Métodos de regularización, regresión logística, modelo lineal generalizado y clasificación supervisada

Departamento de Estadística e I.O.
Universitat de València

Problemas regresión lineal múltiple

- Si el número de variables predictoras es grande o hay correlaciones altas entre las variables predictoras.
  - Dificultad en la interpretación del modelo.
  - Colinealidad. Modelo con mucha varianza y sobreajustado.

  Solución. Regularización: Considerar todas las variables predictoras pero forzando que algunos de los parámetros se estimen mediante valores muy próximos a cero (o directamente cero).
- La variable explicada es categórica.
  Solución. Regresión logística o análisis discriminante.

Regresión Ridge

- El objetivo es obtener los coeficientes (β_i) que minimizan:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2,$$

  donde λ se denomina parámetro de sintonía.
- Al igual que en la regresión lineal clásica, el objetivo es minimizar el error en media cuadrático. En este caso el modelo contiene el segundo término, denominado penalización por contracción (es pequeño cuando los coeficientes estan cerca del cero y sirve para controlar el impacto de los coeficientes en la regresión).

Regresión Ridge

- A medida que aumenta λ aumenta el sesgo en las variables, pero disminuye la varianza de las predicciones

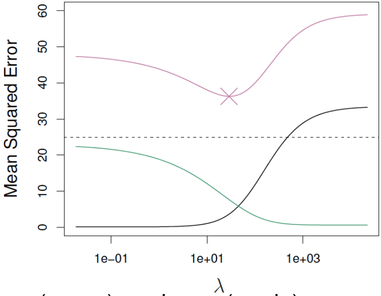


Figura: Sesgo (negro), varianza (verde), y error (rojo)

- Tiene una gran desventaja, utiliza todos los predictores en el modelo final.

Regresión Lasso

- El objetivo es obtener los coeficientes (β_i) que minimizan:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|.$$

- Al utilizar la norma l_1, tiene la ventaja de forzar algunos coeficientes a exactamente cero cuando el parámetro λ es suficientemente grande.
- Problemas: Si $p > n$, Lasso selecciona como máximo n variables. Además, si se tiene un grupo de variables entre las cuales las relaciones a pares son muy altas, Lasso tiende a seleccionar solo una variable de dicho grupo, sin importar cuál se selecciona.

Regresión Elastic net

- El objetivo es obtener los coeficientes (β_i) que minimizan:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{p}\beta_j^2+\alpha\sum_{j=1}^{p}|\beta_j|\right).$$

$0 \leq \alpha \leq 1$.

- Combina las ventajas y evita las desventajas de Ridge y Lasso.

Regresión Ridge, Lasso y Elastic net en R

- Creamos la matrix de diseño X y el vector de valores salario sin los datos no disponibles.

```
library(ISLR)
summary(Hitters)
x <- model.matrix(Salary~., Hitters)[, -1]
y <- Hitters[!is.na(Hitters$Salary),]$Salary
```

- Utilizaremos el paquete glmnet (alpha $\in [0, 1]$, 0: Ridge; 1: Lasso)

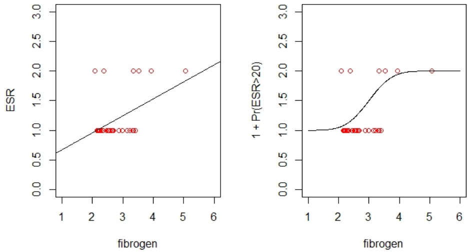
```
library(glmnet)
grid<-10^seq(10,-2,length=100)
ridge.mod <-glmnet(x, y,alpha=0, lambda =grid)
```

- Para encontrar el mejor lambda podemos utilizar la validación cruzada (cv.glmnet)

```
cv.out<-cv.glmnet(x,y,alpha=0,nfolds=6) #por defecto es 10
bestlam<-min(cv.out$lambda)
```

Fracaso del modelo lineal

- El modelo lineal no puede utilizarse aquí. Una fórmula numérica no puede dar por resultado una categoría



- Modelos de clasificación: Estimar/predecir una variable categórica.

Modelos de clasificación

- En general, los métodos de clasificación se basan en la estimación de la probabilidad asociada a cada categoría de la variable categórica.
- Existen diferentes procedimientos de clasificación:
  - Regresión logística
  - Análisis discriminante

Ejemplo

Queremos saber si se ha cometido fraude (Y=Fraude) a partir de datos de la actividad de la tarjeta de crédito (X1=Balance), la renta anual (X2=Income), estatus de estudiante (X3=student (sí/no))

Modelos de clasificación

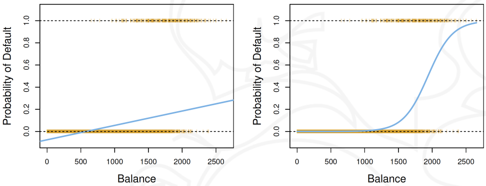
Ejemplo

No vamos a modelizar la variable respuesta Y, sino que vamos a modelizar la distribución de probabilidad

- P(Y=1), P(Y=0)
- P(Y=1|X1), P(Y=0|X1)
- P(Y=1|X2), P(Y=0|X2)
- P(Y=1|X3), P(Y=0|X3)
- P(Y=1|X1,X2), P(Y=0|X1,X2)
- ....

Regresión logística: Transformación logit

- Sea el modelo más sencillo: Y como respuesta binaria y X como variable predictora.
- Denotamos por $p(X) = P(Y = 1|X)$. Queremos un modelo que relacione $p(X)$ con X.
- Primera opción (gráfica de la izquierda): $p(X) = \beta_0 + \beta_1 X$

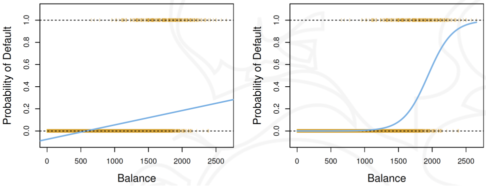


- Para valores de balance cercanos a cero predecimos probabilidades menores que cero!
- Se trata de un problema general: con el modelo lineal, podemos obtener estimaciones de p(X) < 0 para algunos valores de X y p(X) > 1 para otros.

Regresión logística: Transformación logit

- Sea el modelo más sencillo: Y como respuesta binaria y X como variable predictora.
- Denotamos por $p(X) = P(Y = 1|X)$. Queremos un modelo que relacione $p(X)$ con X.
- Segunda opción (gráfica de la derecha): Buscar una función que solo de valores entre 0 y 1 para cualquier X. Sea la función logística (no única)

$$p(X) = \frac{\mathrm{e}^{\beta_0+\beta_1 X}}{1+\mathrm{e}^{\beta_0+\beta_1 X}}$$



- Siempre genera curvas de tipo sigmoidal y captura el rango [0,1].

Regresión logística: Transformación logit

- Sea el modelo más sencillo: Y como respuesta binaria y X como variable predictora.
- Denotamos por $p(X) = P(Y = 1|X)$. Queremos un modelo que relacione $p(X)$ con X.
- El modelo de regresión logística: $(Y|X) \sim \text{Ber}(p(X))$

$$p(X) = \frac{\mathrm{e}^{\beta_0+\beta_1 X}}{1 + \mathrm{e}^{\beta_0+\beta_1 X}}$$

$$\frac{p(X)}{1-p(X)} = \mathrm{e}^{\beta_0+\beta_1 X} \longrightarrow ln\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

- $\frac{p(X)}{1-p(X)}$: odds, valores entre 0 e infinito. Valores cercanos a cero (uno) indican probabilidades muy bajas (altas) de p(X) y muy altas (bajas) de 1-p(X).
- $ln\left(\frac{p(X)}{1-p(X)}\right)$: log-odds o logit.
- Un aumento de una unidad en X implica un aumento de e^{β_1} unidades en la escala de los odds y un aumento de β_1 unidades en la escala logit.

Regresión logística simple

Vamos a estimar los parámetros de la regresión logística

- Como $Y|X \sim \text{Ber}(p(X))$, la función de verosimilitud es

$$\prod_{i=1}^{n} \text{Pr}(y_i|x_i) = \prod_{i:y_i=1} \text{Pr}(x_i) \prod_{i:y_i=0} (1 - \text{Pr}(x_i))$$

$$= \prod_{i:y_i=1} \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \prod_{j:y_j=1} \frac{1}{1 + \exp(\beta_0 + \beta_1 x_j)}$$

- Los estimadores máximo verosímiles de los parámetros β_0 y β_1 son los que maximizan esa expresión.

Regresión logística en R

- La función `glm()`, con la opción `family=binomial` realiza un análisis estadístico utilizando regresión logística
  - Un ejemplo, con los datos `plasma`:

```
ajuste <- glm(ESR ~ fibrinogen, data= plasma, family=
binomial())
summary(ajuste)
```

  - Tiene asociadas las mismas funciones que `lm()`
    - Los coeficientes y sus intervalos de confianza

```
coef(ajuste)
confint(ajuste)
```

    - Predicción de nuevos datos. Por defecto log odds, las probabilidades con `type = 'response'`

```
nuevos <- data.frame(fibrinogen=c(2.2,3.0,4.5))
predict(ajuste, nuevos, se.fit = TRUE)
predict(ajuste, nuevos, type = 'response')
```

Regresión logística en R

A partir de los coeficientes obtenidos, podemos estimar probabilidades asociadas a Y para valores dados de X. En el ejemplo:

```
coef(plasma_glm_1)
(Intercept) fibrinogen
-6.845075 1.827081
```

$$P(Y=1|X=\hat{x}) = \frac{e^{-6{,}845075+1{,}827081\hat{x}}}{1+e^{-6{,}845075+1{,}827081\hat{x}}}.$$

¿Es el ajuste adecuado?

- La desvianza (deviance) es una medida de la bondad del ajuste de un modelo lineal generalizado (sería equivalente a la suma de cuadrados residual de un modelo lineal; valores más altos indican peor ajuste).
- Resolvemos el contraste de hipótesis donde la hipótesis nula es que la deviance es cero (lo que nos interesa para tener un buen ajuste).

```
pchisq(plasma_glm_1$deviance,plasma_glm_1$df.residual,
lower=FALSE)
```

- También podemos hacer el contraste de hipótesis donde la hipótesis nula es que mi modelo sea nulo.

```
anova(plasma_glm_1, test ='Chisq')
```

Regresión logística múltiple

- Y como respuesta binaria y $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ como variables explicativas.
- Formulación: $(Y|\mathbf{X}) \sim \text{Ber}(p(\mathbf{X}))$

$$P(Y = 1|\mathbf{X}) = p(\mathbf{X}), \quad P(Y = 0|\mathbf{X}) = 1 - p(\mathbf{X})$$

$$p(X) = \frac{\mathrm{e}^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + \mathrm{e}^{\beta_0 + \beta_1 X + \cdots + \beta_p X_p}}$$

$$\text{logit}(p(\mathbf{X})) = ln\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X + \cdots + \beta_p X_p$$

Regresión logística múltiple en R

- La utilización de estos modelos con R es sencilla
  - Ejemplo: con los datos `plasma`

```
library('HSAUR3')
ajuste <- glm(ESR ~ fibrinogen*globulin,
                              data = plasma, family=binomial)
summary(ajuste)
```

  - En este caso la interacción es no significativa, podríamos eliminarla del modelo
- La selección de predictores puede realizarse como en regresión múltiple
  - Con medidas de ajuste, utilizando la función `step()`
  - Con medidas de predicción, utilizando validación cruzada

Regresión logística, otro ejemplo

- Datos `Default` en paquete `ISLR`. Objetivo: predecir impagos en tarjetas de crédito
  - Utilizando solo el predictor `student` se deduce que los estudiantes son más propensos a impagos

```
aj1 <- glm(default ~ student, data = Default, family=binomial)
pred1<-predict(aj1,data.frame(student=c('Yes','No')), type =
'response')
```

    - La predicción de impago dado que es estudiante es $0{,}043$. En caso de ser no estudiantes es $0{,}029$
  - Si utilizamos todos los predictores, los estudiantes son menos propensos a impagos (dado que el coeficiente al hacer la regresión es negativo -0.65)

```
aj2<-glm(default ~ .,data=Default,family=binomial)
exp(coef(aj2)['studentYes'])
```

Regresión logística, otro ejemplo

- Calculamos la probabilidad de cometer fraude cuando **balance=1500** e **income=40000** para estudiantes y no estudiantes:

$$P(\text{default} = 1|\text{student} = 1, \text{balance} = 1500, \text{income} = 40000) =$$

$$= \frac{\text{e}^{-10{,}87-0{,}65+0{,}0057\cdot 1500+0{,}000003\cdot 40000}}{1+\text{e}^{-10{,}87-0{,}65+0{,}0057\cdot 1500+0{,}000003\cdot 40000}} = 0{,}058$$

  Para no estudiantes es análogo pero sin el -0.65, obteniendo como resultado 0.105.

- Lo anterior es igual que hacer en R:

```
predict(aj2,data.frame(student=c('No','Yes'),balance=1500,income=40000),
type = 'response')
```

Algunos comentarios

- Siempre, al analizar una variable respuesta, debemos utilizar todos los predictores que creamos que son interesantes. No se deben analizar uno a uno, por separado
  - Si existe algún predictor importante que no hemos utilizado, podemos obtener conclusiones completamente erróneas
  - En el ejemplo `Default`, `balance` es importante, al no tenerlo en cuenta se obtienen conclusiones erróneas
- Regresión Logística también se puede extender a situaciones donde la variable respuesta tenga más de dos categorías. Pero, en esos casos, se utiliza mucho más el análisis discriminante que estudiaremos a continuación.

Modelos Lineales Generalizados

- Regresión logística es un caso particular de los modelos lineales generalizados
- Mediante un función de enlace, link function, relacionan la esperanza de la variable respuesta, $\mathrm{E}(Y)$, con una combinación lineal de los predictores
  - En regresión logística, $g(\mathrm{E}(Y))$ es la transformación logit
  - Si $g(\mathrm{E}(Y))$ es la identidad, tendremos un modelo lineal, que es un caso particular de modelos lineales generalizados
- El equivalente a los residuos son los residuos deviance o desviaciones. Se obtienen a partir de la log-verosimilitud de cada dato.
  - En modelos lineales, residuos deviance y mínimos cuadrados coinciden
- Regresión Poisson es otro modelo lineal generalizado, útil si la variable respuesta puede suponerse Poisson.
  - Incidencia de una enfermedad, por ejemplo

Análisis discriminante

- Sea Y variable categórica con g categorías.
- La regresión logística modeliza directamente la probabilidad de pertenencia a cada grupo: $\Pr(Y = k|\mathbf{X} = \mathbf{x})$
- El análisis discriminante modeliza la distribución de los predictores en cada uno de los grupos, $f(\mathbf{x}|Y = k)$ (función de densidad conjunta de los predictores $\mathbf{X}$ condicionada por $Y = k$)
  - Estima, si no se conoce previamente, las frecuencias de cada grupo en la población, $\pi_k = P(Y = k)$
  - Calcula las probabilidades de clasificación con el Teorema de Bayes:

$$\Pr(Y = k|\mathbf{X} = \mathbf{x}) = \frac{f(\mathbf{x}|Y = k)\pi_k}{\sum_{i=1}^{g} f(\mathbf{x}|Y = i)\pi_i}$$

Análisis discriminante

¿Por qué queremos conocer otro método alternativo a la regresión logística?

- La estimación del modelo de regresión logística es muy inestable cuando las categorías están muy bien separadadas. El análisis discriminante lineal (ADL) no tiene esos inconvenientes.
- Si n es pequeño y la distribución de los predictores X es aproximadamente normal en cada una de las categorías de la variable respuesta Y, el análisis discriminante es más estable que el modelo de regresión logística.
- El análisis discriminante es muy conocido y se usa mucho en el mundo científico cuando el número de categorías de la variable respuesta es mayor que dos.

Análisis lineal discriminante

- El análisis lineal discriminante supone que $\mathbf{X}|Y = k$ son normales multivariantes con distintas medias, pero la misma matriz de covarianzas

$$f(x|Y=k) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)'\Sigma^{-1}(x-\mu_k)\right)$$

$$\exp\left(-\tfrac{1}{2}(x-\mu_k)'\Sigma^{-1}(x-\mu_k)\right) = \exp\left(-\tfrac{1}{2}x'\Sigma^{-1}x\right)\exp\left(x'\Sigma^{-1}\mu_k - \tfrac{1}{2}\mu_k'\Sigma^{-1}\mu_k\right)$$

- Así, las probabilidades de clasificación son:

$$\Pr(Y=k|x) = \frac{\exp\left(x'\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k'\Sigma^{-1}\mu_k\right)\pi_k}{\sum_{i=1}^{g}\exp\left(x'\Sigma^{-1}\mu_i - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i\right)\pi_i}$$

- La clase más probable es la tenga mayor valor en:

$$x'\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k'\Sigma^{-1}\mu_k + \log(\pi_k)$$

Estimación de los parámetros

- Los parámetros desconocidos se sustituyen por sus estimadores puntuales habituales
  - La media μ_k se estima con la media muestral $\bar{x}_k$
  - Si S_k es la matriz de covarianzas muestral del grupo k,

$$\hat{\Sigma} = \frac{1}{n-g}\sum_{i=1}^{g}(n_i - 1)S_i$$

  - Si el banco de datos es una muestra representativa de toda la población, π_k puede estimarse con $\hat{\pi}_k = n_k/n$
- Así, la regla de clasificación se obtiene con

$$\begin{aligned}\delta_k(x) &= x'\hat{\Sigma}^{-1}\bar{x}_k - \frac{1}{2}\bar{x}'_k\hat{\Sigma}^{-1}\bar{x}_k + \log(\hat{\pi}_k) = \\ &= b_{k0} + b_{k1}x_1 + \cdots + b_{kp}x_p\end{aligned}$$

- Y las probabilidades de clasificación:

$$Pr(Y = k|x) = \frac{\exp(\delta_k(x))}{\sum_{i=1}^{g}\exp(\delta_i(x))}$$

Análisis lineal discriminante en R

La base de datos wine de la librería (datasetsICR) contiene el análisis químico de vinos cultivados en la misma región de Italia pero procedentes de 3 cultivares diferentes. Vamos a considerar 5 de los variables consideradas (Class, Alcohol, Malic acid, Ash y Alcalinity of ash)

- En análisis discriminante lineal lo hemos realizado en los comandos.R.

Análisis discriminante cuadrático

- El análisis discriminante cuadrático supone que $\mathbf{X}|Y=k$ son normales multivariantes con distintas medias y con distintas matrices de covarianzas

$$f(x|Y=k) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k)\right)$$

- Así, hay menos factores comunes. La clase más probable es la que tenga mayor valor en:

$$-\frac{1}{2}x'\Sigma_k^{-1}x + x'\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k'\Sigma_k^{-1}\mu_k - \frac{1}{2}\log|\Sigma_k| + \log(\pi_k)$$

- Estimando los parámetros:

$$\delta_k(x) = -\frac{1}{2}x'S_k^{-1}x + x'S_k^{-1}\bar{x}_k - \frac{1}{2}\bar{x}_k'S_k^{-1}\bar{x}_k - \frac{1}{2}\log|S_k| + \log(\hat{\pi}_k)$$

Comparación de los métodos de clasificación

- Regresión logística es la primera opción para clasificar entre dos grupos
  - El análisis lineal discriminante suele proporcionar resultados muy parecidos
  - Si los predictores son normales, el análisis discriminante suele ser algo mejor
  - Si no son normales, puede ser bastante peor que regresión logística
- Análisis discriminante es la primera opción para clasificar entre más de dos grupos
  - Existe regresión logística para más de dos grupos, pero apenas se utiliza