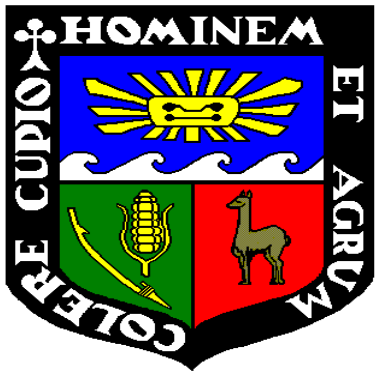
Técnicas Multivariadas Avanzadas

Selección de Modelos

Ms Carlos López de Castilla Vásquez

Universidad Nacional Agraria La Molina

2014-2



Introducción

- El modelo de regresión lineal:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \in$$

- A pesar de su simplicidad el modelo anterior tiene ventajas distintivas en términos de su *interpretabilidad*.
- En muchas ocasiones un modelo lineal presenta un buen *comportamiento predictivo*.
- En este capítulo se discuten algunos métodos que permiten mejorar el modelo lineal reemplazando el método de mínimos cuadrados ordinarios por algún otro método alternativo.
- Más adelante se consideran también modelos *no lineales*.

¿Por que considerar alternativas a MCO?

- *Precisión en las observaciones*: si la relación entre la variable respuesta y los predictores es lineal, las estimaciones por MCO tienen poco sesgo y variancia cuando $n > p$. Por otro lado, si $p > n$ los coeficientes por MCO tienen variancia infinita.
- *Interpretabilidad del modelo*: es usual que algunas de las variables usadas en un modelo de regresión múltiple no se encuentren asociadas con la variable respuesta. Incluir estas variables irrelevantes aumentan de forma innecesaria la complejidad del modelo resultante.
- Se presentan algunas metodologías para la reducción de la variancia y el proceso automático de selección de variables (*feature selection*).

Tres clases de métodos

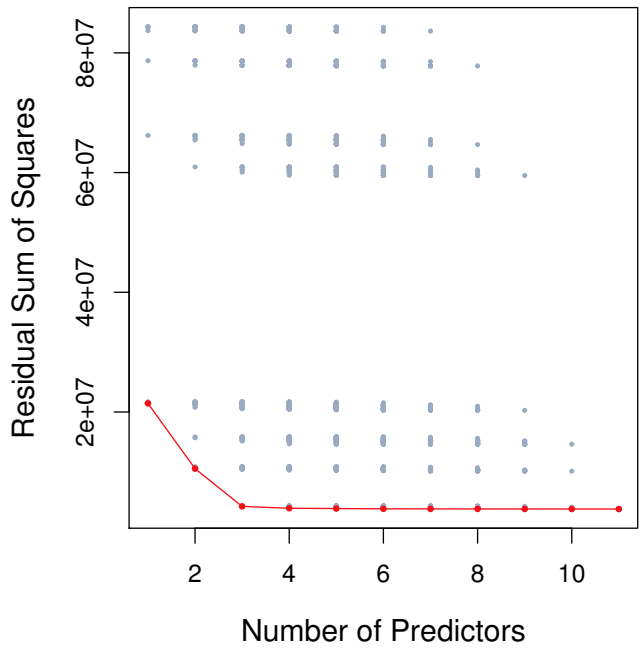
- *Subset selection*: Se identifica un subconjunto de p predictores que se encuentran relacionados con la variable respuesta. El modelo se estima con MCO.
- *Shrinkage*: Se estima el modelo con los p predictores pero los coeficientes estimados son encogidos hacia cero en relación con las estimaciones por MCO. Este encogimiento (llamado también regularización) tiene el efecto de reducir la variancia.
- *Dimension reduction*: Se proyectan los p predictores en un subespacio M-dimensional donde $M < p$, calculando M combinaciones lineales, o proyecciones, de las variables. Luego estas proyecciones son usadas como predictores para estimar un modelo de regresión lineal por MCO.

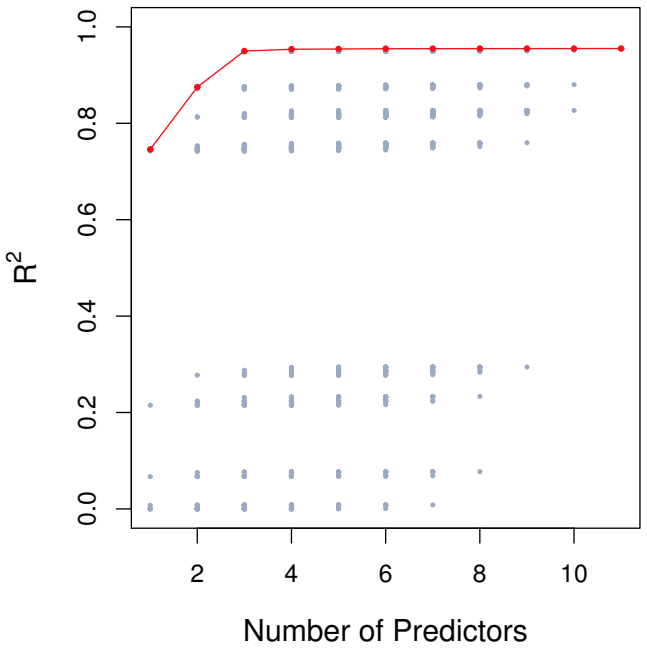
Best Subset Selection

1. Sea $\mathcal{M}_0$ que denota el *modelo nulo*, aquel que no tiene predictores.
2. Para $k = 1, 2, \cdots, p$:
   - 2.1 Estimar los $\binom{p}{k}$ modelos que contienen exactamente k predictores.
   - 2.2 Tomar el mejor de los modelos anteriores, llamado $\mathcal{M}_k$. El mejor modelo es aquel que tiene el menor RSS o equivalentemente el mayor R^2.
3. Seleccionar el mejor modelo a partir de $\mathcal{M}_0, \mathcal{M}_1, \cdots, \mathcal{M}_p$ usando el error de predicción por validación cruzada, C_p, AIC, BIC o R^2 ajustado.

Data Automóvil

- La línea en color rojo define el mejor modelo para un determinado número de predictores de acuerdo al RSS y R^2.





Extensiones hacia otros modelos

- A pesar que se presento *best subset selection* para la regresión por MCO, se pueden aplicar las mismas ideas a otro tipo de modelos como la regresión logística.
- En el caso de la regresión logística en lugar de ordenar los modelos por RSS se usa la *deviancia*, una medida que juega el mismo rol de RSS para una clase más amplia de modelos.
- La deviancia es igual a menos dos veces el logaritmo de la verosimilitud maximizada.
- Mientras menor sea la deviancia el modelo se ajusta mejor a los datos.

Stepwise Selection

- Por razones computacionales *best subset selection* no puede ser aplicado cuando p es grande.
- Cuanto mayor es el espacio de búsqueda mayor es también la posibilidad de encontrar modelos aparentemente satisfactorios en la data de entrenamiento pero que quizás no tengan ningún poder predictivo en la data de prueba.
- Con un espacio de búsqueda grande podría caerse en el problema de *sobreestimación* y tener una gran variancia en la estimación de los coeficientes.
- Por estas razones los métodos *stepwise*, que exploran un conjunto restringido de modelos, son alternativas atractivas al *best subset selection*.

Forward Stepwise Selection

1. Sea $\mathcal{M}_0$ que denota el *modelo nulo*, aquel que no tiene predictores.
2. Para $k = 0, 1, \cdots, p-1$:
   - 2.1 Considere todos los $p-k$ modelos que aumentan en uno los predictores en $\mathcal{M}_k$.
   - 2.2 Escoger el mejor modelo entre los $p-k$ y llamarlo $\mathcal{M}_{k+1}$. El mejor modelo es aquel que tiene el menor RSS o equivalentemente el mayor R^2.
3. Seleccionar el mejor modelo a partir de $\mathcal{M}_0, \mathcal{M}_1, \cdots, \mathcal{M}_p$ usando el error de predicción por validación cruzada, C_p, AIC, BIC o R^2 ajustado.

Forward Stepwise Selection

- Se tiene una ventaja computacional clara sobre *best subset selection*.
- A diferencia de *best subset selection* que requiere estimar 2^p modelos, f*orward stepwiese selection* requiere solamente $1 + p(p+1)/2$ modelos.
- Lo anterior representa una diferencia sustancial: cuando $p = 20$ *best subset selection* requiere estimar 1048576 modelos mientras que *forward stepwiese selection* requiere estimar solo 211 modelos.
- Sin embargo no se garantiza que se encuentre el mejor modelo posible fuera de los 2^p modelos que contienen subconjuntos de los predictores.

Backward Stepwise Selection

- *Backward stepwise selection* también proporciona una alternativa eficiente sobre *best subset selection*.
- A diferencia de *forward stepwiese selection* se empieza con un modelo completo que contiene los p predictores y luego se va eliminando de forma iterativa el predictor menos útil.
- Así como en *forward stepwiese selection* este método busca entre $1 + p(p+1)/2$ modelos y puede ser aplicado cuando p es grande. Sin embargo tampoco se garantiza que se encuentre el mejor modelo.
- *Backward stepwiese selection* requiere que $n > p$, lo cual no representa una restricción para *forward stepwiese selection* que puede aplicarse aun cuando $n < p$.

Backward Stepwise Selection

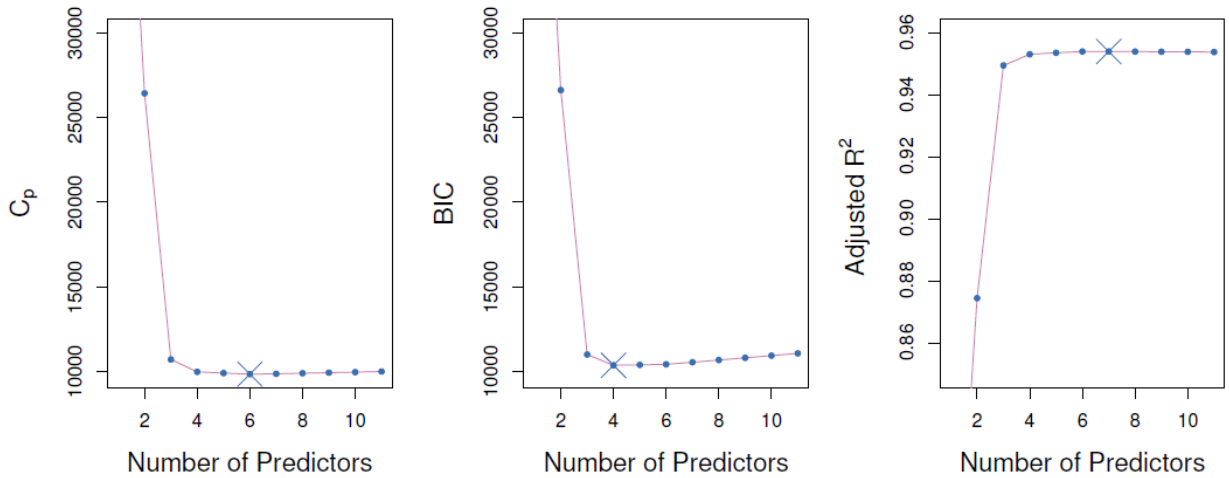
1. Sea $\mathcal{M}_p$ que denota el *modelo completo*, aquel que contiene los p predictores.
2. Para $k = p, p-1, \cdots, 1$:
   - 2.1 Considere todos los k modelos que contienen todos los predictores menos uno en $\mathcal{M}_k$.
   - 2.2 Escoger el mejor modelo entre los k y llamarlo $\mathcal{M}_{k-1}$. El mejor modelo es aquel que tiene el menor RSS o equivalentemente el mayor R^2.
3. Seleccionar el mejor modelo a partir de $\mathcal{M}_0, \mathcal{M}_1, \cdots, \mathcal{M}_p$ usando el error de predicción por validación cruzada, C_p, AIC, BIC o R^2 ajustado.

Eligiendo el modelo óptimo

- El modelo completo siempre es aquel que tiene el menor RSS y el mayor R^2 y se trata de indicadores relacionados con el error de entrenamiento.
- Se desea elegir un modelo con menor error de prueba y, como ya se menciono previamente, el error de entrenamiento es un estimador deficiente.
- Por esta razón R^2 y RSS no son indicadores apropiados para seleccionar entre un conjunto de modelos con diferente número de predictores.
- Para estimar el error de prueba *indirectamente* se puede realizar un ajuste al error de entrenamiento para medir el sesgo debido al sobreajuste. Es posible estimar el error de prueba *directamente* usando validación cruzada.

C_p, AIC, BIC y R^2ajustado

- Estos indicadores ajustan el error de entrenamiento de acuerdo al tamaño del modelo y pueden ser usados para elegir entre un conjunto de modelos con diferente número de variables.
- El siguiente gráfico muestra C_p, BIC y R^2 ajustado para el mejor modelo producido por *best subset selection* usando la data Credit.



C_p, AIC, BIC y R^2ajustado

- C_p de Mallow:

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

donde d es el número total de parámetros usados y $\hat{\sigma}^2$ es una estimación de la variancia del error.

- El criterio de información de Akaike es definido para un modelo estimado por máxima verosimilitud:

$$AIC = -2\log L + 2d$$

donde L es el valor maximizado de la función de verosimilitud.

- En el caso de un modelo lineal con errores gaussianos C_p y AIC son equivalentes.

C_p, AIC, BIC y R^2ajustado

$$BIC = \frac{1}{n}(RSS + d\hat{\sigma}^2 \log n)$$

- De la misma forma que C_p, BIC tiende a tomar un menor valor para un modelo con un bajo error de prueba y por consiguiente se elige el modelo que tiene el menor valor para este indicador.
- BIC reemplaza $2d\hat{\sigma}^2$ usado en C_p por $d\hat{\sigma}^2 \log n$ donde n es el número de observaciones.
- Como $\log n > 2$ para $n > 7$ BIC aplica una penalidad más severa sobre modelos con muchas variables por lo que permite elegir modelos más pequeños que C_p.

C_p, AIC, BIC y R^2ajustado

- Para un modelo con d variables se define R^2 ajustado por:

$$R^2aj = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

- A diferencia de los otros indicadores un valor grande de R^2aj indica un modelo con menor error de prueba.
- El R^2 ajustado paga un precio por la inclusión de variables irrelevantes en el modelo.

Validación y validación cruzada

- Los procedimientos anteriores brindan una secuencia de modelos $\mathcal{M}_k$ para $k = 0, 1, \cdots$. Se requiere hallar $\hat{k}$ para poder obtener $\mathcal{M}_{\hat{k}}$.
- Se calcula el error por *validación cruzada* para cada modelo en $\mathcal{M}_k$. Luego se elige el valor de k para el cual se obtuvo el menor error.
- Este procedimiento permite estimar directamente el error de prueba y tiene una ventaja relativa con respecto de C_p, AIC, BIC o R^2 ajustado ya que no requiere estimar σ^2.

Ejemplo data Credit

- Los errores de validación fueron calculados seleccionando al azar tres cuartos de las observaciones como data de entrenamiento y el cuarto restante como conjunto de validación.
- Los errores por validación cruzada fueron calculados usando $K = 10$. En ambos casos los métodos llevan a considerar un modelo con seis variables.
- Sin embargo los tres procedimientos anteriores sugieren que los modelos con 4, 5 y 6 variables son equivalentes en términos de sus errores de prueba.

Introducción

- Los métodos de selección de subconjuntos usan la técnica de mínimos cuadrados ordinarios para estimar modelos lineales.
- Como alternativa es posible estimar un modelo con p predictores usando una técnica que *restringe* o *regulariza* las estimaciones de los coeficientes. o equivalentemente que *encogen* las estimaciones hacia cero.
- La razón por la que una restricción mejora el ajuste esta en que el encogimiento de los coeficientes podría reducir significativamente su varianza.

Regresión Ridge

- El método de MCO estima los parámetros minimizando:

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

- Los coeficientes estimados para la regresión *ridge* son los valores que minimizan:

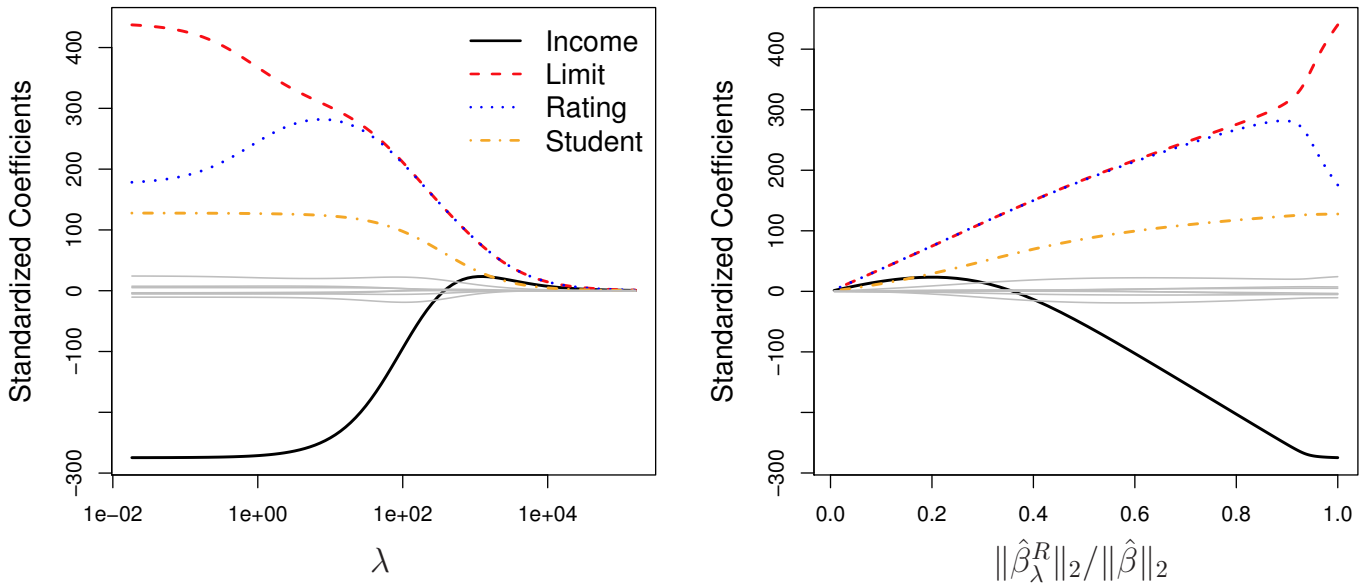
$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

donde $\lambda \geq 0$ es un *parámetro de sintonización* que debe ser determinado de forma separada.

Regresión Ridge

- Al igual que el método de mínimos cuadrados ordinarios la regresión ridge busca las estimaciones de los coeficientes minimizando RSS.
- Sin embargo el segundo término, llamado *penalidad por encogimiento*, es pequeño cuando los coeficientes se encuentran cerca de cero.
- El parámetro de sintonización controla el impacto sobre los coeficientes estimados.
- Elegir un buen valor para λ es crítico. Puede usarse validación cruzada.

Regresión Ridge



- La notación $||\beta||_2 = \sqrt{\sum_{i=1}^{p} \beta_j^2}$ denota la norma l_2.

Regresión Ridge

- Las estimaciones por MCO son *equivariantes en escala*, es decir si se multiplica X_j por una constante c su coeficiente estimado queda dividido por c. En otras palabras $X_j\hat{\beta}_j$ no cambia su valor.
- En contraste los coeficientes estimados con la regresión ridge pueden cambiar sustancialmente cuando se multiplica un predictor por una constante debido al término de penalidad por encogimiento.
- Por esta razón se debe aplicar la regresión ridge luego de estandarizar los predictores usando:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}}$$

Regresión Lasso

- La regresión ridge tiene una desventaja obvia: incluye a los p predictores en el modelo final.
- La regresión *lasso* es una alternativa que supera esta desventaja. Sus coeficientes minimizan:

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

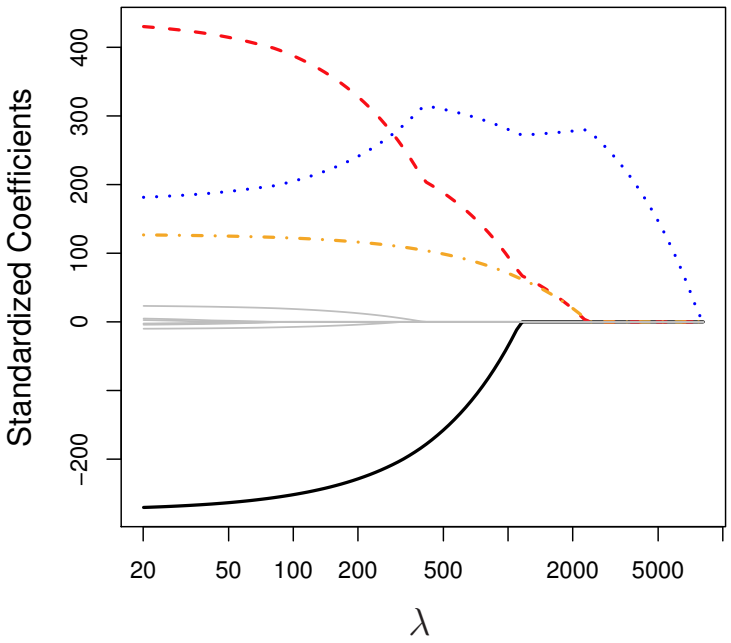
- La regresión lasso usa la norma l_1 definida por:

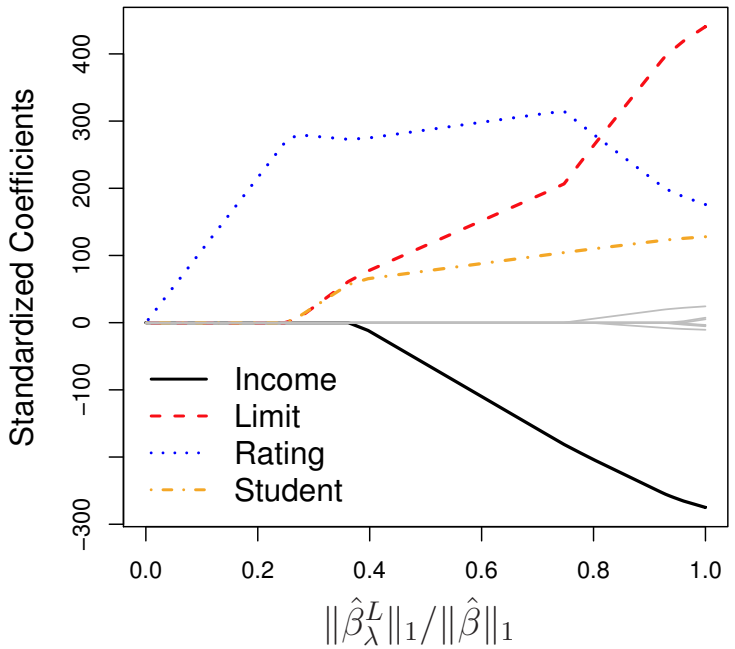
$$||\beta||_1 = \sum_{i=1}^{p} |\beta_j|$$

Regresión Lasso

- Al igual que la regresión ridge, la regresión lasso encoge las estimaciones de los parámetros hacia cero.
- Sin embargo, en el caso de la regresión lasso la penalidad tiene el efecto de forzar que ciertos coeficientes estimados sean exactamente iguales a cero cuando λ es lo suficientemente grande.
- Lo anterior sugiere que la regresión lasso puede usarse para el proceso de selección de variables.
- Para la regresión lasso escoger el valor de λ es crítico. Nuevamente puede usarse validación cruzada.

Regresión Lasso





Introducción

- Los métodos discutidos anteriormente estiman modelos de regresión lineal usando mínimos cuadrados ordinarios o algún método de encogimiento con los predictores originales $X_1, X_2, \cdots, X_p$.
- En esta sección se considera la transformación de los predictores que luego son usados para estimar el modelo vía MCO.
- Estas técnicas son llamadas *métodos de reducción de dimensión*.

Introducción

- Sean $Z_1, Z_2, \cdots, Z_M$ que representan $M < p$ combinaciones lineales de los p predictores originales, es decir:

$$Z_m = \sum_{j=1}^{p} \phi_{mj} X_j \tag{1}$$

para algunas constantes $\phi_{m1}, \cdots, \phi_{mp}$.

- Luego se puede estimar el modelo de regresión lineal:

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i \tag{2}$$

usando MCO.

Introducción

- A partir de la ecuación (1):

$$\begin{aligned}\sum_{m=1}^{M}\theta_m z_{im} &= \sum_{m=1}^{M}\theta_m\sum_{j=1}^{p}\phi_{mj}x_j \\ &= \sum_{j=1}^{p}\sum_{m=1}^{M}\theta_m\phi_{mj}x_j \\ &= \sum_{j=1}^{p}\beta_j x_j\end{aligned}$$

donde:

$$\beta_j = \sum_{m=1}^{M}\theta_m\phi_{mj} \qquad (3)$$

Regresión por Componentes Principales

- Se puede usar el *análisis por componentes principales* (PCA) para definir las combinaciones lineales de los predictores a usar en la regresión.
- El primer componente principal es aquella combinación lineal normalizada de las variables que tiene la mayor variancia.
- El segundo componente principal tiene la siguiente mayor variancia tal que no se encuentra correlacionado con el primer componente, y así sucesivamente.
- Cuando las variables originales se encuentran correlacionadas pueden ser reemplazadas por un conjunto pequeño de componentes principales que capturan su variación conjunta.

Mínimos Cuadrados Parciales

- La regresión por componentes principales identifica las combinaciones lineales o *direcciones* que mejor representan a los predictores $X_1, X_2, \cdots, X_p$.
- Estas direcciones se identifican de forma *no supervisada* desde que Y no es usada para determinar estas direcciones.
- Por otro lado la regresión por componentes principales tiene un potencial inconveniente: no existe garantía que las direcciones que mejor expliquen a los predictores nos permitan explicar de manera apropiada la variable respuesta.

Mínimos Cuadrados Parciales

- Al igual que la PCR, PLS es un método de reducción de dimensión que primero identifica un nuevo conjunto de variables $Z_1, Z_2, \cdots, Z_M$ que son combinaciones lineales de las variables originales y luego estima el modelo lineal vía MCO.
- Pero a diferencia de PCR, PLS identifica estas nuevas variables de forma *supervisada*, es decir se usa la variable Y que permita identificar las nuevas variables que no solamente aproximen las variables originales sino que también se encuentren relacionadas la variable respuesta.
- La técnica de PLS busca las direcciones que permitan explicar la variable respuesta y los predictores.

Mínimos Cuadrados Parciales

- Después de estandarizar los p predictores, PLS calcula la primera dirección Z_1 tomando cada ϕ_{1j} en (1) igual al coeficiente de la regresión lineal simple de Y sobre X_j.
- Se puede demostrar que este coeficiente es proporcional a la correlación entre Y y X_j.
- Por consiguiente al calcular $Z_1 = \sum_{j=1}^{p} \phi_{1j} X_j$, PLS considera mayor peso a las variables que tienen fuerte relación con la variable respuesta.
- Las siguientes direcciones se encuentran tomando los residuales y luego repitiendo el procedimiento anterior.