

MODELOS LINEALES

SELECCIÓN DE MODELOS

Fernando Massa; Bruno Bellagamba

martes 23 de abril 2024



FACULTAD DE
CIENCIAS ECONÓMICAS
Y DE ADMINISTRACIÓN

UESTA INSTITUTO
DE ESTADÍSTICA




UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY



- 1 INTRODUCCIÓN
- 2 INDICADORES BASADOS EN $SCRes$
- 3 INDICADORES BASADOS EN LA VEROSIMILITUD
- 4 PROCEDIMIENTOS BASADOS EN PRUEBAS DE HIPÓTESIS
- 5 REGULARIZACIÓN
- 6 VALIDACIÓN CRUZADA
- 7 PRÓXIMA CLASE



- Familiarizarnos con un conjunto de indicadores para determinar qué modelos son “mejores” o “peores”.
- Conocer las herramientas disponibles en  para obtener estos indicadores.
- Pensar en construir una lista o *ranking* de modelos en lugar de obtener el mejor modelo disponible. Esto será de vital importancia cuando entremos a la etapa de diagnóstico.

El uso de modelos lineales no siempre está ligado a la búsqueda del modelo que realice las predicciones más precisas. En ocasiones, los investigadores están interesados en determinar si la relación entre variables que dictamina la teoría se observa o no en la práctica.

El uso de modelos lineales no siempre está ligado a la búsqueda del modelo que realice las predicciones más precisas. En ocasiones, los investigadores están interesados en determinar si la relación entre variables que dictamina la teoría se observa o no en la práctica.

En estos casos, el modelo *RLM* parte de un conjunto de variables explicativas y el análisis se encamina a determinar si estas variables tienen un aporte significativo (en su conjunto y/o de manera individual) sobre la variable de respuesta Y .

El uso de modelos lineales no siempre está ligado a la búsqueda del modelo que realice las predicciones más precisas. En ocasiones, los investigadores están interesados en determinar si la relación entre variables que dictamina la teoría se observa o no en la práctica.

En estos casos, el modelo *RLM* parte de un conjunto de variables explicativas y el análisis se encamina a determinar si estas variables tienen un aporte significativo (en su conjunto y/o de manera individual) sobre la variable de respuesta Y .

No obstante, como se mencionaba al principio, existen ocasiones donde la predicción es el objetivo primordial del modelado y en estos casos, se parte de un conjunto de variables con la idea de seleccionar aquellas que aportan una mayor contribución a la predicción de Y .

En cualquier caso, la selección de un conjunto reducido de variables explicativas es un problema complicado.

- Si consideramos un número **demasiado pequeño**, es posible que se estén omitiendo variables relevantes a la hora de explicar Y , lo cual puede acarrear sesgos en estimaciones y predicciones.
- Si consideramos un número **demasiado grande**, se complica la utilidad práctica del modelo y, aunque mejora el ajuste aparente, aumenta la varianza de los estimadores de los parámetros.

En cualquier caso, la selección de un conjunto reducido de variables explicativas es un problema complicado.

- Si consideramos un número **demasiado pequeño**, es posible que se estén omitiendo variables relevantes a la hora de explicar Y , lo cual puede acarrear sesgos en estimaciones y predicciones.
- Si consideramos un número **demasiado grande**, se complica la utilidad práctica del modelo y, aunque mejora el ajuste aparente, aumenta la varianza de los estimadores de los parámetros.

Decidir el mejor conjunto de variables es prácticamente un arte. Por lo cual, presentaremos algunos indicadores y procedimientos que nos asistan a la hora de comparar modelos con distintas variables explicativas.

En clases anteriores definimos el R^2 como un indicador que nos brinda una idea de la *bondad de ajuste* de un modelo.

$$R^2 = 1 - \frac{SCRes}{SCTotal} = \frac{SCExp}{SCTotal} = \frac{\sum_i (\hat{Y} - \bar{Y})^2}{\sum_i (Y - \bar{Y})^2}$$

Este indicador tiene la virtud de que, de manera sencilla, nos permite interpretar cual es la proporción de la variabilidad de Y que es explicada por el conjunto de variables explicativas en la matriz X .

En clases anteriores definimos el R^2 como un indicador que nos brinda una idea de la *bondad de ajuste* de un modelo.

$$R^2 = 1 - \frac{SCRes}{SCTotal} = \frac{SCExp}{SCTotal} = \frac{\sum_i (\hat{Y} - \bar{Y})^2}{\sum_i (Y - \bar{Y})^2}$$

Este indicador tiene la virtud de que, de manera sencilla, nos permite interpretar cual es la proporción de la variabilidad de Y que es explicada por el conjunto de variables explicativas en la matriz X .

No obstante, hace unas clases, vimos como el valor del indicador, esta fuertemente influenciado por el número de variables incluidas en el modelo. Ya que el agregar variables explicativas, la $SCExp$ jamás decrece. Por este motivo, este indicador **NO debe utilizarse para comparar modelos con distinta cantidad de variables explicativas.**

Una forma de solventar este inconveniente es trabajar con una versión *ajustada* de este indicador, donde se penaliza por la incorporación de variables explicativas.

$$R_{aj}^2 = 1 - \frac{SCRes/(n-k-1)}{SCTotal/(n-1)}$$

El precio que pagamos es que se pierde la interpretación que tiene el R^2 y que en casos extremos, es posible que $R_{aj}^2 < 0$.

No obstante, a partir de este indicador es posible comparar modelos de una manera más *justa* con el criterio de que valores más altos (cercanos a 1) indican un mejor ajuste con **menos** variables.

Estos indicadores son una forma alternativa de solucionar el problema que vimos con el R^2 . En la medida que se añaden variables al modelo, SCR_{es} disminuye, por lo tanto, la log-verosimilitud se incrementa.

Estos indicadores son una forma alternativa de solucionar el problema que vimos con el R^2 . En la medida que se añaden variables al modelo, $SCRes$ disminuye, por lo tanto, la log-verosimilitud se incrementa.

$$\begin{aligned}\ell(\hat{\beta}, \hat{\sigma}^2) &= -\frac{n}{2} \log(2\pi\hat{\sigma}^2) - \frac{1}{2} \sum_i \frac{(y_i - x_i' \hat{\beta})^2}{\hat{\sigma}^2} \\ \ell(\hat{\beta}, \hat{\sigma}^2) &= -\frac{n}{2} \log(2\pi SCRes / (n - k - 1)) - \frac{1}{2} \frac{SCRes}{SCRes / (n - k - 1)} \\ \ell(\hat{\beta}, \hat{\sigma}^2) &= -\frac{n}{2} \log(2\pi SCRes / (n - k - 1)) - \frac{1}{2} (n - k - 1)\end{aligned}$$

Estos indicadores son una forma alternativa de solucionar el problema que vimos con el R^2 . En la medida que se añaden variables al modelo, $SCRes$ disminuye, por lo tanto, la log-verosimilitud se incrementa.

$$\begin{aligned}\ell(\hat{\beta}, \hat{\sigma}^2) &= -\frac{n}{2} \log(2\pi\hat{\sigma}^2) - \frac{1}{2} \sum_i \frac{(y_i - x_i' \hat{\beta})^2}{\hat{\sigma}^2} \\ \ell(\hat{\beta}, \hat{\sigma}^2) &= -\frac{n}{2} \log(2\pi SCRes / (n - k - 1)) - \frac{1}{2} \frac{SCRes}{SCRes / (n - k - 1)} \\ \ell(\hat{\beta}, \hat{\sigma}^2) &= -\frac{n}{2} \log(2\pi SCRes / (n - k - 1)) - \frac{1}{2} (n - k - 1)\end{aligned}$$

Por este motivo, se proponen una variedad de *criterios de información* que **penalizan** la verosimilitud por el número de variables. Los más conocidos son el *AIC* y el *BIC*.

- *AIC*: $-2\ell(\hat{\beta}, \hat{\sigma}^2) + 2k$.
- *BIC*: $-2\ell(\hat{\beta}, \hat{\sigma}^2) + \log(n)k$

Estos indicadores son una forma alternativa de solucionar el problema que vimos con el R^2 . En la medida que se añaden variables al modelo, $SCRes$ disminuye, por lo tanto, la log-verosimilitud se incrementa.

$$\begin{aligned}\ell(\hat{\beta}, \hat{\sigma}^2) &= -\frac{n}{2} \log(2\pi\hat{\sigma}^2) - \frac{1}{2} \sum_i \frac{(y_i - x_i' \hat{\beta})^2}{\hat{\sigma}^2} \\ \ell(\hat{\beta}, \hat{\sigma}^2) &= -\frac{n}{2} \log(2\pi SCRes / (n - k - 1)) - \frac{1}{2} \frac{SCRes}{SCRes / (n - k - 1)} \\ \ell(\hat{\beta}, \hat{\sigma}^2) &= -\frac{n}{2} \log(2\pi SCRes / (n - k - 1)) - \frac{1}{2} (n - k - 1)\end{aligned}$$


Por este motivo, se proponen una variedad de *criterios de información* que **penalizan** la verosimilitud por el número de variables. Los más conocidos son el *AIC* y el *BIC*.

- *AIC*: $-2\ell(\hat{\beta}, \hat{\sigma}^2) + 2k$.
- *BIC*: $-2\ell(\hat{\beta}, \hat{\sigma}^2) + \log(n)k$

Al utilizar estos indicadores, es preferible obtener **valores más bajos**.

En el siguiente ejemplo trataremos de seleccionar las variables más relevantes para explicar la tasa estandarizada de fertilidad en 47 provincias franco-parlantes de Suiza a finales del siglo XIX en base a 5 variables socioeconómicas.



Vayamos al 

PROCEDIMIENTOS BASADOS EN PRUEBAS DE HIPÓTESIS

La enumeración de todos los posibles modelos se puede hacer en tanto se disponga de pocas variables explicativas. Si no se cuenta el modelo nulo y se denomina por k el número de variables explicativas.

$$N_{\text{modelos}}^{\circ} = 2^k - 1$$

PROCEDIMIENTOS BASADOS EN PRUEBAS DE HIPÓTESIS

La enumeración de todos los posibles modelos se puede hacer en tanto se disponga de pocas variables explicativas. Si no se cuenta el modelo nulo y se denomina por k el número de variables explicativas.

$$N_{modelos}^o = 2^k - 1$$

Si se dispone de 10 variables, el número de modelos a ajustar es 1023...

PROCEDIMIENTOS BASADOS EN PRUEBAS DE HIPÓTESIS

La enumeración de todos los posibles modelos se puede hacer en tanto se disponga de pocas variables explicativas. Si no se cuenta el modelo nulo y se denomina por k el número de variables explicativas.

$$N_{modelos}^o = 2^k - 1$$

Si se dispone de 10 variables, el número de modelos a ajustar es 1023...

Por este motivo surgen algunos procedimientos que buscan *dirigir* la búsqueda de forma de no tener que recorrer todo el abanico de posibles modelos.

Los más comunes son:

- Backward.
- Forward.
- Stepwise.

Estos procedimientos son de carácter iterativo. En particular, el método *Backward* conlleva los siguientes pasos.

Estos procedimientos son de carácter iterativo. En particular, el método *Backward* conlleva los siguientes pasos.

- 1 En primer lugar se fija un nivel de significancia α_0 .

Estos procedimientos son de carácter iterativo. En particular, el método *Backward* conlleva los siguientes pasos.

- 1 En primer lugar se fija un nivel de significancia α_0 .
- 2 Se ajusta el modelo con todas las variables explicativas y se realiza la prueba de significación parcial de cada variable explicativa registrando el conjunto de p-valores $\{pv_j\}_j$.

Estos procedimientos son de carácter iterativo. En particular, el método *Backward* conlleva los siguientes pasos.

- 1 En primer lugar se fija un nivel de significancia α_0 .
- 2 Se ajusta el modelo con todas las variables explicativas y se realiza la prueba de significación parcial de cada variable explicativa registrando el conjunto de p-valores $\{pv_j\}_j$.
 - Si todos los p-valores son menores a α_0 ($pv_j \leq \alpha_0 \forall j$) \Rightarrow el proceso concluye.
 - Si uno o más p-valores son mayores a α_0 ($\max\{pv_j\} \geq \alpha_0$) \Rightarrow se remueve la variable con el mayor p-valor, se vuelve a ajustar el modelo y se vuelve al paso anterior.

Estos procedimientos son de carácter iterativo. En particular, el método *Backward* conlleva los siguientes pasos.

- 1 En primer lugar se fija un nivel de significancia α_0 .
- 2 Se ajusta el modelo con todas las variables explicativas y se realiza la prueba de significación parcial de cada variable explicativa registrando el conjunto de p-valores $\{pv_j\}_j$.
 - Si todos los p-valores son menores a α_0 ($pv_j \leq \alpha_0 \forall j$) \Rightarrow el proceso concluye.
 - Si uno o más p-valores son mayores a α_0 ($\max\{pv_j\} \geq \alpha_0$) \Rightarrow se remueve la variable con el mayor p-valor, se vuelve a ajustar el modelo y se vuelve al paso anterior.
- 3 El método concluye cuando todas las variables explicativas remanentes son significativas al nivel α_0 .

Este método es similar al anterior, pero procede de manera inversa.

- 1 En primer lugar se fija un nivel de significancia α_0 .

Este método es similar al anterior, pero procede de manera inversa.

- 1 En primer lugar se fija un nivel de significancia α_0 .
- 2 El método comienza agregando al modelo la variable que mejor explica a la variable de respuesta según algún criterio (por ejemplo el R^2).

Este método es similar al anterior, pero procede de manera inversa.

- ➊ En primer lugar se fija un nivel de significancia α_0 .
- ➋ El método comienza agregando al modelo la variable que mejor explica a la variable de respuesta según algún criterio (por ejemplo el R^2).
- ➌ Se lleva a cabo la significación de dicha variable.
 - Si su aporte es significativo, el método continúa.
 - Si su aporte no fuese significativo, el modelo queda vacío.

Este método es similar al anterior, pero procede de manera inversa.

- ➊ En primer lugar se fija un nivel de significancia α_0 .
- ➋ El método comienza agregando al modelo la variable que mejor explica a la variable de respuesta según algún criterio (por ejemplo el R^2).
- ➌ Se lleva a cabo la significación de dicha variable.
 - Si su aporte es significativo, el método continúa.
 - Si su aporte no fuese significativo, el modelo queda vacío.
- ➍ Se añade (por separado) cada variable explicativa maximizando el valor del indicador elegido, teniendo en cuenta las variables presentes en el modelo. Una vez seleccionada la variable de mayor aporte, se realiza la significación parcial.

Si $p\text{-valor} \leq \alpha_0 \Rightarrow$ se incorpora al modelo y se vuelve a iterar.

Si $p\text{-valor} \geq \alpha_0 \Rightarrow$ el método concluye.

Se trata de una combinación de los anteriores. La idea es que (en el procedimiento Forward) es posible que la incorporación de variables conlleve la pérdida de significación de variables incorporadas en pasos anteriores.

Se trata de una combinación de los anteriores. La idea es que (en el procedimiento Forward) es posible que la incorporación de variables conlleve la pérdida de significación de variables incorporadas en pasos anteriores.

Para esto, es necesario emplear un criterio de *entrada* y un criterio de *salida*.

- El criterio de **entrada** es que la variable seleccionada para ingresar al modelo debe cumplir que $p - \text{valor} \leq \alpha_{\text{entra}}$.
- El criterio de **salida** es que la variable seleccionada para salir del modelo debe cumplir que $p - \text{valor} \geq \alpha_{\text{sale}}$.

Se trata de una combinación de los anteriores. La idea es que (en el procedimiento Forward) es posible que la incorporación de variables conlleve la pérdida de significación de variables incorporadas en pasos anteriores.

Para esto, es necesario emplear un criterio de *entrada* y un criterio de *salida*.

- El criterio de **entrada** es que la variable seleccionada para ingresar al modelo debe cumplir que $p - \text{valor} \leq \alpha_{\text{entra}}$.
- El criterio de **salida** es que la variable seleccionada para salir del modelo debe cumplir que $p - \text{valor} \geq \alpha_{\text{sale}}$.

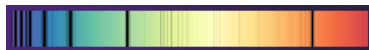
En la práctica suele elegirse $\alpha_{\text{entra}} \leq \alpha_{\text{sale}}$ para evitar situaciones en las que el método entre en un bucle infinito.



En este caso empleamos un conjunto de datos donde se desea obtener un método para predecir el contenido graso de una muestra de carne mediante espectroscopia. Para ello se dispone del contenido de grasa de 215 muestras y la absorbancia en 100 frecuencias del espectro en longitudes de onda cercanas al infrarrojo.




En este caso empleamos un conjunto de datos donde se desea obtener un método para predecir el contenido graso de una muestra de carne mediante espectroscopia. Para ello se dispone del contenido de grasa de 215 muestras y la absorbancia en 100 frecuencias del espectro en longitudes de onda cercanas al infrarrojo.





En este caso empleamos un conjunto de datos donde se desea obtener un método para predecir el contenido graso de una muestra de carne mediante espectroscopia. Para ello se dispone del contenido de grasa de 215 muestras y la absorbancia en 100 frecuencias del espectro en longitudes de onda cercanas al infrarrojo.



Vayamos al  y veamos como es el desempeño de estos métodos.

Cuando en la próxima clase comencemos a estudiar la etapa de diagnóstico, veremos que cuando tenemos *muchas* variables explicativas en el modelo, pueden surgir problemas de multicolinealidad.

Cuando en la próxima clase comencemos a estudiar la etapa de diagnóstico, veremos que cuando tenemos *muchas* variables explicativas en el modelo, pueden surgir problemas de multicolinealidad.

Un signo de que estamos ante un caso donde esto sucede es cuando tenemos un modelo con un valor del R^2 muy alto pero ninguna de las variables explicativas tiene un aporte significativo.

Lo que está sucediendo en estos casos es que una (o varias) de las variables explicativas son *casi* combinación lineal de algunas de las demás.

Cuando en la próxima clase comencemos a estudiar la etapa de diagnóstico, veremos que cuando tenemos *muchas* variables explicativas en el modelo, pueden surgir problemas de multicolinealidad.

Un signo de que estamos ante un caso donde esto sucede es cuando tenemos un modelo con un valor del R^2 muy alto pero ninguna de las variables explicativas tiene un aporte significativo.

Lo que está sucediendo en estos casos es que una (o varias) de las variables explicativas son *casi* combinación lineal de algunas de las demás.

Esto se traduce en que:

- Los estimadores y sus varianzas se vuelven inestables.
- Suelen adoptar valores exageradamente grandes.
- Lo cual conlleva a que las pruebas de hipótesis llevadas a cabo a partir de estos, sean poco confiables.

Lo que sucede es que la matriz $X'X$ es *casi* singular.

Lo que sucede es que la matriz $X'X$ es *casi* singular.

Una primera manera de solventar este problema consiste en agregar un cierto valor positivo λ a la diagonal de esta matriz, de forma de asegurar su invertibilidad. De esta manera, el estimador del vector de coeficientes es:

$$\hat{\beta}_{ridge} = (X'X + \lambda I_n)^{-1} X'Y$$

Este estimador es el estimador de la regresión *ridge* y surge del siguiente problema de minimización.

$$\min_{\beta} (Y - X\beta)'(Y - X\beta) + \lambda \beta' \beta$$

Lo que sucede es que la matriz $X'X$ es *casi* singular.

Una primera manera de solventar este problema consiste en agregar un cierto valor positivo λ a la diagonal de esta matriz, de forma de asegurar su invertibilidad. De esta manera, el estimador del vector de coeficientes es:

$$\hat{\beta}_{ridge} = (X'X + \lambda I_n)^{-1} X'Y$$

Este estimador es el estimador de la regresión *ridge* y surge del siguiente problema de minimización.

$$\min_{\beta} (Y - X\beta)'(Y - X\beta) + \lambda \beta' \beta$$

Nótese como se trata de la misma función objetivo que se minimiza en el método de mínimos cuadrados, pero con el agregado de que se está penalizando por el *tamaño* de los coeficientes.

La principal característica de este estimador es que

$$\mathbb{E}(\hat{\beta}_{ridge}) \neq \beta$$

Siendo esto característico de todos los modelos donde se incluye una penalización sobre la función a minimizar.

La principal característica de este estimador es que

$$\mathbb{E}(\hat{\beta}_{ridge}) \neq \beta$$

Siendo esto característico de todos los modelos donde se incluye una penalización sobre la función a minimizar.

No obstante, no necesariamente es algo malo. Esto sucede ya que al aumentar el sesgo de un estimador, disminuye su varianza y viceversa ($ECM = sesgo^2 + varianza$).

$$\begin{aligned} ECM(\hat{\beta}_{ridge}) &= \mathbb{E} \left(\hat{\beta}_{ridge} - \beta \right)^2 \\ &= \underbrace{\left[\mathbb{E} \left(\hat{\beta}_{ridge} - \beta \right) \right]^2}_{Sesgo^2(\hat{\beta}_{ridge})} + \underbrace{\mathbb{E} \left(\hat{\beta}_{ridge} - \mathbb{E}(\hat{\beta}_{ridge}) \right)^2}_{Var(\hat{\beta}_{ridge})} \end{aligned}$$

Entonces, ya estamos dispuestos a utilizar un método que produce estimadores sesgados
¿cuál es el beneficio?

Entonces, ya estamos dispuestos a utilizar un método que produce estimadores sesgados ¿cuál es el beneficio?

SELECCIONAR VARIABLES

La idea fundamental de este método es que si bien todos los componentes del vector $\hat{\beta}$ se *encojen* al aumentar el valor de λ , aquellos correspondientes a variables explicativas irrelevantes, lo hacen más rápido.

Entonces, ya estamos dispuestos a utilizar un método que produce estimadores sesgados ¿cuál es el beneficio?

SELECCIONAR VARIABLES

La idea fundamental de este método es que si bien todos los componentes del vector $\hat{\beta}$ se *encojen* al aumentar el valor de λ , aquellos correspondientes a variables explicativas irrelevantes, lo hacen más rápido.

Adicionalmente, debido a que estamos relajando los criterios para buscar estimadores, es posible que obtengamos predicciones con menor error cuadrático medio que las que obtenemos con *MCO*.

SELECCIÓN DEL PARÁMETRO λ

El problema que resta por resolver es la selección del parámetro de penalización.

SELECCIÓN DEL PARÁMETRO λ

El problema que resta por resolver es la selección del parámetro de penalización.

Si bien existen teoremas que garantizan la existencia de un valor óptimo, no hay métodos que nos permitan estimar de manera simultánea todos los parámetros del modelo (incluido λ).

SELECCIÓN DEL PARÁMETRO λ

El problema que resta por resolver es la selección del parámetro de penalización.

Si bien existen teoremas que garantizan la existencia de un valor óptimo, no hay métodos que nos permitan estimar de manera simultánea todos los parámetros del modelo (incluido λ).

En la práctica se suelen emplear métodos computacionales como *validación cruzada*

VALIDACIÓN CRUZADA

El conjunto de datos se particiona en m subconjuntos de forma tal que la evaluación del modelo procede realizando las predicciones de cada subconjunto usando un modelo ajustado con todas las observaciones **excepto** las de dicho subconjunto.

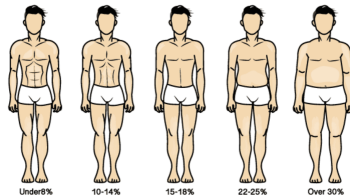
Al final de este proceso, el modelo es evaluado con algún indicador (por ejemplo el *ECM*).

Este proceso se repite para un conjunto de valores de λ y se selecciona el valor que produce el menor valor del indicador.

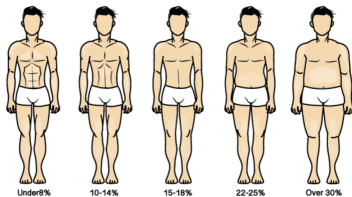


Ahora pasamos a un conjunto de datos donde interesa predecir el porcentaje de grasa corporal a partir de un conjunto de medidas corporales.

Ahora pasamos a un conjunto de datos donde interesa predecir el porcentaje de grasa corporal a partir de un conjunto de medidas corporales.



Ahora pasamos a un conjunto de datos donde interesa predecir el porcentaje de grasa corporal a partir de un conjunto de medidas corporales.



Vayamos al  y seleccionemos variables.

El último método que veremos hoy es una modificación del anterior en la medida de que se cambia la penalización. El método LASSO (**L**east **A**bsolute **S**hrinkage and **S**election **O**perator).

$$\min_{\beta} (Y - X\beta)' (Y - X\beta) + \lambda \sum_j |\beta_j|$$

El cual puede ser reformulado en la forma:

$$\begin{aligned} \min_{\beta} (Y - X\beta)' (Y - X\beta) \\ \text{s.a.} \quad \sum_j |\beta_j| \leq \lambda \end{aligned}$$

COMPARACIÓN RIDGO VS LASSO

En un problema donde solo se cuente con 2 variables explicativas, el problema de minimización se podría visualizar de la siguiente manera.

COMPARACIÓN RIDGE VS LASSO

En un problema donde solo se cuente con 2 variables explicativas, el problema de minimización se podría visualizar de la siguiente manera.

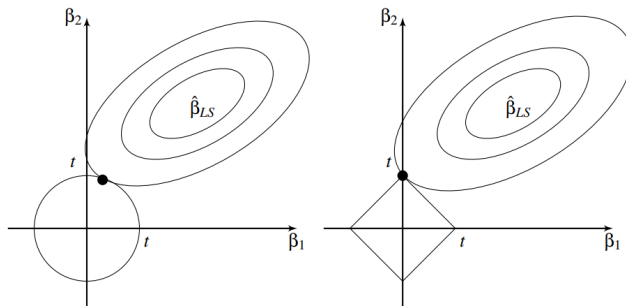


Figure 11.9 Ridge and lasso regression are illustrated. On the left, confidence ellipses of increasing level are plotted around the least squares estimate. The largest ellipse intersects the circle of radius t at the ridge estimate. On the right, the largest ellipse intersects the square at the lasso estimate.

La principal diferencia entre estos métodos se debe al tipo de penalización.

- Debido que en **ridge** la penalización es cuadrática, los coeficientes tienden a cero al aumentar la penalización, pero nunca llegan 0.
- En el caso de **LASSO**, como la penalización se da en términos absolutos, la solución del problema de minimización se da en uno de los vértices del hipercubo definido por la restricción $\sum_j |\beta_j| \leq \lambda$.

La principal diferencia entre estos métodos se debe al tipo de penalización.

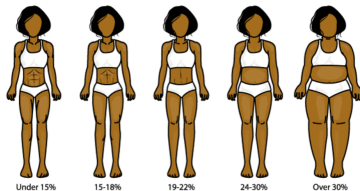
- Debido que en **ridge** la penalización es cuadrática, los coeficientes tienden a cero al aumentar la penalización, pero nunca llegan 0.
- En el caso de **LASSO**, como la penalización se da en términos absolutos, la solución del problema de minimización se da en uno de los vértices del hipercubo definido por la restricción $\sum_j |\beta_j| \leq \lambda$.

En los vértices de dicho hipercubo, algunos coeficientes tienen valor 0, por lo tanto, la penalización selecciona variables.

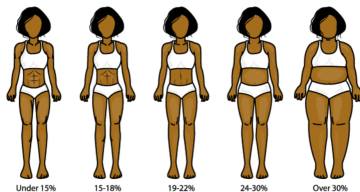



Volvamos al ejemplo de la grasa corporal pero en esta oportunidad, utilicemos LASSO.

Volvamos al ejemplo de la grasa corporal pero en esta oportunidad, utilicemos LASSO.



Volvamos al ejemplo de la grasa corporal pero en esta oportunidad, utilicemos LASSO.



Vayamos al  y (ahora si) seleccionemos variables.



La próxima hablaremos de:

- Comenzaremos a hablar de la etapa de diagnóstico.
- El primer aspecto a evaluar será la multicolinealidad.
- El segundo aspecto a evaluar será la homoscedasticidad.
- El tercer aspecto a evaluar será la normalidad.



Carmona, Francesc (2003). *Modelos Lineales (notas de curso)*. Departament d'Estadística.



Faraway, Julian (2014). *Linear Models with R, second edition*. Chapman Hall/CRC.



Rencher, Alvin y Bruce Schaalje (2008). *Linear Models in Statistics, second edition*. John Wiley Sons, Inc.

¿Preguntas?

Muchas Gracias