Análisis Avanzado de Datos

Nicolás López

URosario

Introducción a la clase

Introducción a la clase

Detalles del curso:

- Análisis avanzado de datos (AAD)
- 8 sesiones. Sede Claustro UR. Salón Turing. Tercer piso de Calatrava El Tiempo (Claustro).
- Horario: Sábados: 7am a 10am.
- Modalidad: Presencial.
- Profesor: Nicolás López.

Pre requisitos

- Fundamentos de programación en R.
- Fundamentos AED (Análisis Estadístico de Datos): Estadística introductoria.
- Contenido AED.
  - Modelos lineales: RLS y RLM.
  - Análisis en componentes principales.
  - Distribución normal multivariada.
  - Métodos de clasificación: kmeans y aglomeramiento jerárquico.

Programa: Disponible en eaulas.

Principios de R

Principios de R

Herramienta de gran importancia en el análisis de datos, particularmente en el contexto de AAD:

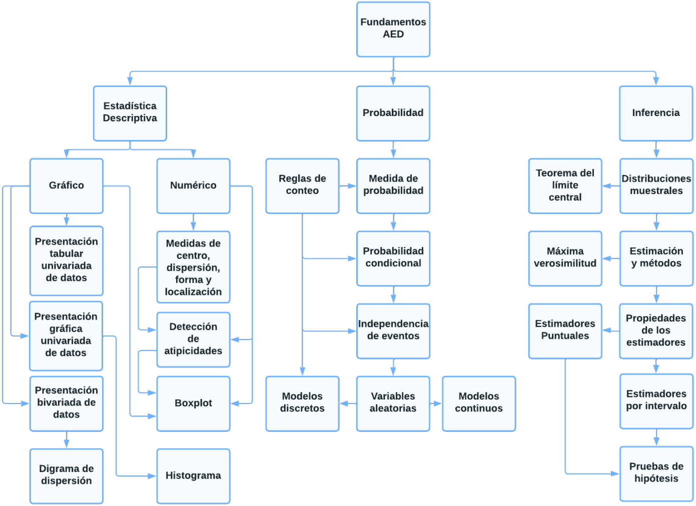
- Innovación en investigación y modelamiento: ej BTM.
- Gran trayectoria en visualización efectiva y manipulación de datos: ej tidyverse y ggplot.
- Modelamiento avanzado altamente documentado: gamlss.

Instrucciones de instalación local disponible en eaulas. También existe opción remota (antiguo Rstudio Cloud).

Laboratorios introductorios están disponibles en eaulas. Se recomienda realizar los laboratorios introductorios 0 y 1.

Estadística introductoria

Estadística introductoria



Aprendizaje estadístico

Introducción

Aprender el proceso subyacente generador de datos. El proceso es formalizado matemáticamente en el aprendizaje estadístico y se clasifica en dos grupos:

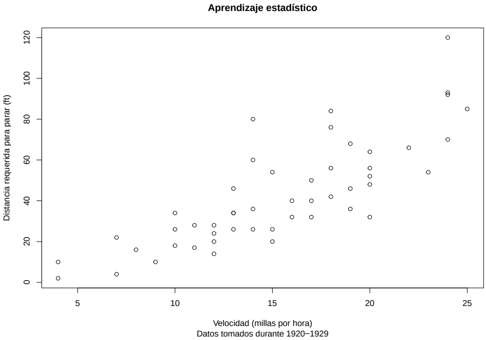
- Aprendizaje supervisado: Se tiene un resultado (*outcome*) que guía el proceso de aprendizaje (ej. identificación de dígitos).
- Aprendizaje no supervisado. No se tiene una medición de un resultado para guiar el aprendizaje (ej. clasificación de carros basado en características).

En ambos escenarios se cuenta con un conjunto de covariables (*features*) que permiten el aprendizaje.

¿Aprendizaje supervisado o no supervisado?

Ejemplo 1

```
plot(cars, xlab = "Velocidad (millas por hora)",
           ylab = "Distancia requerida para parar (ft)",
           main  = "Aprendizaje estadístico",
           sub   = "Datos tomados durante 1920-1929")
```



Ejemplo 2

```
#mnist = dslabs::read_mnist()
#i  = 10
#image(1:28, 1:28, matrix(mnist$test$images[i,], nrow=28)[ , 28:1],
#                  col  = gray(seq(0, 1, 0.05)),
#                  xlab = paste0("Número ",mnist$test$labels[i]),
#                  ylab = "", main = "Aprendizaje estadístico")
```

Tipos de variables

En el aprendizaje estadístico contamos con dos clases principales de variables:

- Cuantitativas
- Cualitativas

Esto tanto para las covariables como para la variable respuesta. Existe un mayor refinamiento en la categorización, pero por ahora basta entender que **el mismo problema de aprendizaje (supervisado o no) puede darse para diferente naturaleza de las variables**:

- Análisis de regresión lineal (simple/múltiple): Supervisado con respuesta cuantitativa.
- Análisis de componentes principales: No supervisado con covariables cuantitativas.
- Análisis de correspondencias (simple/múltiple): No supervisado con covariables cualitativas.
- Árbol de decisión: Supervisado con respuesta cualitativa.

Definiciones importantes

Se destacan 4 elementos fundamentales en el aprendizaje estadístico dada la revisión anterior:

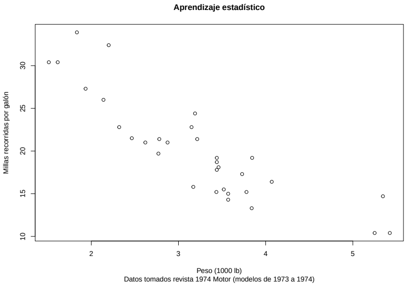
- Proceso generador P.
- Variable de entrada/covariable/input: X (uni/multivariada).
- Variable de salida/variable respuesta/output: Y (univariada usualmente).
- Observaciones/realizaciones/mediciones: $(x_1, y_1), ..., (x_n, y_n)$.

Estas mediciones son arregladas en una matriz dise~no $\mathbf{X}$.

RLS

Introducción

Si analizamos con detenimiento la gráfica de dispersión de los datos de velocidad podemos establecer con claridad una relación entre estas dos variables.



Si Y representa 'Millas recorridas por galón' y X es igual a 'Peso (1000 lb)', podemos representar una relación **determinística** entre las variables de la siguiente forma:

$$Y = \beta_0 + \beta_1 X$$

Por lo tanto una relación **aleatoria** que ajusta por el error está formulado por:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

La ecuación anterior se llama *Modelo de Regresión Lineal Simple*.

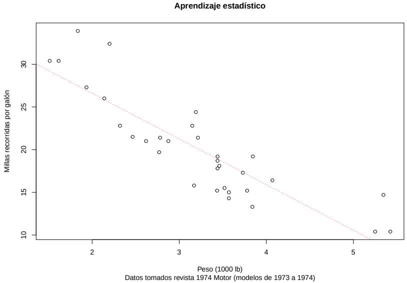
El modelo aleatorio esta completamente especificado cuando definimos las características aleatorias del error, suponemos que $\epsilon \sim N(0, \sigma^2)$. Tanto el error como la respuesta son aleatorias, bajo normalidad de ϵ, Y también es normal (¿por qué?). Entonces la respuesta esperada de Y dado X es:

$$E(Y|X) = \mu_{Y|X} = E(\beta_0 + \beta_1 X + \epsilon) = \beta_0 + \beta_1 X$$

Con una varianza igual a

$$V(Y|X) = \sigma_{Y|X} = V(\beta_0 + \beta_1 X + \epsilon) = V(\epsilon) = \sigma^2$$

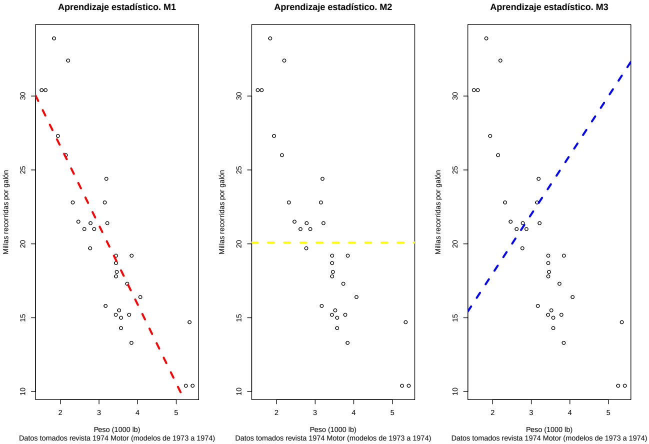
El modelo de regresión $\mu_{Y|X}$ es una línea recta de valores promedios, esto es, la altura de la línea de regresión en X es el valor esperado $\beta_0 + \beta_1 X$. Esto implica que hay una distribución de valores de Y para cada X, y que la varianza σ^2 de esta distribución es igual en cada X.



Con una realización de n observaciones $(x_1, y_1), ..., (x_n, y_n)$ se estiman los parámetros del modelo: $\hat{\beta}_0$, $\hat{\beta}_1$, y $\hat{\sigma}$.

Ajuste por mínimos cuadrados

Para encontrar la línea que se ajusta mejor a los datos, necesitamos una medida de calidad del ajuste. Bajo estos tres candidatos es claro cual resulta en una menor **suma de cuadrados**:



Es claro que:

$$SC(M1) < SC(M2) < SC(M3)$$

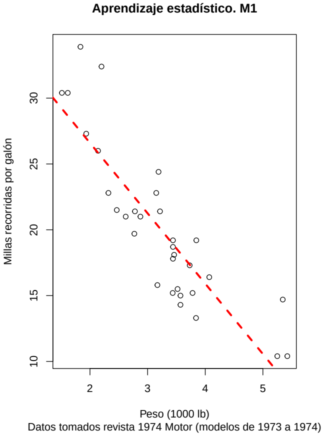
Con

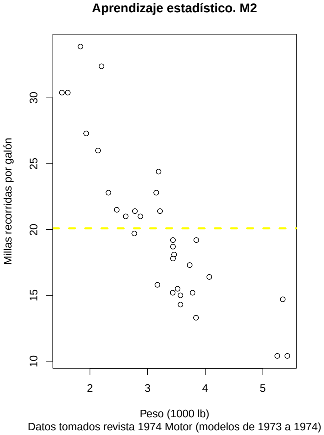
$$SC(Mj) = \sum_i (y_i - \hat{y}_{i,Mj})^2$$

En dónde

- El **modelo simple o reducido** (M2) no utiliza información de X para encontrar el valor de Y, asume el valor promedio de Y como modelo marginal. Este es considerado la línea base (*baseline*).
- La estimación por mínimos cuadrados consiste en encontrar β_0 y β_1 de tal forma que minimicen la suma de cuadrados, es decir, los **residuales cuadrados**. (¿qué sucedió con σ en la estimación?).

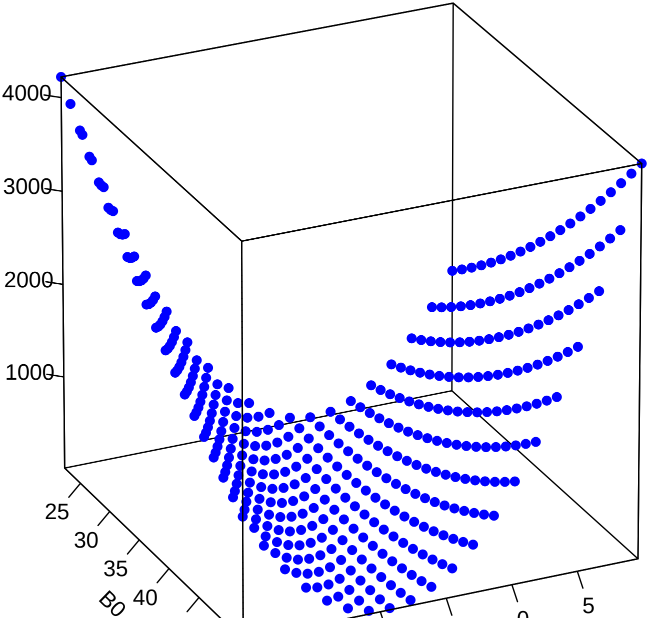
Es evidente que hay menor variabilidad alrededor de $M1$ que alrededor de $M2$, es decir que la variación de las millas recorridas es explicada por el peso del vehículo. ¿Cómo formalizar esta noción?





La SC del modelo simple ($SC(M2) = SC_T$) cuantifica la variabilidad total de Y respecto a su media. Por su parte $SC(M1) = SC_E$ mide la variación remanente al ajustar el modelo mediante mínimos cuadrados.

Fíjese que para cada modelo se tiene un intercepto y una pendiente (β_0, β_1), y con ello se obtiene SC_E, es decir $SC_E(\beta_0, \beta_1)$. Los valores estimados $(\hat{\beta}_0, \hat{\beta}_1)$ por mínimos cuadrados minimizan la función de error:



Para cualquier pareja de parámetros, comparar SC_T con SC_E cuantifica la reducción de la variabilidad bajo el modelo lineal en X. La reducción de la varianza en Y explicada por X bajo el modelo es igual a:

$$SC_M = SC_T - SC_E$$

Así SC_M cuantifica la reducción en la variación total al ajustar el modelo lineal en X. Al igual que SC_E, SC_M es función de (β_0, β_1), es decir $SC_M(\beta_0, \beta_1)$, la cual es maximizada en $(\hat{\beta}_0, \hat{\beta}_1)$ por mínimos cuadrados (¿qué unidades tiene SC_M?).

La SC_M es estandarizada como:

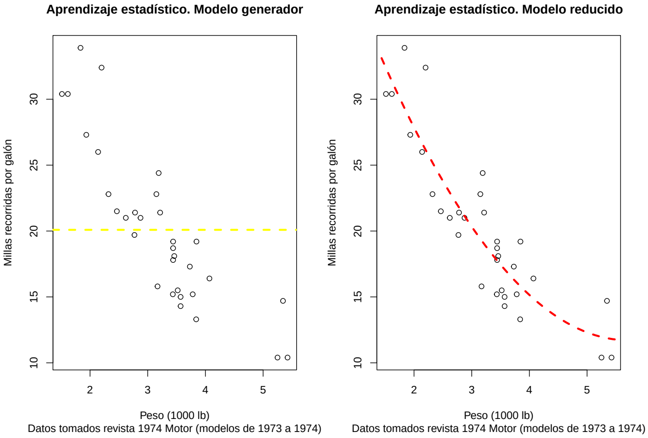
$$R^2 = \frac{SC_M}{SC_T}$$

- Con R^2 cuantificamos la **proporción** de la varianza en Y explicada por el regresor X.
- Al ser R^2 cercano a 1, SC_M se acerca a SC_T, es decir que el *modelo explica* la variabilidad en Y (¿cómo lo medimos objetivamente?).
- Al ser R^2 cercano a 0, SC_M se aleja de SC_T, es decir que el *modelo no explica* la variabilidad en Y (¿cómo lo medimos objetivamente?).

En nuestro ejemplo $R^2 = 0.75$, con lo cual hay una reducción de la varianza de un 75% en las millas recorridas al considerar linealmente el peso del vehículo.

Note que esta definición de R^2 aplica para situaciones aún más generales, por ejemplo, para un modelo con componente lineal y **cuadrático** en X:

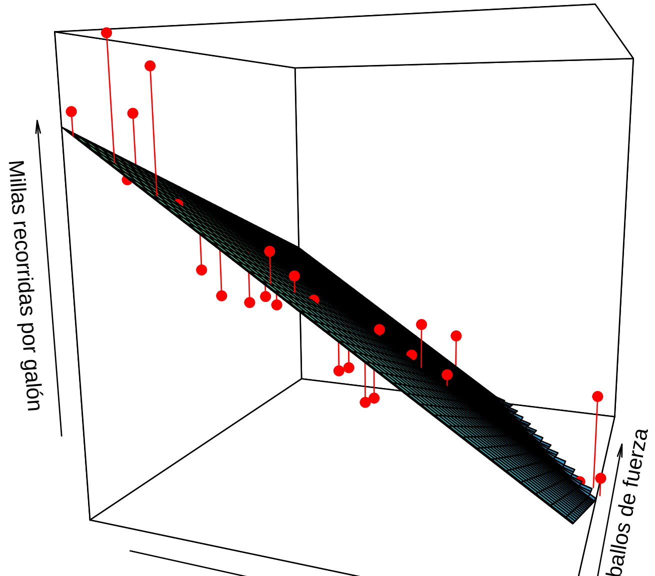
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2$$



Y bajo este modelo, tenemos un $R^2 = 0.81$.

También para el escenario multivariado, con Z igual a los caballos de fuerza del vehículo

$$Y = \beta_0 + \beta_1 X + \beta_2 Z$$



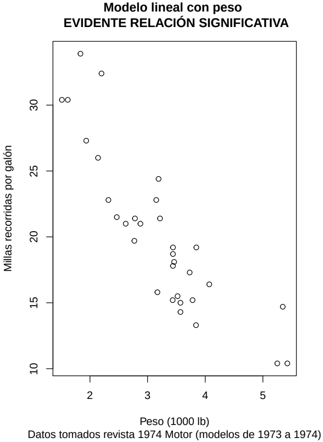
Note que incorporar una variable al modelo, esta puede o no ser relevante para explicar la variabilidad en Y. Si la variable Z no tiene efecto en la respuesta:

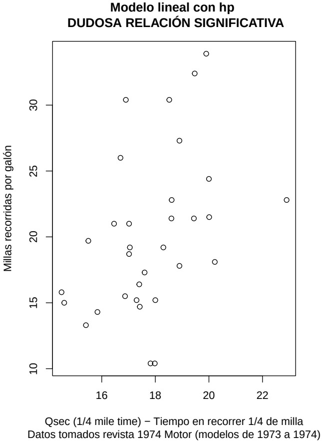
- Al minimizar SC_M, se lleva a $\beta_2 = 0$ y así $Y = \beta_0 + \beta_1 X + 0Z = Y = \beta_0 + \beta_1 X$ tenemos el modelo de RLS.
- SC_M es la misma bajo los dos modelos, es decir añadir Z no mejoró, ni empeoró R^2.
- Añadir variables mantiene igual o incluso **mejora** R^2 aunque no sean de utilidad para explicar Y.
  - Variables irrelevantes pueden estar correlacionadas con la variable Y por coincidencia.
  - Mas variables irrelevantes aumentan la probabilidad de que esto suceda, mejorando artificialmente R^2.

En la práctica se reporta usualmente un R^2 ajustado por en número de variables.

Signifincancia de la regresión

Volviendo a RLS, no es claro cómo determinar o medir qué tan significativo resulta el valor de R^2, ¿en qué punto de R^2 el modelo es realmente mejor con o sin la covariable?





Recuerde que:

- La variación o error total en los datos, SC_T, corresponde a la variación total al asumir el modelo reducido.
- Al considerar la covariable mediante el modelo lineal disminuimos este error, la nueva variación la llamamos SC_E.
- La diferencia entre SC_T y SC_E corresponde a la variabilidad explicada por el modelo, o SC_M.

SC_T es una suma de cuadrados, por lo cual tiene **grados de libertad** (gl) asociados. Los $gl(SC_T)$ indican la cantidad de unidades de información relacionadas con los n números independientes $y_1, \ldots y_n$ necesarios para calcular SC_T:

- Para el cálculo de SC_T hacen faltan $gl(SC_T) = n - 1$ unidades para determinarlo:
  - $y_1' = y_1 - \bar{y}$.
  - $\ldots$
  - $y_{n-1}' = y_n - \bar{y}$.
  - Como $\sum_i y_i' = 0$ se tiene $y_n' = -\sum_{i-1} y_i'$.
  - $y_n' = f(y_1', ..., y_{n-1}')$.

Con lo cual $SC_T = \sum_i (y_i - \bar{y})^2 = \sum_i y_i'^2 = f(y_1', ..., y_{n-1}')$ tiene $n - 1$ grados de libertad. Note que hace falta un parámetro (el promedio), para estimar SC_T, por eso se pierde un gl de los n que tienen los datos $y_1, \ldots, y_n$.

SC_E es una suma de cuadrados, por lo cual tiene **grados de libertad** (gl) asociados. Los $gl(SC_E)$ indican la cantidad de unidades de información relacionadas con los n números independientes $y_1, \ldots y_n$ necesarios para calcular SC_E:

- Para el cálculo de SC_E hacen falta $p = 2$ parámetros (pendiente e intercepto, en RLS) para ser estimado, por lo cual perdemos 2 grados de libertad, es decir $gl(SC_E) = n - p$.

SC_M es una suma de cuadrados, por lo cual tiene **grados de libertad** (gl) asociados. Los $gl(SC_M)$ indican la cantidad de unidades de información relacionadas con los n números independientes $y_1, \ldots y_n$ necesarios para calcular SC_M:

- Como vimos, $SC_M = SC_T - SC_E$, se tiene de la misma forma $gl(SC_M) = gl(SC_T) - gl(SC_E) = (n-1) - (n-p) = p - 1$, en RLS, $p - 1 = 2 - 1 = 1$.

De manera semejante a R^2, podemos definir una relación entre las sumas de cuadrados, esta vez entre SC_M y SC_E, para establecer un estadístico que caracterice la calidad del ajuste:

$$F' = \frac{SC_M}{SC_E}$$

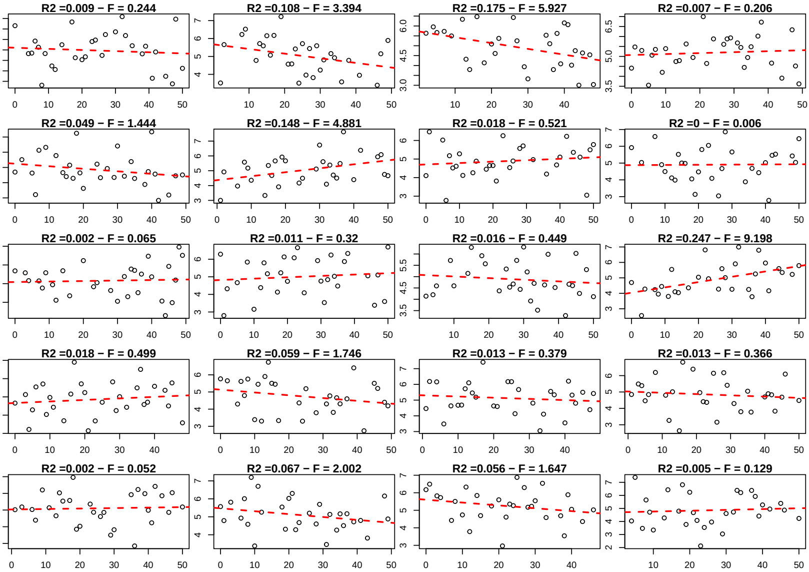
- F' es fácilmente interpretable: a medida que aumente, la variabilidad explicada por el modelo aumenta respecto a la que este deja de explicar.
- F' es una razón en lugar de una proporción, pero su diferencia con R^2 es de forma, más no de fondo. De hecho sus numeradores son iguales.

Al normalizar por los gl de cada SC en F', tenemos el estadístico F como un cociente de varianzas:

$$F = \frac{SC_M/gl(SC_M)}{SC_E/gl(SC_E)}$$

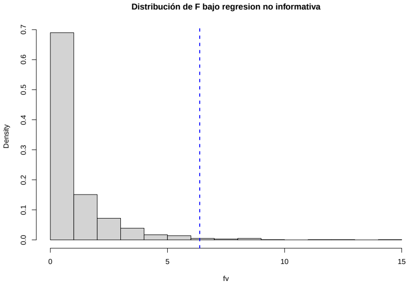
- Tanto F como F' tienen la misma interpretación.
- Su diferencia radica en que, cuándo la regresión no es informativa (es decir, los parámetros de las covariables son iguales a cero), F sigue una distribución estadística conocida, la distribución F-
- Los parámetros de F bajo la *hipótesis nula* son $gl(SC_M)$ en el numerador y $gl(SC_E)$ en el numerador y denominador.

Si la regresión no es informativa $\beta_1 = 0$ y los datos podrían verse como:



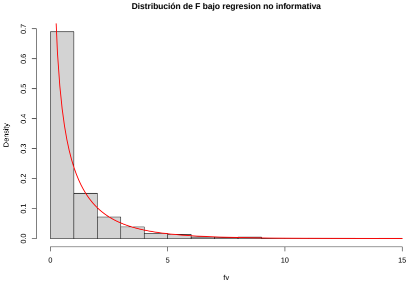
Podemos encontrar la distribución empírica de F bajo dicha **hipótesis** y agregar a esta los valores F observados para los modelos de regresión encontrados en la motivación de la sección

```
fv = NULL
set.seed(10)
for(i in 1:1000){
  xv    = sample(0:50,size=32) ; yv    = rnorm(32)
  lm_0  = lm(yv~xv)
  fv[i] = summary(lm_0)$fstatistic['value']
}
lm_sig  = lm(mpg~wt,data=mtcars)
lm_nsig = lm(mpg~qsec,data=mtcars)
hist(fv,prob=TRUE,main = 'Distribución de F bajo regresion no informativa')
abline(v = summary(lm_sig)$fstatistic['value'],col='red',lty=2,lwd=2)
abline(v = summary(lm_nsig)$fstatistic['value'],col='blue',lty=2,lwd=2)
```



No es necesario encontrar manualmente la distribución, ya que bajo la **hipótesis nula** F sigue una distribución estadística conocida, la distribución F, con $gl(SC_M) = 1$ en el numerador y $gl(SC_E) = n - 2 = 32 - 2 = 30$ en el denominador:

```
hist(fv,prob=TRUE,main = 'Distribución de F bajo regresion no informativa')
xval = seq(0,15,by=0.1)
yval = df(xval,df1=1,df2=30)
lines(xval,yval,col='red',lty=1,lwd=2)
```



Y con esta hacer inferencia (cálculo de p valor). Note que, nuevamente, esta cuantificación es también aplicable para RLM.

Finalmente se destaca que hay una lista de premisas bajo el modelo lineal:

- Relación (aproximadamente) lineal.
- Error con media cero.
- Error con varianza constante.
- Errores no correlacionados - correlación bajo RLS/RLM implica una disminución artificial de la varianza - falsa significancia.
- Errores normalmente distribuídos - necesaria para probar, entre otras, la hipótesis sobre F.

Los cuales no son detectados mediante R^2, o F, al ser propiedades globales del modelo. Un modelo inadecuado puede resultar en conclusiones incluso opuestas a las reales bajo el proceso real generador de datos.

Regresión ridge

Regresión ridge

Un problema que no es comúnmente mencionado para el modelo de regresión lineal es la **multicolinealidad**, en la cual una o varias covariables se encuentran linealmente relacionadas entre ellas de manera significativa. Esto incrementa la varianza en las estimaciones de los parámetros: $V(\hat{\beta}_i) = f(R^2_{X_i})$

```
##         mpg     wt   disp
## mpg   1.000 -0.868 -0.848
## wt   -0.868  1.000  0.888
## disp -0.848  0.888  1.000
```

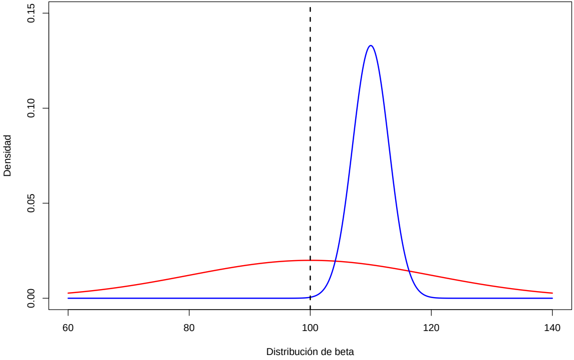
```
##
## Call:
## lm(formula = mpg ~ wt + disp, data = data_car)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4087 -2.3243 -0.7683  1.7721  6.3484
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.96055    2.16454  16.151 4.91e-16 ***
## wt          -3.35082    1.16413  -2.878  0.00743 **
## disp        -0.01773    0.00919  -1.929  0.06362 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.917 on 29 degrees of freedom
## Multiple R-squared:  0.7809, Adjusted R-squared:  0.7658
## F-statistic: 51.69 on 2 and 29 DF,  p-value: 2.744e-10
```

A su vez, esto aumenta la magnitud absoluta de los parámetros, ya que, con $\beta' = (\beta_1, ..., \beta_p)$ se tiene $(\hat{\beta} - \beta)'(\hat{\beta} - \beta)$ como la distancia al cuadrado entre el parámetro y su estimación por mínimos cuadrados:

$$E(\hat{\beta} - \beta)'(\hat{\beta} - \beta) = \sigma^2 \sum_i \frac{1}{\lambda_i}$$

Con $\{\lambda\}_i$ conjunto de valores propios de $\mathbf{X}'\mathbf{X}$. Este valor esperado es llamado **error cuadrático medio** de $\hat{\beta}$ (o $ECM(\hat{\beta})$), es decir que la multicolinealidad hace que $ECM(\hat{\beta})$ aumente. Esto no es bueno, ya que coeficientes muy grandes positivos se cancelarían con sus contrapartes correlacionadas con coeficientes muy grandes negativos.

Los estimadores de los parámetros β_i encontrados minimizando SC_E resultan ser de **varianza mínima** respecto a los demás estimadores lineales **insesgados** (BLUE). Esto es cierto tanto para RLS como para RLM (teorema de Gauss-Markov). Sin embargo, esto no es siempre bueno, pues no garantiza que la varianza sea pequeña.



Permitir menor varianza a costa de admitir sesgo en la estimación puede resultar en una estimación más consistente

En este escenario agregar sesgo en la estimación de los parámetros es muy conveniente, El modelo lineal es el mismo que antes:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z$$

Esta vez con con Z igual al cilindraje del vehículo. Sin embargo, en contraste con mínimos cuadrados, buscaremos ajustar los valores a los datos y además reducir **la magnitud** de los coeficientes. **No porque exista multicolinealidad el modelo lineal es un mal modelo para los datos**.

Como se mencionó anteriormente, $ECM(\hat{\beta})$ aumenta bajo multicolinealidad. Después de una manupulación algebráica, este puede ser escrito como:

$$ECM(\hat{\beta}) = V(\hat{\beta}) + (E(\hat{\beta}) - \beta)^2 = V(\hat{\beta}) + B(\hat{\beta})^2$$

Suponiendo que se tiene otro estimador de β, llamado $\hat{\beta}^*$ y permitimos que sea sesgado, $E(\hat{\beta}^*) - \beta \neq 0$, podríamos encontrar una menor varianza de aquella del BLUE. La **regresión ridge** es uno de los métodos para obtener estimadores sesgados de los coeficientes de regresión.

Surgen dos preguntas respecto al método:

- Qué tanto sesgo es necesario?
- Cómo a∼nadir el sesgo en la estimación de mínimos cuadrados?

Surgen dos preguntas respecto al método:

- Qué tanto sesgo es necesario?. Aquél que resulte en una menor varianza y menor sesgo.
  - Para la varianza, imagine que tiene un **conjunto de datos de prueba** disponible. Con dichos datos puede calcular la SC_E bajo varios sesgos. Aquel que incurra en el menor SC_E será un buen candidato.
- Cómo a∼nadir el sesgo en la estimación de mínimos cuadrados?. Depende el método: Ridge, Lasso, Elastic net son ejemplos de ello.

Agregar el sesgo en la estimación de β resulta en una generalización del estimador de mínimos cuadrados.

En regresión Ridge tenemos una constante $\lambda \geq 0$ llamada **parámetro de sesgo**. Este parámetro penaliza la estimación de mínimos cuadrados:

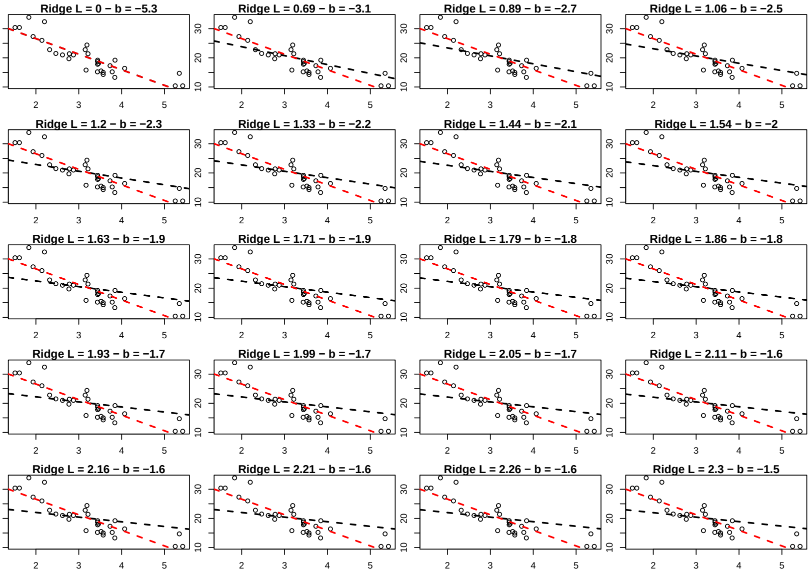
$$SC_{E_R} = \sum_i (y_i - \hat{y}_i)^2 + \lambda \sum_j \beta_j^2 = SC_E + \lambda \sum_j \beta_j^2$$

Note que

- $SC_{E_R} \geq SC_E$, y como SC_T se mantiene fijo, la regresión ridge disminuye el R^2 de mínimos cuadrados a medida que λ aumenta. Esto no es necesariamente malo -> Puede resultar en mayor poder de predicción fuera de la muestra (reduce **sobreajuste**).
- Al minimizar SC_{E_R} motivamos no solo a un buen ajuste a los datos, sino amdeás una menor magnitud de los parámetros.
- Si $\lambda = 0$ volvemos a RLS.

A $\sqrt{\sum_j x_j^2} = |x|_2$ se le llama la norma l^2 (minúscula) del vector x. A $\sqrt{\sum_j |x_j|} = |x|_1$ la norma l^1 de x.

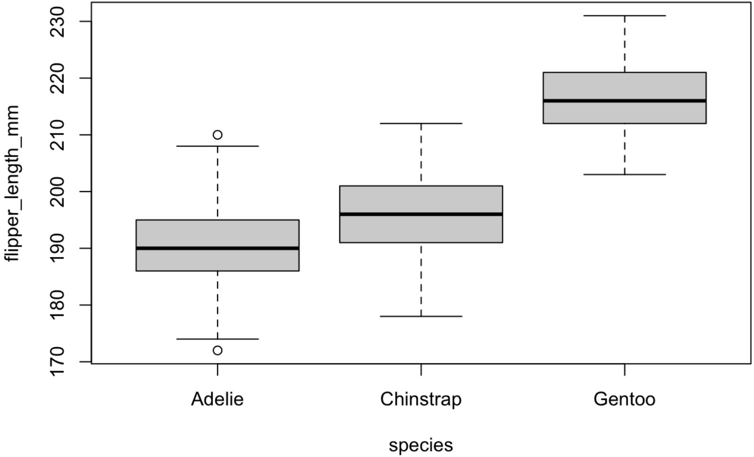
```r
lambda_val = c(0,log(seq(2,4,length.out = 19)))
lm_cars    = lm(mpg~wt,data=mtcars)
X          = cbind(rep(1,nrow(mtcars)),mtcars$wt)
y          = mtcars$mpg

ridge <- function(beta,X,y,lambda = 0) {
  beta_pen = beta[2:length(beta)]
  crossprod(y - X %*% beta) + lambda * length(y) * crossprod(beta_pen)
}

old_mai = par()$mai
par(mfrow=c(5,4), mai = c(0.35, 0.15, 0.15, 0.15))
for(i in 1:length(lambda_val)){
result_ridge = optim(rep(0, ncol(X)),
                     ridge,
                     X = X,
                     y = y,
                     lambda = lambda_val[i],
                     method = 'BFGS')
plot(mtcars$wt,
     mtcars$mpg,
     main  = paste0("Lambda = ", round(lambda_val[i],2)," - m = ",  round(result_ridge$par[2],2)))
abline(a=result_ridge$par[1],b=result_ridge$par[2],lty=2,lwd=2)
abline(lm_cars,col='red',lty=2,lwd=2)
}

par(mfrow=c(1,1),mai = old_mai)
```


Ridge L = 0 – b = −5.3
Ridge L = 0.69 – b = −3.1
Ridge L = 0.89 – b = −2.7
Ridge L = 1.06 – b = −2.5
Ridge L = 1.2 – b = −2.3
Ridge L = 1.33 – b = −2.2
Ridge L = 1.44 – b = −2.1
Ridge L = 1.54 – b = −2
Ridge L = 1.63 – b = −1.9
Ridge L = 1.71 – b = −1.9
Ridge L = 1.79 – b = −1.8
Ridge L = 1.86 – b = −1.8
Ridge L = 1.93 – b = −1.7
Ridge L = 1.99 – b = −1.7
Ridge L = 2.05 – b = −1.7
Ridge L = 2.11 – b = −1.6
Ridge L = 2.16 – b = −1.6
Ridge L = 2.21 – b = −1.6
Ridge L = 2.26 – b = −1.6
Ridge L = 2.3 – b = −1.5

- A medida que λ aumenta, el efecto del peso del vehículo en las millas recorridas por galón disminuye (la pendiente, en valor absoluto, disminuye), hasta el punto que llega a cero aproximadamente.
- La predicción de las millas recorridas por galón se ve cada vez menos afectada por el peso del vehículo.

Esto es importante, pues al reducir el efecto de la covariable en la variable respuesta, se mitiga un posible sobreajuste a los datos.

- ¿Cómo se penalizan los coeficientes cuando contamos con variables cualitativas? ¿Tiene sentido pensar en un ANOVA ridge?



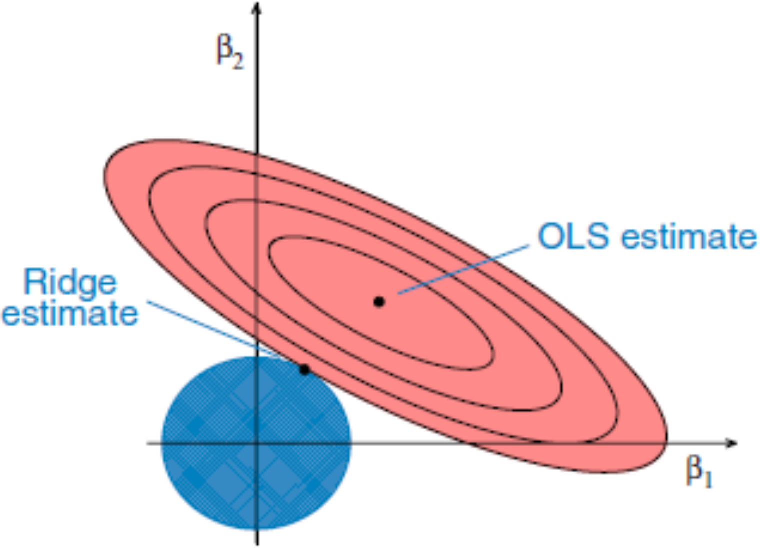
La estimación ridge dada mediante la minimización de SC_{E_R} dada por:

$$\hat{\beta}_R = \text{argmin }_{\beta}\{SC_{E_R}(\beta)\} = \text{argmin }\{\sum_i (y_i - \hat{y})^2 + \lambda \sum_j \beta_j^2\}$$

Puede verse de manera equivalente como una optimización con restricción (note la generalización de Lagrange con desigualdad):

$$\hat{\beta}_R = \text{argmin }_{\beta}\{\sum_i (y_i - \hat{y})^2)\} \text{ restr. } \sum_j \beta_j^2 \leq t$$

Hay una correspondencia 1-1 entre λ y t. Esta representación del problema ridge con restricción permite una representación geométrica del problema de minimización de SC_E para $p = 2$.


β_2
OLS estimate
Ridge estimate
β_1

Bajo el modelo de regresión lineal $Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$, con X igual a la matriz diseño y y el vector de respuestas, se tiene que:

- Estimador sesgado de β es $\hat{\beta} = (X'X)^{-1}Xy$.
- Estimador insesgado de β con sesgo ridge es $\hat{\beta}_R = (X'X + \lambda I)^{-1}Xy$.
- De manera Bayesiana $\beta_i \sim N(0, \sigma^2/\lambda)$, a medida que λ aumenta, penalizamos la varianza de β, que con valor esperado 0, tiende a 0. $E(\beta|(y, \lambda)) = \hat{\beta}_R$.

```
library('glmnet')
lambda_val = c(0,log(seq(2,10,length.out = 19)))

x1 <- mtcars$wt
x2 <- mtcars$disp
y  <- mtcars$mpg


fit_ols   <- lm(y ~ x1 + x2)
fit_ridge <- glmnet(cbind(x1,x2),y,alpha=0,lambda=lambda_val)

round(fit_ols$coefficients,3)
```

```
## (Intercept)          x1          x2
##      34.961      -3.351      -0.018
```

```
round(coef(fit_ridge),3)
```

```
## 3 x 20 sparse Matrix of class "dgCMatrix"
##
## (Intercept) 31.958 32.000 32.045 32.092 32.142 32.196 32.253 32.315 32.382
## x1          -2.445 -2.455 -2.465 -2.476 -2.488 -2.501 -2.514 -2.529 -2.545
## x2          -0.017 -0.017 -0.017 -0.017 -0.018 -0.018 -0.018 -0.018 -0.018
##
## (Intercept) 32.455 32.534 32.621 32.718 32.827 32.952 33.096 33.268 33.478
## x1          -2.562 -2.582 -2.603 -2.627 -2.654 -2.686 -2.724 -2.770 -2.828
## x2          -0.018 -0.018 -0.018 -0.018 -0.018 -0.018 -0.018 -0.018 -0.019
##
## (Intercept) 33.748 34.953
## x1          -2.907 -3.346
## x2          -0.019 -0.018
```

Estimación para un valor determinado de λ

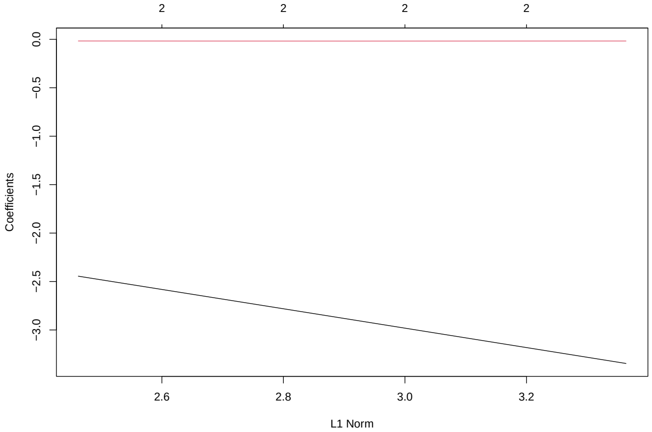
```
coef(fit_ridge,s=0)
```

```
## 3 x 1 sparse Matrix of class "dgCMatrix"
##                      s1
## (Intercept) 34.95294458
## x1          -3.34606887
## x2          -0.01775809
```

Hay una diferencia sutíl con MCO (mínimos cuadrados ordinarios), dado el algoritmo de estimación implementado en glmnet (coordinado descendente).

Y gráficas de las trazas en función de λ

```
plot(fit_ridge)
```



No siempre se tienen suficientes UE para la regresión tradicional (por ejemplo: modelos de cáncer en líneas celulares). Una de las grandes ventajas de la regresión ridge es que no requiere mas datos que parámetros en la recta para ser estimado. En RLS y RLM, se necesita $n \geq p$ para estimar los parámetros de la ecuación. Piense para la regresión tradicional:

- $n = 1$ y $p = 2$.
- $n = 2$ y $p = 2$.
- $n = 2$ y $p = 3$.
- $n = 3$ y $p = 3$.
- $n = p$ y $p = p$.

Para la regresión ridge con $n \leq p$, los puntos no requieren "caer" en el espacio estimado. Basta que minimicen el error en un conjunto de datos de prueba.

Regresión lasso

Es una alternativa al problema de alta varianza en los parámetros de la estimación por mínimos cuadrados. También introduce un sesgo en la estimación, como la regresión ridge, pero penaliza los parámetros de manera diferente:

$$\hat{\beta}_L = \text{argmin }_{\beta}\{SC_{E_L}(\beta)\} = \text{argmin }\{\sum_i (y_i - \hat{y})^2 + \lambda \sum_j |\beta_j|\}$$

Puede verse de manera equivalente como una optimización con restricción (note la generalización de Lagrange con desigualdad):

$$\hat{\beta}_L = \text{argmin }_{\beta}\{\sum_i (y_i - \hat{y})^2)\} \text{ restr. } \sum_j |\beta_j| \leq t$$

Nuevamente hay una correspondencia 1-1 entre λ y t. Esta representación del problema lasso con restricción permite una representación geométrica del problema de minimización de SC_E para $p = 2$.

La única diferencia con la regresión ridge es que la restricción tiene forma de diamante.

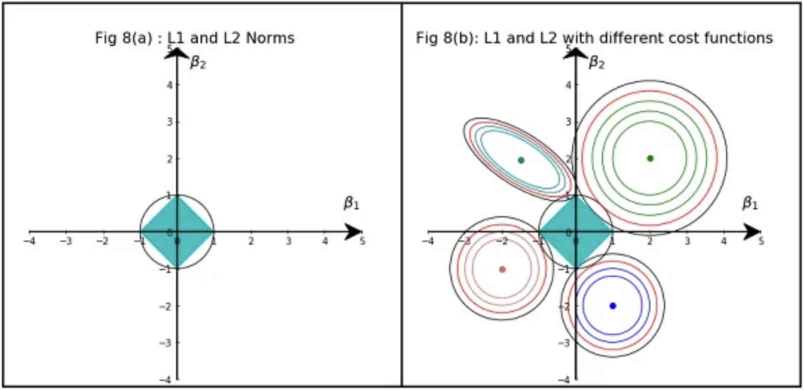
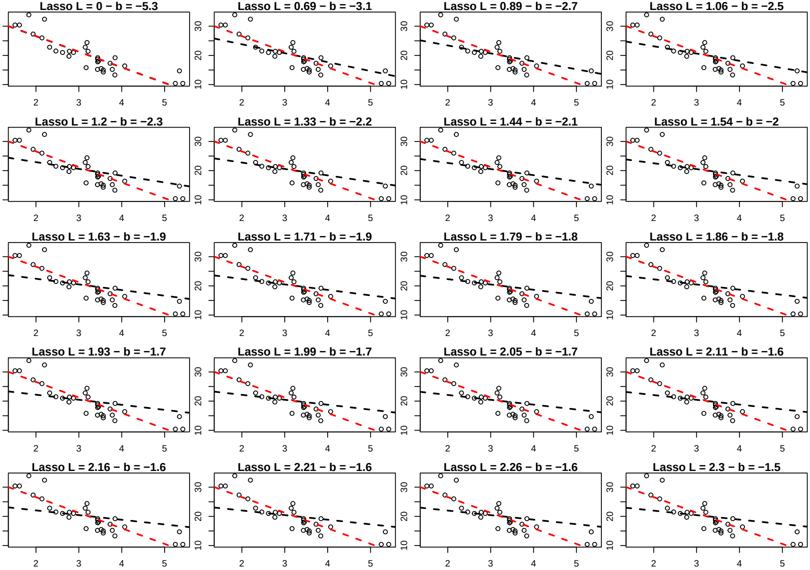


Figure 1: Funciones de costo lasso y ridge


Lasso L = 0 − b = −5.3
Lasso L = 0.69 − b = −3.1
Lasso L = 0.89 − b = −2.7
Lasso L = 1.06 − b = −2.5
Lasso L = 1.2 − b = −2.3
Lasso L = 1.33 − b = −2.2
Lasso L = 1.44 − b = −2.1
Lasso L = 1.54 − b = −2
Lasso L = 1.63 − b = −1.9
Lasso L = 1.71 − b = −1.9
Lasso L = 1.79 − b = −1.8
Lasso L = 1.86 − b = −1.8
Lasso L = 1.93 − b = −1.7
Lasso L = 1.99 − b = −1.7
Lasso L = 2.05 − b = −1.7
Lasso L = 2.11 − b = −1.6
Lasso L = 2.16 − b = −1.6
Lasso L = 2.21 − b = −1.6
Lasso L = 2.26 − b = −1.6
Lasso L = 2.3 − b = −1.5

En resumen:

- Ambos procedimientos hacen que la predicción de Y sea menos sensible a las covariabes $X_1, \ldots, X_p$
- Lasso lleva las pendientes a 0, mientras que ridge las lleva asintóticamente a 0. Esto permite seleccion de variables que bajo el modelo de RLS/RLM son poco importantes, pero sus coeficientes son diferentes a cero. Esto es muy útil cuándo se esperan covariables que por azar inflan el R^2 de RLS/RLM.
- Ridge nunca lleva las pendientes a 0, por lo cual es útil en el escenario en el que se considera que todas las variables son mas o menos informativas.

Extensiones del método

Extensiones del método

- Regresión Elastic Net.
- Grouped Lasso.
- Reducción de dimensionalidad penalizada: NMF.