EST-25134: Aprendizaje Estadístico

Profesor: Alfredo Garbuno Iñigo — Primavera, 2022 — Métodos de selección.
Objetivo: Detalles en métodos de selección de variables. Veremos las estrategias de regularización y penalización para ajustar modelos controlando el sesgo predictivo hacia el conjunto de entrenamiento. Utilizaremos validación cruzada para probar configuraciones y elegir la *mejor*. Hablaremos sobre reducción de dimensiones y su combinación con métodos predictivos.
Lectura recomendada: Capítulo 6 de [1]. Sección 6.4 de [2]. Aunque el enfoque es regresión los principios de validación cruzada para escoger modelos penalizados en el contexto de clasificación son análogos. Puedes leer la sección 12.5 de [2].

1. INTRODUCCIÓN

Ya hemos visto cómo cuantificar el `error de generalización` en un proceso de aprendizaje. Es decir, cuantificar los `errores de predicción` sobre nuevas observaciones y, además, `cuantificar la variabilidad` de esta predicción. En esta parte estudiaremos cómo incorporar lo aprendido para `escoger un modelo sobre otro`.

Por ejemplo, hemos visto que es natural `extender` el modelo lineal

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon\,. \tag{1}$$

Veremos (mas adelante) la idea de incorporar relaciones `no lineales` manteniendo el supuesto de `aditividad`.

Incluso aunque el modelo lineal es sencillo, tiene sus ventajas pues nos ayuda a tener un modelo `interpretable` y al mismo tiempo con buena `capacidad predictiva`.

El libro de Kuhn and Johnson [2] tiene una buena discusión sobre las ventajas algorítmicas de un modelo lineal. Usualmente en la práctica queremos tener nuestro modelo en un ambiente productivo. Lo cual necesita que las predicciones sean fácilmente calculables. ¿Qué pasaría si en la plataforma de Netflix o Amazon se tarda mucho en aparecer las sugerencias? Los modelos lineales son fácilmente calculables en prácticamente cualquier ambiente productivo.

Estudiaremos estrategias para mejorar modelos lineales a través de procedimientos alternativos de ajuste.

1.1. Opciones para ajustar modelos

- Basados en `precisión de ajuste`, ideal cuando $p > n$ con el objetivo de *reducir* varianza.
- Basados en `interpretabilidad`. Por ejemplo, eliminar variables que no tengan capacidad predictiva.

2. ESTRATEGIAS DE SELECCIÓN DE VARIABLES

- Selección por subconjuntos.
- Reducción de coeficientes (regularización).
- Reducción de dimensiones.

2.1. Selección por subconjuntos

El mecanismo sería el siguiente.

1. Utilizar el `modelo nulo` $\mathcal{M}_0$ (sin predictores).
2. Para $k = 1, \ldots, p$:
   a) Ajustar todos modelos posibles con k predictores.
   b) Elegir el *mejor* de esa colección de modelos, le pondremos $\mathcal{M}_k$.
3. Elegir el mejor modelo dentro de la colección $\mathcal{M}_0, \ldots, \mathcal{M}_p$ utilizando un `criterio de comparación de modelos`.

2.1.1. Para pensar: ¿Por qué no puedes utilizar el criterio de `RSS` para escoger entre las opciones $\mathcal{M}_1, \ldots, \mathcal{M}_p$?

2.2. En el contexto de clasificación

La `devianza` –el negativo de dos veces la log-verosimilitud– se utiliza como una métrica de bondad de ajuste (como el `RSS`) para una clase mas amplia de modelos.

2.3. Selección iterativa

Podemos elegir empezar con el modelo mas sencillo e ir incorporando una variable a la vez mas predictores. En cada paso podemos evaluar la `mejora adicional` de haber incorporado estas nuevas características.

2.4. Pseudo-código (selección hacia adelante)

El proceso sería el siguiente.

1. Denotamos por $\mathcal{M}_0$ el modelo `nulo`.
2. Para $k = 0, \ldots, p-1$:
   a) Considera todos los $p - k$ modelos que aumentan el modelo en la iteración anterior $\mathcal{M}_k$ con un predictor adicional.
   b) Escoge el `mejor` de estos $p - k$ modelos y llámale $\mathcal{M}_{k+1}$.
3. Escoge el mejor de los modelos entre $\mathcal{M}_0, \ldots, \mathcal{M}_p$ utilizando un criterio de comparación de modelos.

2.5. Aplicación: créditos

```
  Income Limit Rating Cards Age Education Gender Student Married Balance
1     32  6375    469     3  25        16 Female      No     Yes    1120
2     28  5619    418     2  78        15 Female      No     Yes     822
3     64  7530    515     1  56        12   Male      No     Yes    1086
4     14  2330    203     5  80        16 Female      No      No       0
5     26  5640    398     3  58        15 Female      No      No     905
```

LISTING 1. *Muestra de datos del conjunto `Credit`.*

El objetivo es predecir `Saldo` utilizando las demás características. El ejemplo de [1] ha implementado la búsqueda por subconjuntos y la búsqueda iterativa hacia adelante. Estos son los mejores modelos encontrados.

> Nota que el mecanismo iterativo no tiene garantía de encontrar el mejor modelo dentro de las $\binom{p}{k}$ posibilidades.

# Variables	Best subset	Forward stepwise
One	`rating`	`rating`
Two	`rating`, `income`	`rating`, `income`
Three	`rating`, `income`, `student`	`rating`, `income`, `student`
Four	`cards`, `income` `student`, `limit`	`rating`, `income`, `student`, `limit`

FIGURA 1. *Método de selección para los datos de créditos. Tomada de [1].*

```
   estrategia sigma r.squared adj.r.squared  AIC deviance
1 subconjunto   100      0.95          0.95 4823  3915058
2    adelante   101      0.95          0.95 4835  4032502
```

LISTING 2. *Métricas de bondad de ajuste para los datos de `Credit`.*

2.5.1. Para pensar: ¿Cuántos modelos en total se ajustan con el procedimiento de búsqueda iterativa hacia adelante? Considera $p = 20$.

2.6. Selección iterativa hacia atrás

Empezamos con el `modelo completo` que contenga los p predictores. Eliminando variables, una a la vez, cuando un predictor no sea tan útil. La única restricción que necesitamos es que $n > p$.

3. MÉTRICAS DE DESEMPEÑO

Si utilizáramos el `RSS` para comparar entre $\mathcal{M}_0, \ldots, \mathcal{M}_k$ tendríamos un problema pues eliminar (aumentar) predictores siempre perjudicaría (beneficia) la capacidad predictiva del modelo. Necesitamos `compensar` por el sesgo de sobre-ajuste. Es decir, considerar una métrica que pueda estimar el error de `generalización`.

3.1. C_p de Mallow

Es un criterio de bondad de ajuste (`menor mejor`) definida como

$$C_p(\mathcal{M}_d) = \frac{1}{n} \left(\mathsf{RSS}(d) + 2d\hat{\sigma}^2 \right) . \tag{2}$$

> Tenemos una penalización a la suma de residuales al cuadrado (`RSS`) que considera un aumento en predictores utilizados.

3.2. El criterio de información de Akaike (AIC)

Se utiliza para evaluar modelos ajustados por máxima verosimilitud (`menor mejor`)

$$\mathsf{AIC}(\mathcal{M}_d) = -2 \log L + 2d \, . \tag{3}$$

3.2.1. Ejercicio: Prueba que en el caso del modelo lineal con errores Gaussianos el criterio de mínimos cuadrados y máxima verosimilitud es el mismo. Además los criterios C_p y AIC son lo mismo.

3.3. R^2 ajustada

Se calcula como

$$R^2_A(\mathcal{M}_d) = 1 - \frac{\mathsf{RSS}/(n-d-1)}{\mathsf{TSS}/(n-1)}\,. \tag{4}$$

Es una métrica de correlación entre predicción ($\hat{y}$) y respuesta (y) (`mayor mejor`). Al contrario de la R^2 tradicional esta métrica si se afecta por la inclusión de variables inecesarias/redundantes.

3.4. Objetivo

Cada uno de los procedimientos de selección de variables regresa una secuencia de modelos $\mathcal{M}_k$. Lo que queremos es escoger la $k^\star$ de acuerdo al `error de generalización`. El error de generalización obtenido por `validación cruzada` tiene la ventaja de no hacer la estimación de σ^2.

> Estimar σ^2 es una tarea complicada. Implica, bajo el modelo de regresión, estimar el mejor modelo y encontrar la precisión de la familia de modelos que estamos utilizando.

3.4.1. Selección de modelo: Datos de crédito El objetivo es predecir el `Saldo` en términos de los demás predictores. Se seleccionarán las variables de acuerdo a un proceso iterativo. En este caso por `búsqueda hacia adelante`.

`Funciones a utilizar`:

```
train ← analysis(split)
valid ← assessment(split)
```

Listing 3. *Separación de muestras*

```
modelo.nulo     ← lm(Balance ~ 1, train)
modelo.completo ← lm(Balance ~ ., train)
```

Listing 4. *Ajuste de modelos `esquina`.*

```
adelante ← step(modelo.nulo,
                direction='forward',
                scope=formula(modelo.completo),
                trace=0)
predictores ← attributes(adelante$terms)$term.labels
```

Listing 5. *Instrucción de ajuste iterativo.*

Escogemos el modelo con el error mas pequeño. Sin embargo, validación cruzada nos puede dar una métrica de incertidumbre (¿cuál?). ¿Y si el problema de decisión lo planteamos como una prueba de hipótesis?

4. REGULARIZACIÓN

Los procedimientos `selección de variables discretos/iterativos` pueden generar una `varianza` muy alta en las estimaciones del error y podría no reducir el error de predicción del modelo completo. Estudiaremos dos métodos de regularización, `Ridge` y `LASSO`, donde ajustamos un modelo con todas las características *penalizando* de alguna manera la complejidad del modelo.

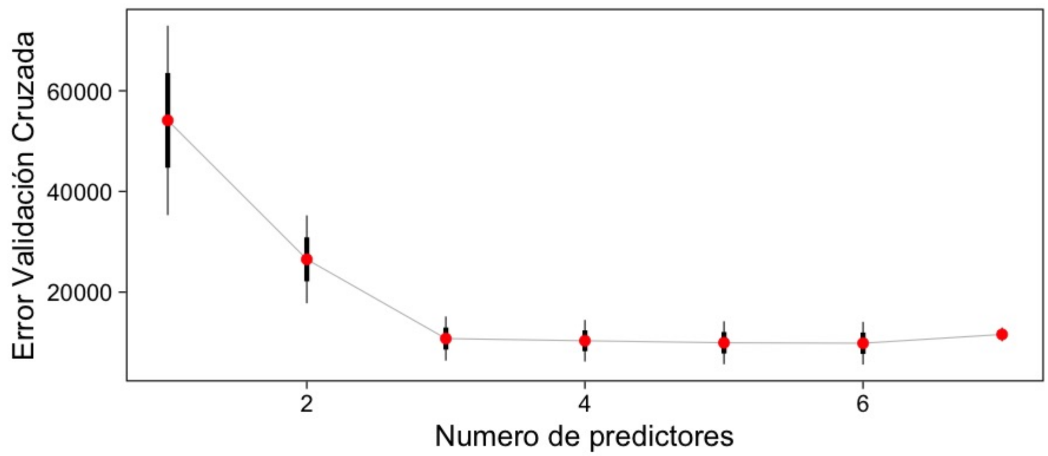

FIGURA 2. *Error de generalización estimado por validación cruzada con $K = 10$. Para los datos de* `Credit`.

4.1. Regresión Ridge

Nuestra formulación anterior consideraba encontrar $\beta_0, \beta_1, \ldots, \beta_n$ minimizando

$$\mathsf{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2 . \tag{5}$$

Lo que haremos ahora será incorporar un `término de penalización` en la función objetivo

$$\mathsf{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2 , \tag{6}$$

donde $\lambda \geq 0$ es un `hiper-parámetro`.

El objetivo sigue siendo el mismo, ajustar el modelo lo mejor posible. El término adicional favorece soluciones con $\beta_1, \ldots, \beta_p$ pequeños. El parámetro λ controla qué tanto `penalizamos` el *tamaño* de los coeficientes.

4.1.1. Para pensar: Un valor muy pequeño para λ implica una penalización `pequeña`, por lo tanto la solución tenderá a ser un modelo `altamente flexible`. Por otro lado un valor de λ grande implica una penalización `fuerte`. Esto se traduce en un solución `poco flexible`.

4.2. Ridge: datos de crédito

Usaremos Ridge como mecanismo de reducción de coeficientes para ajustar modelos parsimoniosos.

Observa que conforme `aumenta la penalización` los `coeficientes disminuyen` gradualmente.

> Al penalizar sobre los coeficientes necesitamos que todos *platiquen* en el mismo idioma. Es por esto que tenemos que estandarizar los predictores. Si queremos estimar el error de generalización métodos de separación de muestras, ¿en qué momento lo hacemos? Es decir, ¿antes de separar los datos o en cada paso del proceso de ajuste?

`Instrucciones útiles`:

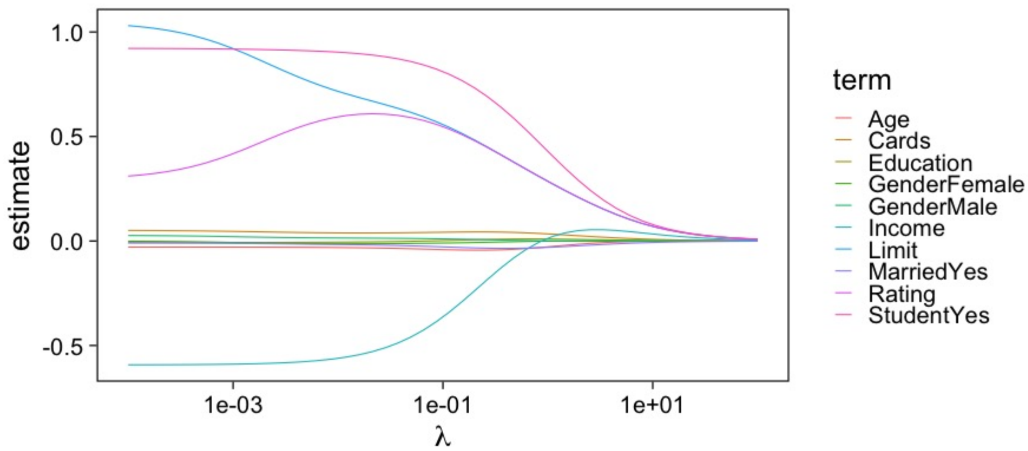

FIGURA 3. *Trayectorias de los coeficientes al aumentar la penalización λ.*

```
## Preparo el objetivo del modelo
rec ← recipe(respuesta ~ ., data = train)
## Defino procesamiento de datos
estandarizador ← rec ▷
  step_normalize(Income, Limit, Rating, Cards, Age, Education, respuesta)
```

LISTING 6. *Definir procesamiento de normalización con* `recipes`.

```
## Calculo medias y desviaciones estandar en entrenamiento
estandarizador.ajustado ← prep(estandarizador, train)
## Normalizo ambos conjuntos
valid.std ← bake(estandarizador.ajustado, valid)
train.std ← bake(estandarizador.ajustado, train)
list(train = train.std, valid = valid.std)
```

LISTING 7. *Aplicación del proceso de normalización.*

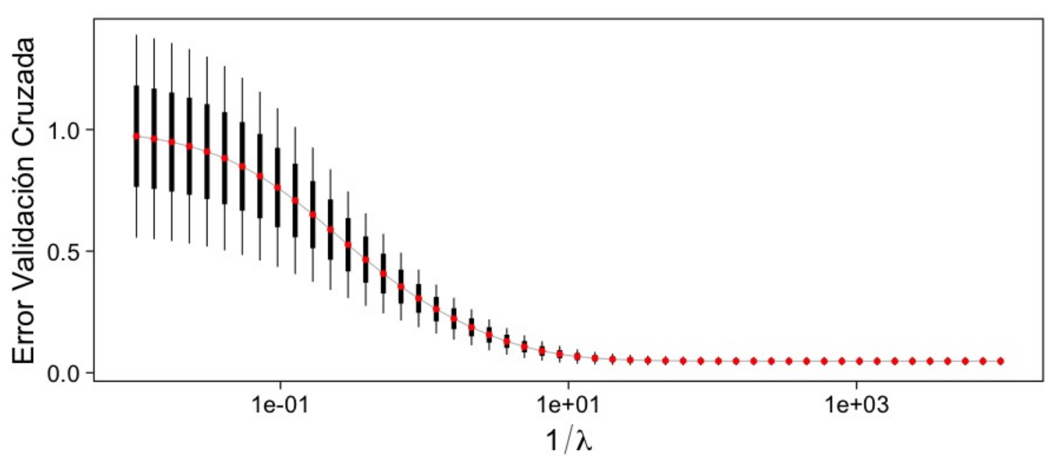

FIGURA 4. *Error de validación calculada con $K = 10$. Nota que graficamos contra $1/\lambda$.*

Con validación cruzada podemos identificar qué valor de λ es el adecuado para penalizar. Una vez realizada esta elección, re-entrenamos el modelo utilizando `todo` el conjunto de datos para predecir situaciones/observaciones futuras.

4.3. Regresión LASSO

En la práctica Ridge no elimina completamente los predictores. Podemos cambiar la penalización para incorporar un `término de penalización` en la función objetivo

$$\mathsf{RSS} + \lambda \sum_{j=1}^{p} |\beta_j| \,, \tag{7}$$

donde $\lambda \geq 0$ es un `hiper-parámetro`.

Igual que antes... el objetivo sigue siendo el mismo, ajustar el modelo lo mejor posible. El término adicional favorece soluciones con $\beta_1, \ldots, \beta_p$ pequeños. El parámetro λ controla qué tanto penalizamos el *tamaño* de los coeficientes.

4.4. LASSO: datos de crédito

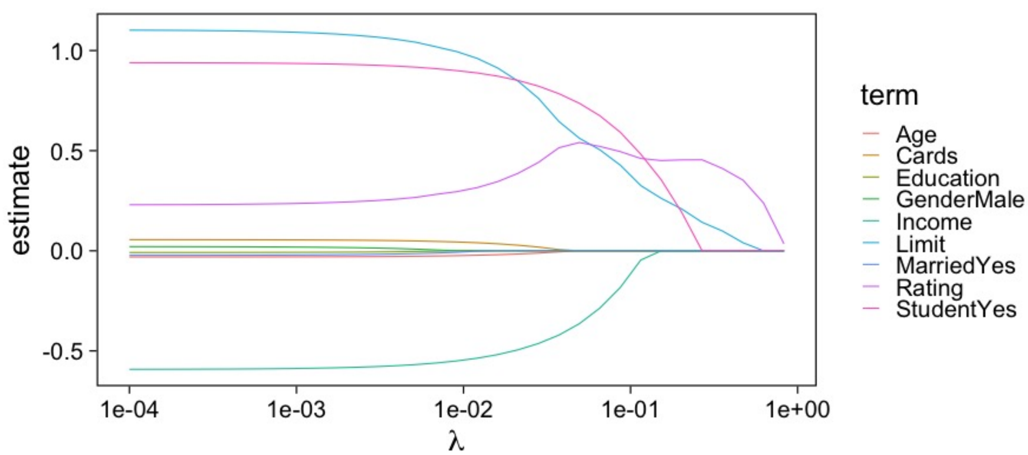

FIGURA 5. *Trayectorias de los coeficientes al aumentar la penalización λ.*

Observa que LASSO tiene la propiedad de eliminar completamente los predictores ($\beta = 0$) por lo que es un mecanismo de `selección automática de variables`.

4.5. Comparación: Ridge v. LASSO

El problema de optimización (Ridge) se puede reescribir de la siguiente manera

$$\text{minimizar } \mathsf{RSS}, \qquad \text{sujeto a} \sum_{j=1}^{p} \beta_j^2 \leq s \,, \tag{8}$$

y el respectivo de LASSO

$$\text{minimizar } \mathsf{RSS}, \qquad \text{sujeto a} \sum_{j=1}^{p} |\beta_j| \leq s \,. \tag{9}$$

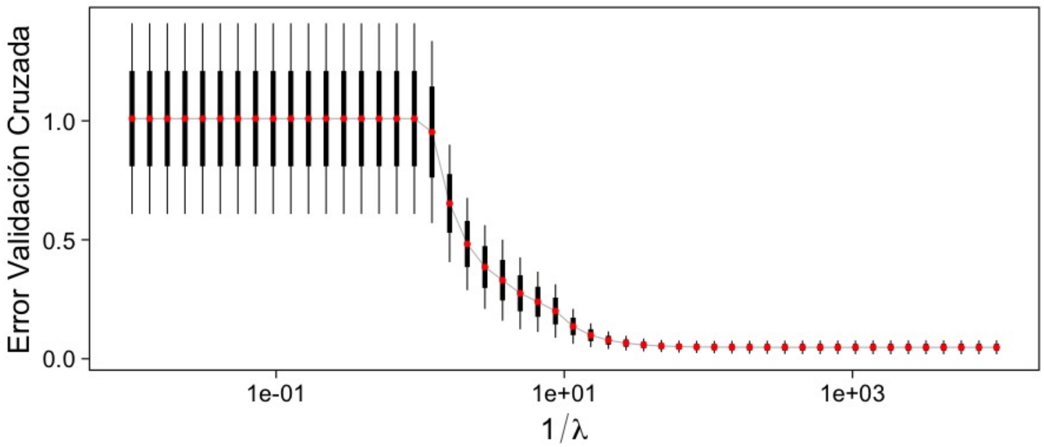

FIGURA 6. *Error de validación calculada con $K = 10$. Nota que graficamos contra $1/\lambda$.*

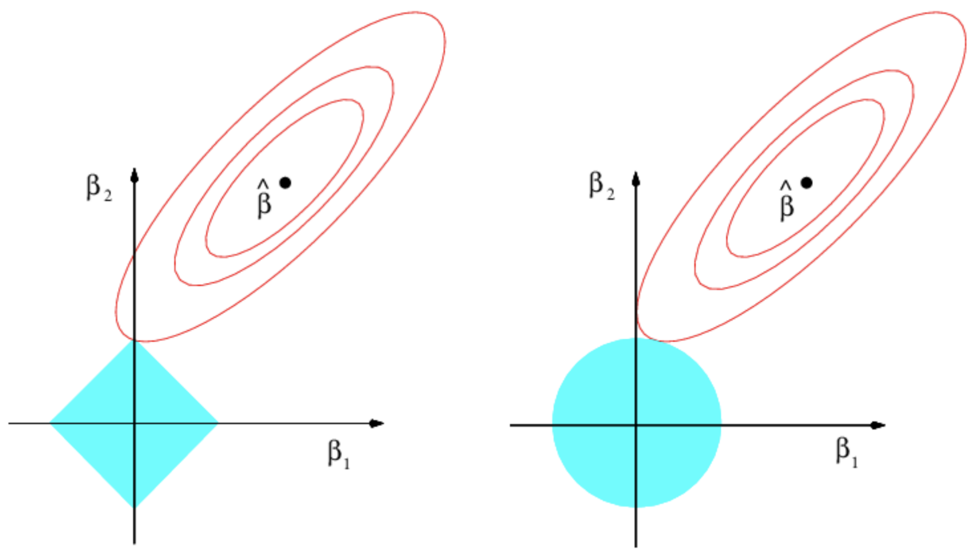

FIGURA 7. *Curvas de nivel de los problemas de optimización. Tomada de [1].*

4.6. Conclusiones

En la práctica no hay una estrategia dominante. LASSO podría ser preferido cuando el número de parámetros es pequeño. Pero eso implica conocer *a priori* el número de predictores para usar en el modelo.

4.6.1. Para pensar: ¿cómo escogerías entre Ridge o LASSO?

5. MÉTODOS DE REDUCCIÓN DE DIMENSIONES

LASSO o Ridge utilizan el concepto de regularización para `restringir` los modelos posibles. Una alternativa es `transformar` primero los predictores (el espacio de los predictores) y `ajustar` un modelo con ese subespacio.

5.1. Regresión con reducción de dimensiones

Denotemos por $Z_1, Z_2, \dots Z_M$ combinaciones lineales de nuestros predictores originales. Lo escribimos como

$$Z_m = \sum_{j=1}^{p} \phi_{mj} X_j, \qquad m = 1, \dots, M\,, \tag{10}$$

con algunas constantes ϕ_{mj} (que se escogen con alguna estrategia).

Podemos ajustar un modelo de regresión por medio de

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i\,, \tag{11}$$

utilizando mínimos cuadrados.

Nota que podemos rescribir

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{mj} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij}\,, \tag{12}$$

donde

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{mj}\,. \tag{13}$$

El modelo restringe automáticamente las β_j pues tienen que tomar una forma muy particular. Si las ϕ_{mj} se escogen bien, incluso pueden realizar un mejor trabajo que el modelo de mínimos cuadrados en las variables originales.

5.2. Otros métodos de reducción de dimensiones

- Utilizar componentes principales (`varianza máxima` entre `predictores`).
- Utilizar *partial least squares* (`varianza máxima` entre `predictores y respuesta`).
- Utilizar *least angle regression* (trayectoria de `contribución lineal predictiva` de atributos).

REFERENCIAS

[1] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R.* Springer Texts in Statistics. Springer US, New York, NY, 2021. ISBN 978-1-07-161417-4 978-1-07-161418-1. . 1, 2, 3, 8

[2] M. Kuhn and K. Johnson. *Applied Predictive Modeling.* Springer New York, New York, NY, 2013. ISBN 978-1-4614-6848-6 978-1-4614-6849-3. . 1