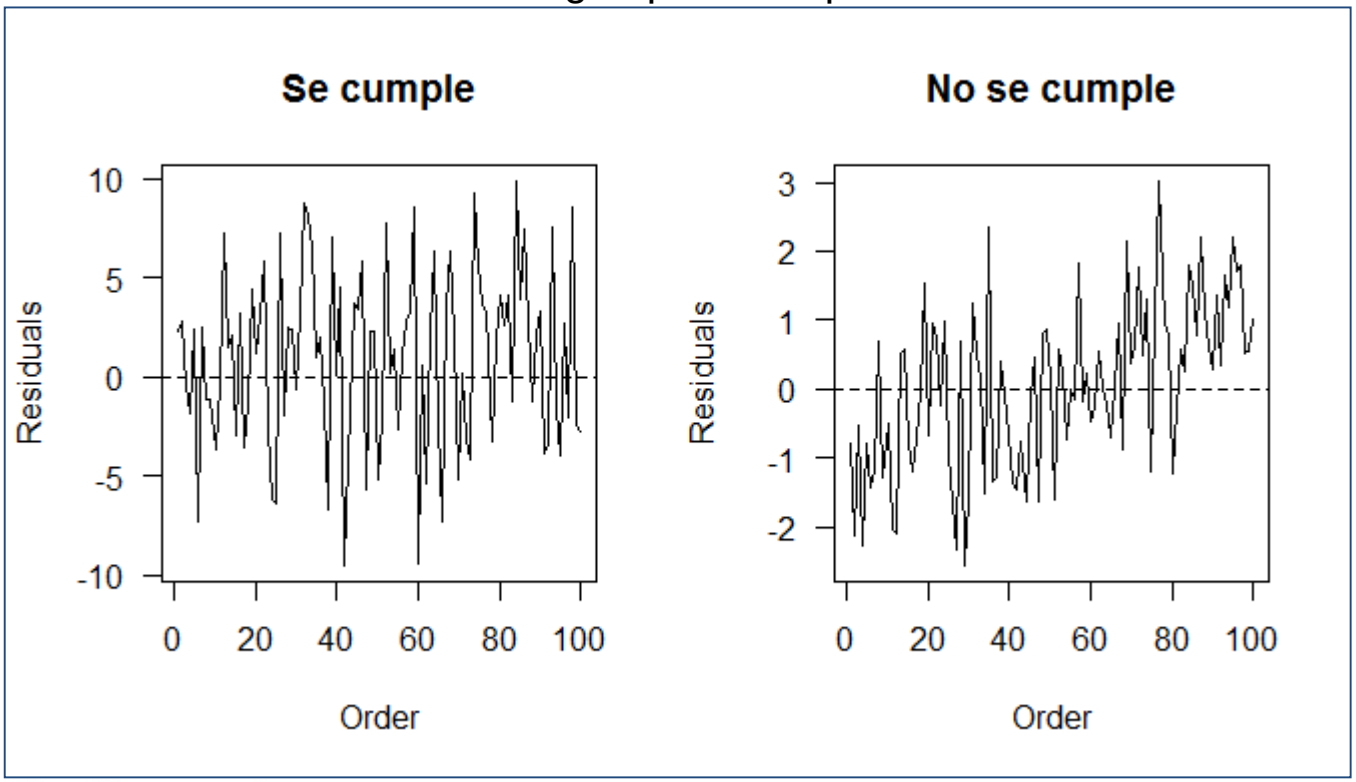
- Se comprueba con el ***qqnorm*** de los residuos:
  - Los residuos deben ajustarse a la recta de Normalidad



Validación

Premisas – Independencia

- No es verificable 100% ya que la independencia se garantiza con un buen diseño de la recogida de datos. El tiempo suele producir dependencias entre valores cercanos cronológicamente. Por este motivo, se estudia la relación entre los residuos y el orden de recogida
  - No se debe observar ningún patrón específico



* En el segundo gráfico se ve una tendencia creciente de los residuos a lo largo del tiempo

Validación

Transformaciones

- Generalmente, aplicar transformaciones sobre la variable respuesta o sobre las variables predictoras puede solucionar el no cumplimiento de alguna de las premisas
- Variable respuesta
  - La transformación logarítmica es la más común

$$Y' = \log(Y) = \beta_0 + \beta_1 \cdot X$$

  - No obstante, la transformación "ideal" sobre la variable respuesta es la de Box-Cox

$$Y' = \begin{cases} (Y^{\lambda} - 1)\ si\ \lambda \neq 0 \\ \log(Y) \quad si\ \lambda = 0 \end{cases}$$

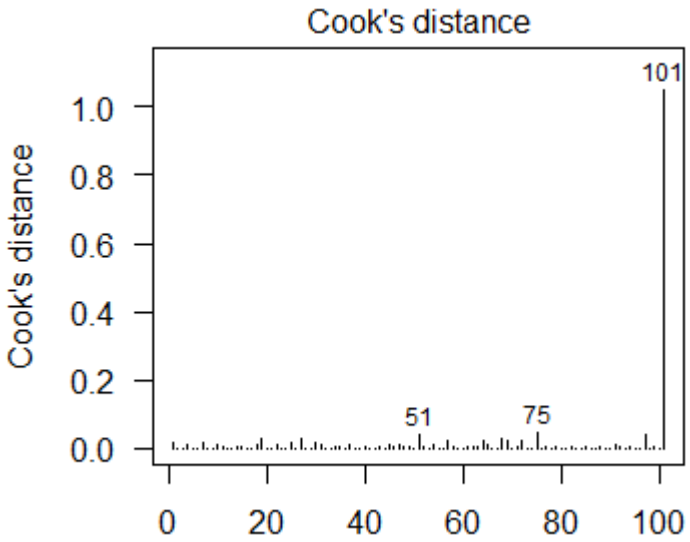
    El valor de λ, se puede obtener con la función *boxCox* del paquete *car*

- Variables predictoras
  - Graficando los residuos para el modelo lineal simple de cada una de ellas en función de su valor para observar si existen patrones
  - En las que se hallen patrones, se aplicarán las transformaciones pertinentes mediante inspección visual para lograr la linealidad.

Validación

Observaciones influyentes

- Pueden existir observaciones que condicionen la estimación sobremanera.
- Conviene estudiar el ajuste sin la presencia de dichas observaciones influyentes
- La **distancia de Cook** mide la influencia de una observación en el modelo. Es decir, mide el efecto que tendría sobre el modelo suprimir dicha observación. Depende de:
  - Cuan alejada esté en el espacio de las variables predictoras (X's)
  - Del residuo de dicha observación
- Las observaciones influyentes no se deben eliminar a no ser que se sepa que se trata de un error
- La distancia de Cook mide para cada observación, la suma de las diferencias entre las predicciones con el modelo completo y sin la observación.
- Se consideraran influyentes aquellas observaciones con una distancia de Cook superior a 1 o superior a 4/n (hacer análisis gráfico)



Modelo lineal

Predicción

- Tipos de predicción
  - **Valor puntual**. Estimar el valor de la respuesta para un/a individuo/observación con unos valores de las variables predictoras concretos.
    - Ej: ¿Cuál es el retraso esperado para el vuelo BCN-ROM de Vueling de las 8:00?

      ```
      R: predict(…, interval = "prediction")
      ```

  - **Valor esperado**. Estima el valor promedio de la respuesta en todos los individuos u observaciones con unos valores de las variables predictoras concretos.
    - Ej: ¿Cuál es el retraso promedio esperado para los vuelos BCN-ROM de Vueling que parten a las 8:00?

      ```
      R: predict(…, interval = "confidence")
      ```

- En ambos casos, la estimación puntual de la predicción es la misma, pero no su incertidumbre. En el caso del valor puntual tenemos un rango más amplio de valores plausibles.

Modelo lineal

Capacidad predictiva

- La capacidad predictiva de un modelo se mide por el coeficiente de determinación (R^2)
- El R^2 es un estadístico que oscila entre 0 y 1 (o entre 0 y 100%) e indica el porcentaje de variabilidad total que queda explicada por el modelo.
- La variabilidad de la respuesta se puede descomponer en 2 términos:

$$Variabilidad\ total = Variabilidad\ explicada + Variabilidad\ residual$$

- El coeficiente de determinación es:

$$R^2 = \frac{variabilidad\ explicada}{variabilidad\ total} = \frac{\sum(\hat{y}_i - \bar{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$$

Anexo: División de la muestra (capacidad predictiva)

Capacidad predictiva – Muestras entrenamiento y test

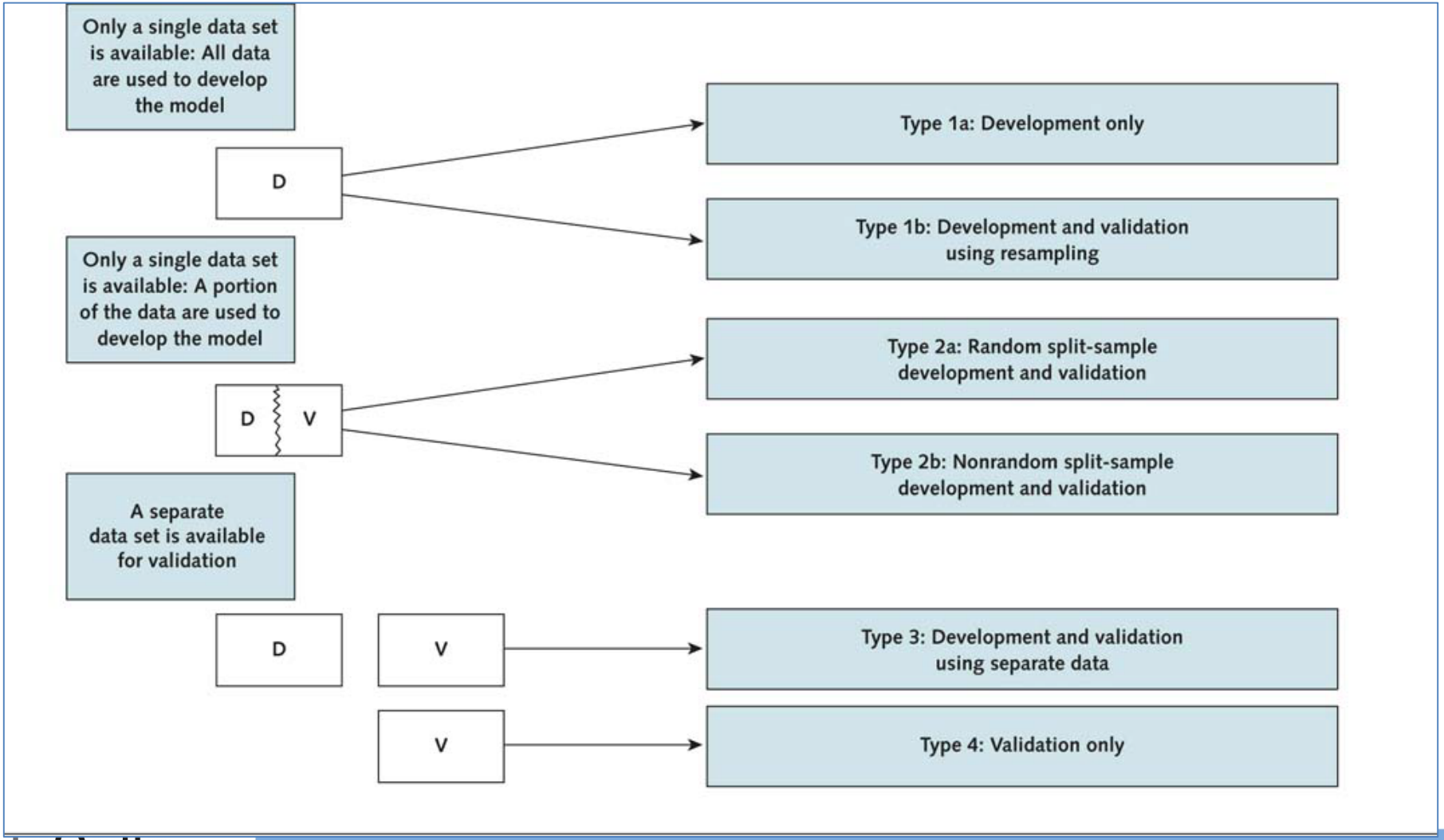
- Para evaluar la capacidad predictiva (en cualquier modelo) es conveniente dividir la muestra en dos submuestras:
  - Entrenamiento. Para construir el modelo
  - Test. Para evaluar la capacidad predictiva
- Generalmente, la submuestra de entrenamiento la compone entre un 60 y 80% de la muestra original y la muestra test, lo compone el porcentaje restante.
- El modelo obtenido con la muestra de entrenamiento se evaluará en la muestra test y se obtendrá una medida de la capacidad predictiva
- El EQM (Error Cuadrático Medio) nos da una estimación de cuan bueno es nuestro modelo:

$$EQM = \sqrt{\frac{\sum(e_i - o_i)^2}{n}}$$

e_i es el valor predicho y o_i el valor observado

Anexo: División de la muestra (capacidad predictiva)

Capacidad predictiva – Muestras entrenamiento y test



Anexo: División de la muestra (capacidad predictiva)

Capacidad predictiva – Muestras entrenamiento y test

Type 1a	Development of a prediction model where predictive performance is then directly evaluated using exactly the same data (apparent performance).
Type 1b	Development of a prediction model using the entire data set, but then using resampling (e.g., bootstrapping or cross-validation) techniques to evaluate the performance and optimism of the developed model. Resampling techniques, generally referred to as "internal validation", are recommended as a prerequisite for prediction model development, particularly if data are limited (6, 14, 15).
Type 2a	The data are randomly split into 2 groups: one to develop the prediction model and one to evaluate its predictive performance. This design is generally not recommended or better than type 1b, particularly in case of limited data, because it leads to lack of power during model development and validation (14, 15, 16).
Type 2b	The data are nonrandomly split (e.g., by location or time) into 2 groups: one to develop the prediction model and one to evaluate its predictive performance. Type 2b is a stronger design for evaluating model performance than type 2a because it allows for nonrandom variation between the 2 data sets (6, 13, 17).
Type 3	Development of a prediction model using 1 data set and an evaluation of its performance on separate data (e.g., from a different study).
Type 4	The evaluation of the predictive performance of an existing (published) prediction model on separate data (13).

Types 3 and 4 are commonly referred to as "external validation studies." Arguably type 2b is as well, although it may be considered an intermediary between internal and external validation.

Source: [TRIPOD guideline]

MUBD

Màster Universitari en Enginyeria de Dades Massives (Big Data)

Estadística

laSalle ENG
Universitat Ramon Llull