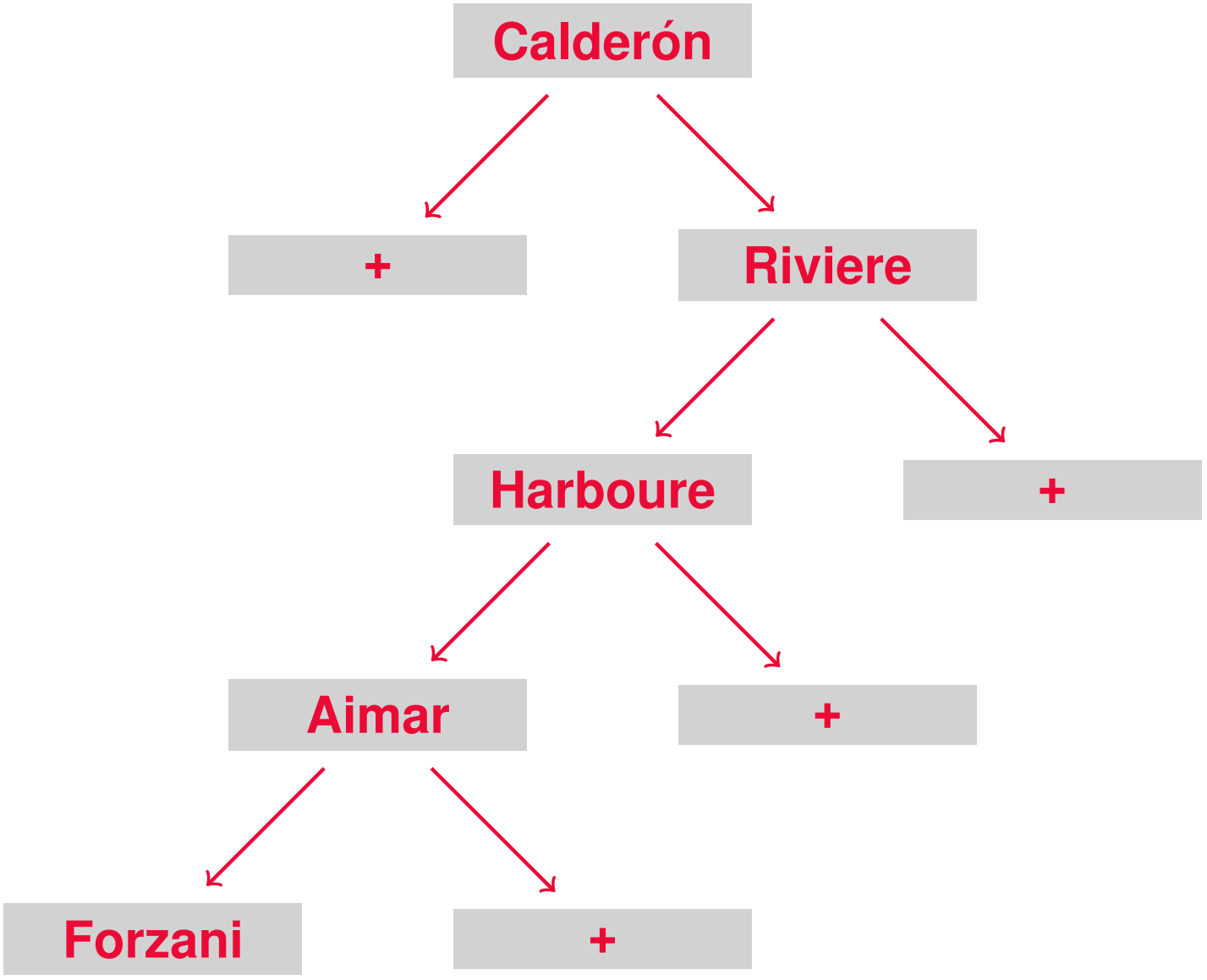
Yo también soy una chica Calderón



Ahorrando dimensiones:

Menos datos, misma información

Liliana Forzani

Facultad de Ingeniería Química, UNL

Bura, **Cook**, Duarte, García A., Gieco, Llop, Pfeiffer, Rodriguez, Rothman, Smucler, Sued, Tolmasky, Tomassi

FIQ UNL

R. Dennis Cook
Estadístico,
U de Minnesota

Efstathia Bura
Estadística, George
Washington University

Rodrigo García
Arancibia
Economista, UNL

Antonella Gieco,
Estadística, UNL

Pamela Llop
Estadística, UNL

R. Ruth Pfeiffer,
Bioestadistica,
Cancer Research Center

Daniela Rodríguez
Estadistica, UBA

Adam Rothman,
Estadístico,
U. de Minnesota

Ezequiel Smucler,
Estadistico, UBA

Mariela Sued
Estadística, UBA

Sabrina Duarte
Estadística, UNL

Carlos Tolmasky,
Matemático,
U. de Minnesota

Diego Tomassi
Computer Science,
UNL

¿Objetivo?

- **Regresión** y **clasificación** supervisada
- Interesa *predecir* o *explicar* una variable respuesta Y en función de un conjunto de predictores $\mathbf{X} = (X_1, X_2, \ldots, X_p)$.

¿Objetivo?

- **Regresión** y **clasificación** supervisada
- Interesa *predecir* o *explicar* una variable respuesta Y en función de un conjunto de predictores $\mathbf{X} = (X_1, X_2, \ldots, X_p)$. $Y = f(X_1, \ldots, X_p)$.
- La solución generalmente es simple cuando $p = 1$ o $p = 2$

La idea básica

Regresión: Estudio de la distribución condicional de una variable respuesta Y dada el vector de predictores $\mathbf{X} \in \mathbb{R}^p$.

La idea básica

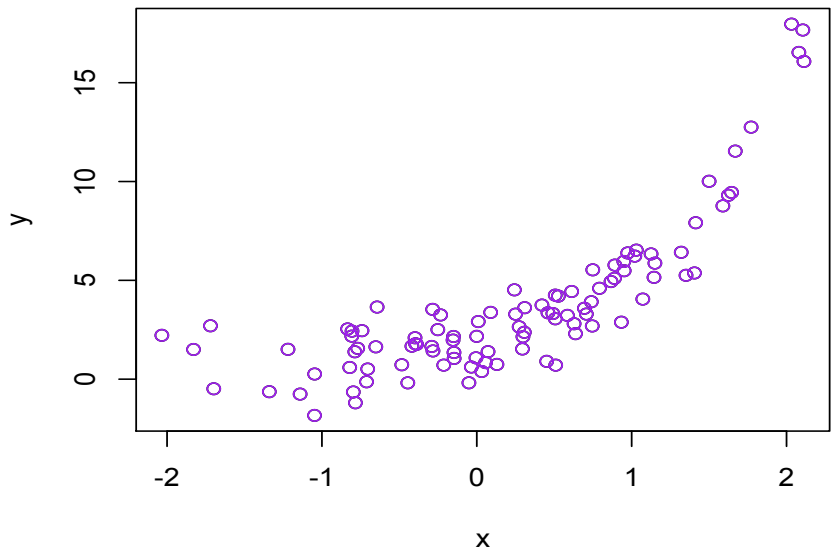
Regresión: Estudio de la distribución condicional de una variable respuesta Y dada el vector de predictores $\mathbf{X} \in \mathbb{R}^p$.

Si $p = 1$, un gráfico nos dice casi todo: grafiquemos los pares (X_1, Y) *datos*

La idea básica

Regresión: Estudio de la distribución condicional de una variable respuesta Y dada el vector de predictores $\mathbf{X} \in \mathbb{R}^p$.

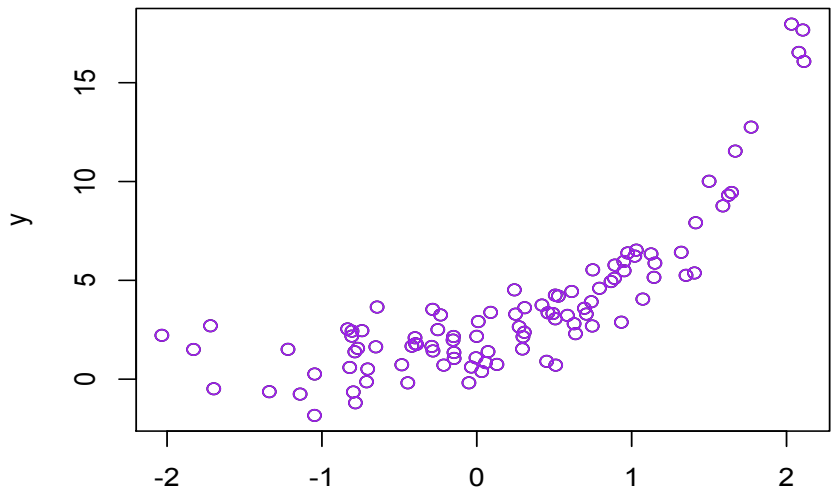
Si $p = 1$, un gráfico nos dice casi todo: grafiquemos los pares (X_1, Y) *datos*



La idea básica

Regresión: Estudio de la distribución condicional de una variable respuesta Y dada el vector de predictores $\mathbf{X} \in \mathbb{R}^p$.

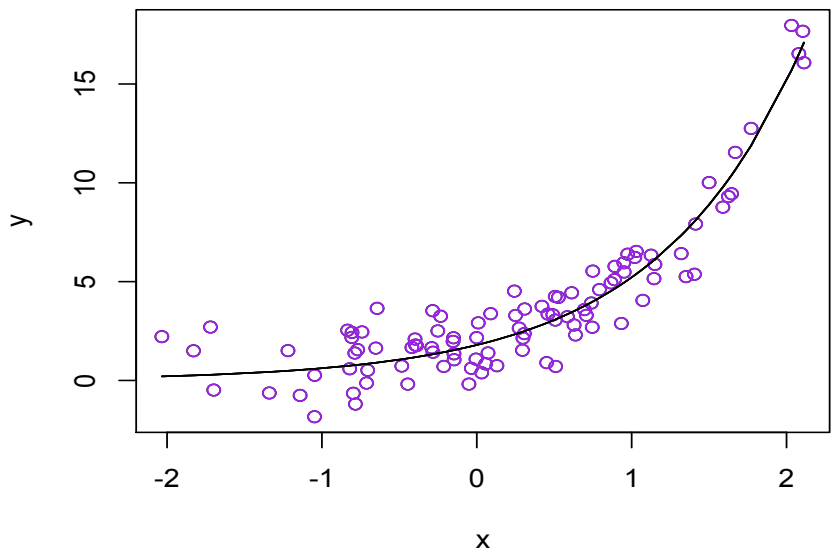
Si $p = 1$, un gráfico nos dice casi todo: grafiquemos los pares (X_1, Y) *datos*



$\mathrm{E}(Y|X = x) = ae^{bx}$. Estimamos a y b.

La idea básica

Regresión: Estudio de la distribución condicional de una variable respuesta Y dada el vector de predictores $\mathbf{X} \in \mathbb{R}^p$.



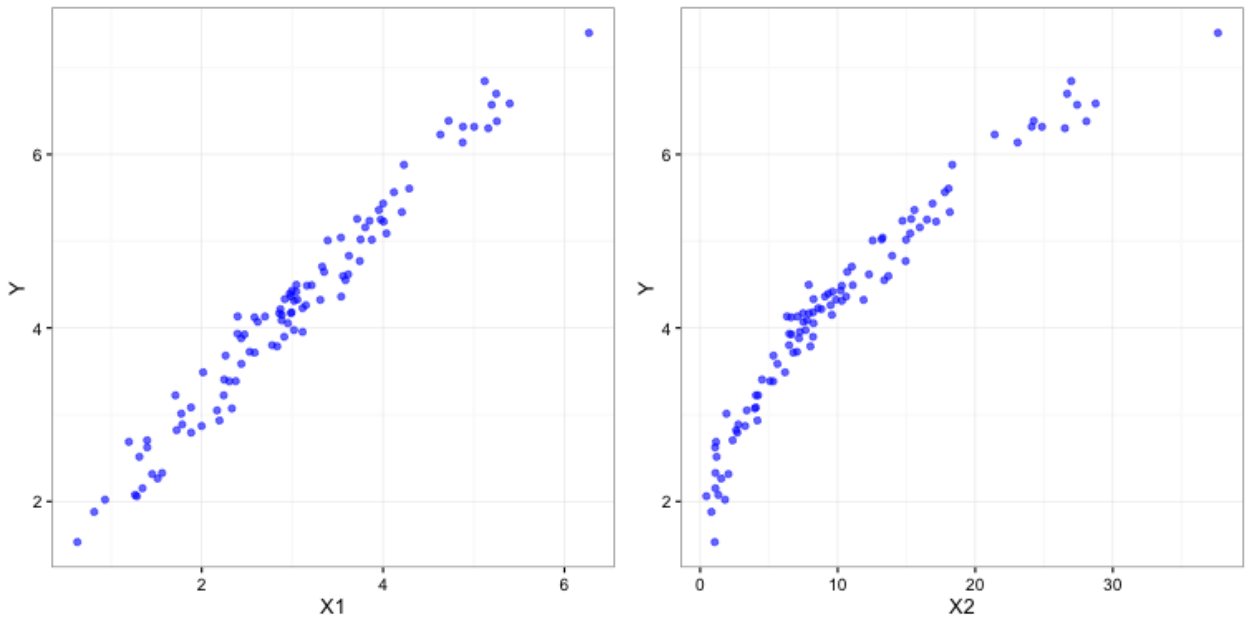
$\mathrm{E}(Y|X = x) = \hat{a}e^{\hat{b}x}$. Observemos que x y $b.x$ nos dan la misma información

¿Qué sucede si tenemos más predictores ?

Queremos estudiar $Y|(X_1, X_2)$, i.e, $Y : \mathbb{R}^2 \to \mathbb{R}$.

¿Qué sucede si tenemos más predictores ?

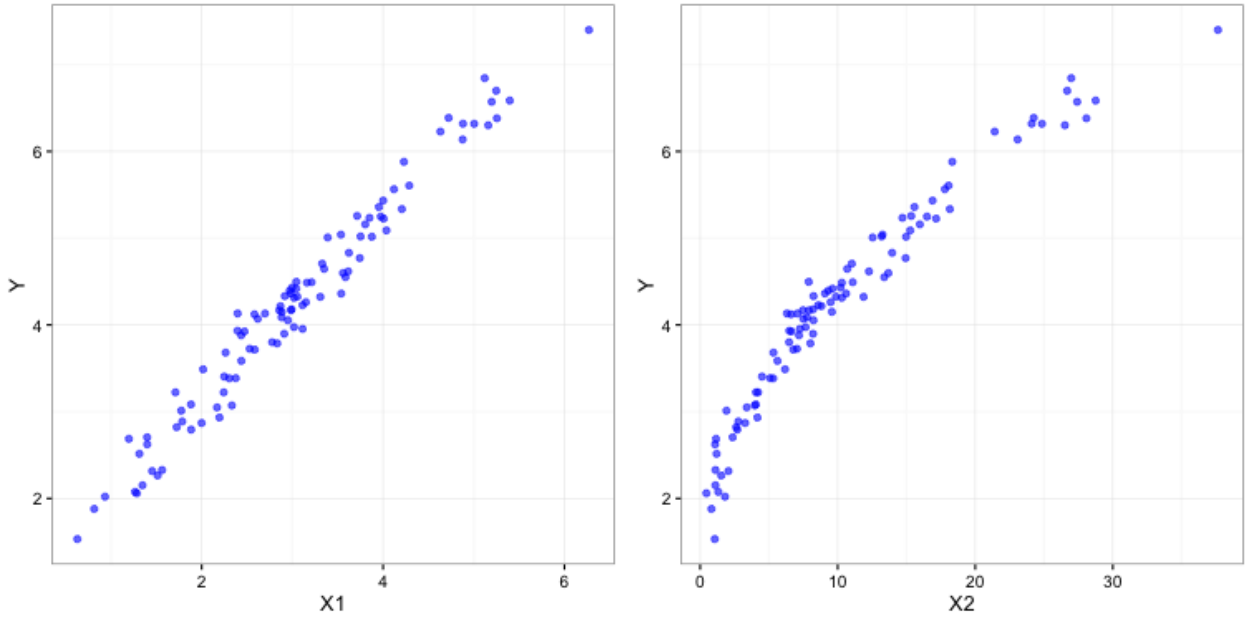
Queremos estudiar $Y|(X_1, X_2)$, i.e, $Y : \mathbb{R}^2 \to \mathbb{R}$. ¿Gráficos marginales que nos dicen?



¿Cómo depende Y de (X_1, X_2)?

¿Qué sucede si tenemos más predictores ?

Queremos estudiar $Y|(X_1, X_2)$, i.e, $Y : \mathbb{R}^2 \to \mathbb{R}$. ¿Gráficos marginales que nos dicen?



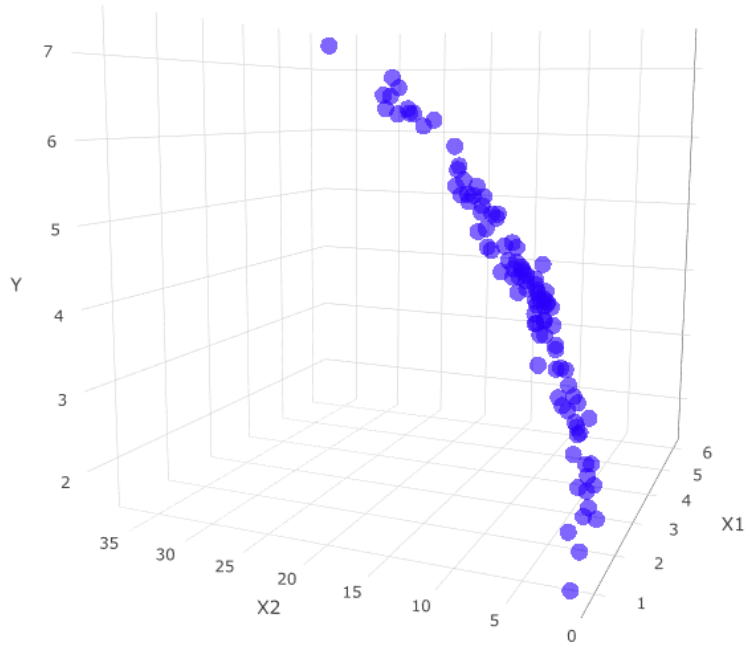
¿Cómo depende Y de (X_1, X_2)? lineal en X_1 y $\sqrt{X_2}$?

Regresión directa

Tenemos que mirar una función $Y : \mathbb{R}^2 \to \mathbb{R}$, i.e. Y vs $\mathbf{X} = (X_1, X_2)$ conjuntamente rotando el gráfico.

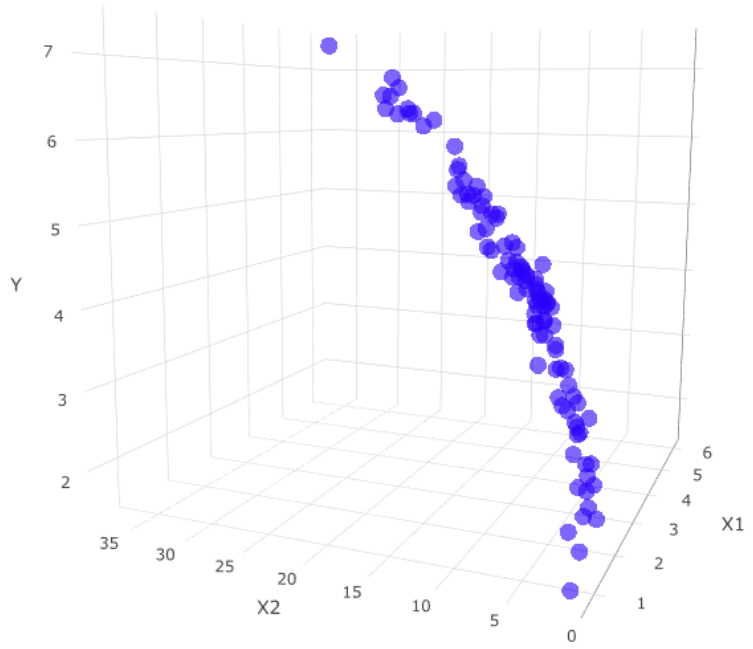
Regresión directa

Tenemos que mirar una función $Y : \mathbb{R}^2 \to \mathbb{R}$, i.e. Y vs $\mathbf{X} = (X_1, X_2)$ conjuntamente rotando el gráfico.



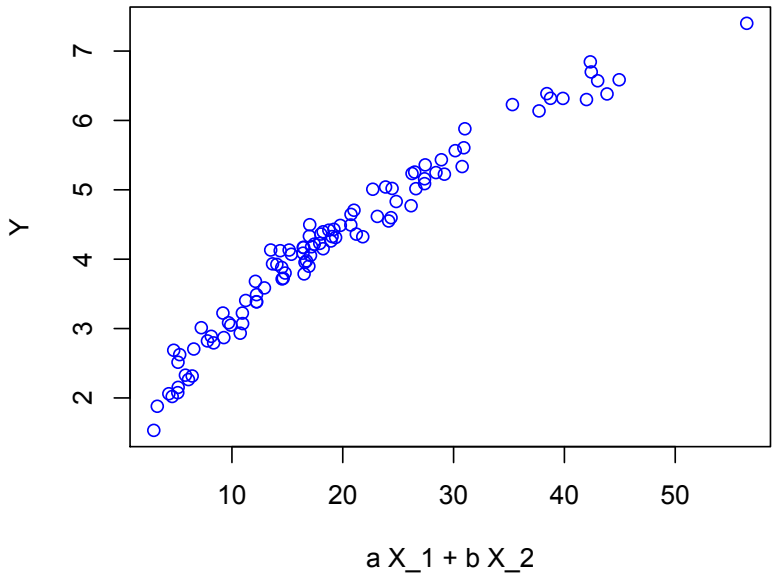
Regresión directa

Tenemos que mirar una función $Y : \mathbb{R}^2 \to \mathbb{R}$, i.e. Y vs $\mathbf{X} = (X_1, X_2)$ conjuntamente rotando el gráfico.



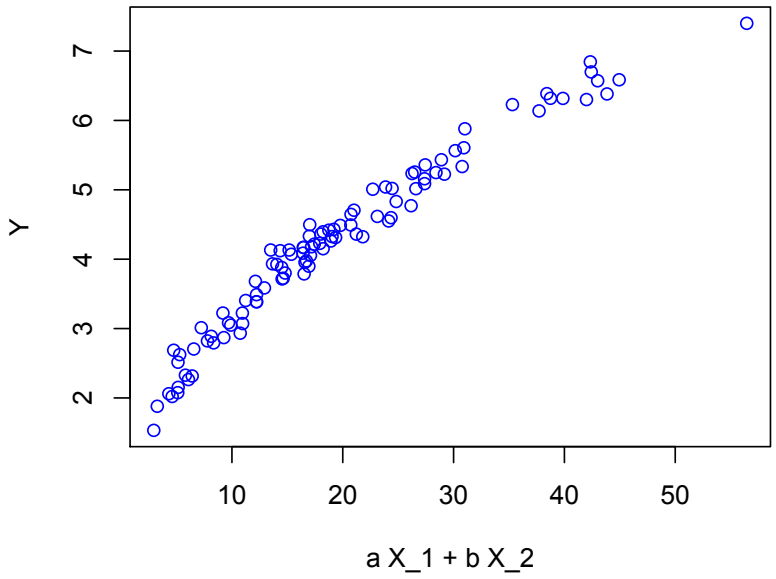
$$Y = f(aX_1 + bX_2).$$

A partir de conocer a y b



Vuelvo al problema de $p = 1$ y modelo $Y = f(aX_1 + bX_2), \ldots$.

A partir de conocer a y b



Vuelvo al problema de $p = 1$ y modelo $Y = f(aX_1 + bX_2), \ldots$.

En este caso:
$Y|(X_1, X_2) =_d Y|\sqrt{3X_1 + X_2} =_d Y|f(3X_1 + X_2) =_d Y|f(\alpha^T \mathbf{X})$
con $\alpha^T = c(3, 1)$.

Ejemplo: caso $p = 2$

Como conclusión $3X_1 + X_2$ no tiene la misma información que el par (X_1, X_2).

Ejemplo: caso $p = 2$

Como conclusión $3X_1 + X_2$ no tiene la misma información que el par (X_1, X_2).

Sin embargo, $3X_1 + X_2$ tiene la misma información que el par (X_1, X_2) sobre Y

Qué pasa cuando la cantidad de predictores aumenta?

Difcil comprensión de la relación entre predictores y respuesta

Qué hacemos?
Como antes: transformamos el problema en uno más simple, encontramos α tal que $\alpha^T \mathbf{X}$ y $\mathbf{X}$ tengan la misma información sobre Y.

Qué pasa cuando la cantidad de predictores aumenta?

Difcil comprensión de la relación entre predictores y respuesta

Qué hacemos?
Como antes: transformamos el problema en uno más simple, encontramos α tal que $\alpha^T \mathbf{X}$ y $\mathbf{X}$ tengan la misma información sobre Y.

Cómo encontramos α?
Estudiamos la función $\mathbb{R}^p \to \mathbb{R}$. ¿Gráficos? ¿Dios nos da α? Y luego sigo con el problema.

Solución: Principal Components in Regression ($2.010.000$ resultados en google)

Si p es grande un paso preliminar a todo estudio es reducir $\mathbf{X}$

Solución: Principal Components in Regression ($2.010.000$ resultados en google)

Si p es grande un paso preliminar a todo estudio es reducir $\mathbf{X}$

La idea

- Reemplazar $\mathbf{X} = (X_1, \ldots, X_p)$ por $k < p$ combinaciones lineales de X_i. Si $k = 1$ sería reemplazar $\mathbf{X}$ por la variable $a_1 X_1 + \cdots + a_p X_p$

Solución: Principal Components in Regression ($2.010.000$ resultados en google)

Si p es grande un paso preliminar a todo estudio es reducir $\mathbf{X}$

La idea

- Reemplazar $\mathbf{X} = (X_1, \ldots, X_p)$ por $k < p$ combinaciones lineales de X_i. Si $k = 1$ sería reemplazar $\mathbf{X}$ por la variable $a_1 X_1 + \cdots + a_p X_p$

- Estudiar la regresión de Y en función de estas combinaciones lineales en lugar de Y en función de $\mathbf{X}$

Solución: Principal Components in Regression ($2.010.000$ resultados en google)

Si p es grande un paso preliminar a todo estudio es reducir $\mathbf{X}$

La idea

- Reemplazar $\mathbf{X} = (X_1, \ldots, X_p)$ por $k < p$ combinaciones lineales de X_i. Si $k = 1$ sería reemplazar $\mathbf{X}$ por la variable $a_1 X_1 + \cdots + a_p X_p$

- Estudiar la regresión de Y en función de estas combinaciones lineales en lugar de Y en función de $\mathbf{X}$

¿Cómo elige PCR estas combinaciones lineales?

Como tomar una foto de una tetera

Como tomar una foto de una tetera

Es la foto que contiene la mayor cantidad de información posible

PCR

PCR of Y on $\mathbf{X} = (X_1, \ldots, X_p)^T$

- Tomar Cov $(\mathbf{X})$
- Encontrar los primeros k autovectores de Cov $(\mathbf{X})$: $\alpha = (\mathbf{a}_1, \ldots, \mathbf{a}_k)$
- Considerar $\alpha^T \mathbf{X} = (\mathbf{a}_1^T \mathbf{X}, \ldots, \mathbf{a}_k^T \mathbf{X})$
- Estudiar Y en función de $\alpha^T \mathbf{X}$ o clasificar usando $\alpha^T \mathbf{X}$.

Algunas ventajas y racionalidad atras de PCR

- Faćil de entender: $2.010.000$ resultados en google vs $3.220.000$ resultados para regresión lineal simple
- Las combinaciones lineales de $\mathbf{X}$ son las que tienen mayor variabilidad
- No necesito modelar $Y|\mathbf{X}$ como paso previo
- Si usamos los PC, en lugar de las variables originales los PC minimizan la varianza de los estimadores

Algunas ventajas y racionalidad atras de PCR

- Faćil de entender: $2.010.000$ resultados en google vs $3.220.000$ resultados para regresión lineal simple
- Las combinaciones lineales de $\mathbf{X}$ son las que tienen mayor variabilidad
- No necesito modelar $Y|\mathbf{X}$ como paso previo
- Si usamos los PC, en lugar de las variables originales los PC minimizan la varianza de los estimadores

PERO

- Error = VARIANZA + SESGO^2

Algunas ventajas y racionalidad atras de PCR

- Faćil de entender: $2.010.000$ resultados en google vs $3.220.000$ resultados para regresión lineal simple
- Las combinaciones lineales de $\mathbf{X}$ son las que tienen mayor variabilidad
- No necesito modelar $Y|\mathbf{X}$ como paso previo
- Si usamos los PC, en lugar de las variables originales los PC minimizan la varianza de los estimadores

PERO

- Error = VARIANZA + SESGO^2
  - La varianza esta conectada con la varianza de $\mathbf{X}$
  - SESGO es la conexión entre $\mathbf{X}$ e Y.

Algunas ventajas y racionalidad atras de PCR

- Faćil de entender: $2.010.000$ resultados en google vs $3.220.000$ resultados para regresión lineal simple
- Las combinaciones lineales de $\mathbf{X}$ son las que tienen mayor variabilidad
- No necesito modelar $Y|\mathbf{X}$ como paso previo
- Si usamos los PC, en lugar de las variables originales los PC minimizan la varianza de los estimadores

PERO

- Error = VARIANZA + SESGO^2
  - La varianza esta conectada con la varianza de $\mathbf{X}$
  - SESGO es la conexión entre $\mathbf{X}$ e Y. Y los PCs NO tienen conexión con Y

Ejemplo de juguete

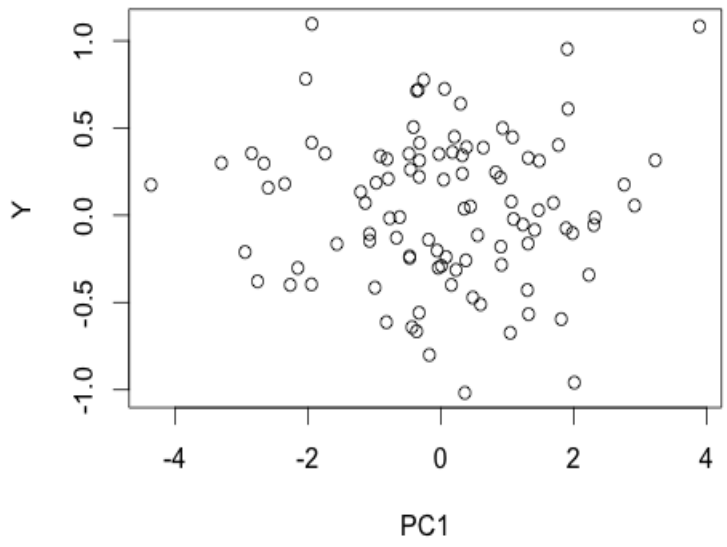
$\mathbf{X} = (X_1, X_2)$, $Y = f(X_1, X_2) + \epsilon$

- Calculo la primera PC y gráfico Y vs PC_1

Ejemplo de juguete

$\mathbf{X} = (X_1, X_2)$, $Y = f(X_1, X_2) + \epsilon$

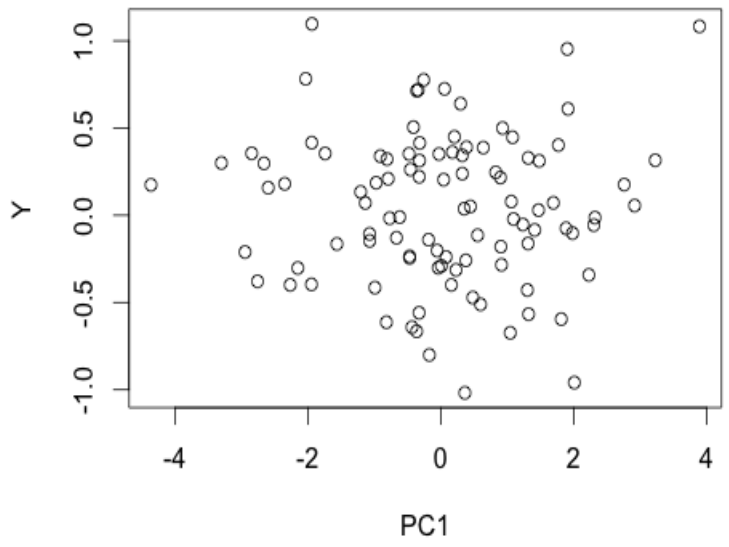
- Calculo la primera PC y gráfico Y vs PC_1



Ejemplo de juguete

$\mathbf{X} = (X_1, X_2)$, $Y = f(X_1, X_2) + \epsilon$

- Calculo la primera PC y gráfico Y vs PC_1

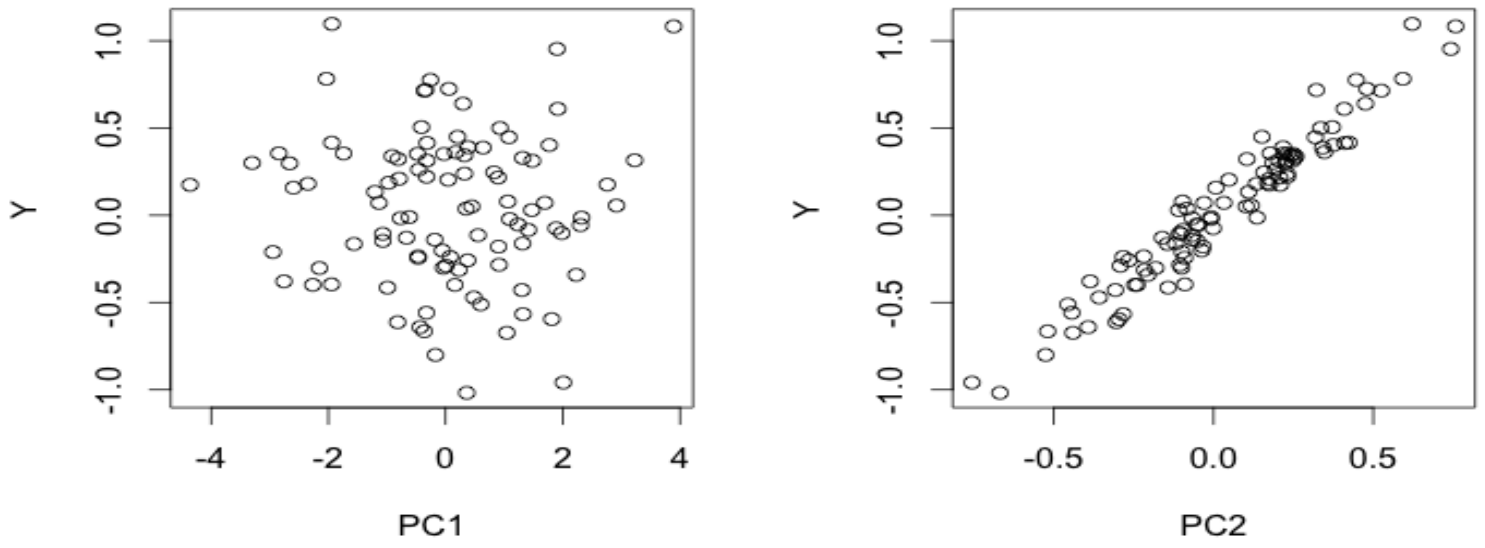


- Pareciera que no hay ninguna relación. Grafico vs PC_2.

Ejemplo de juguete

$\mathbf{X} = (X_1, X_2)$, $Y = f(X_1, X_2) + \epsilon$

- Calculo la primera PC y gráfico Y vs PC_1

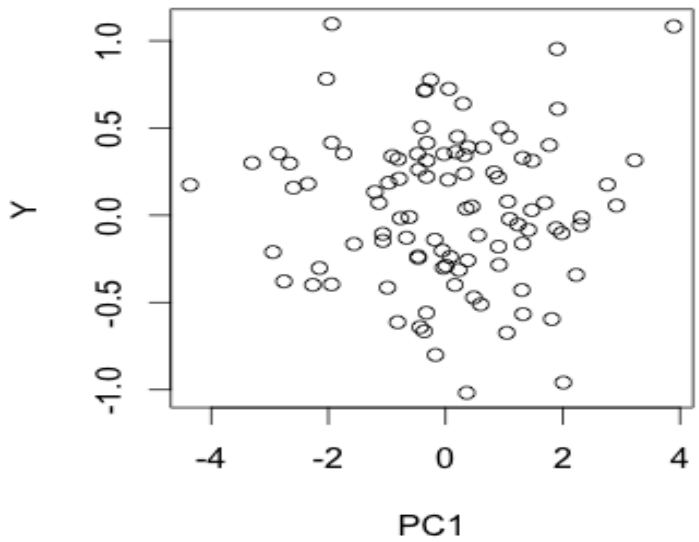


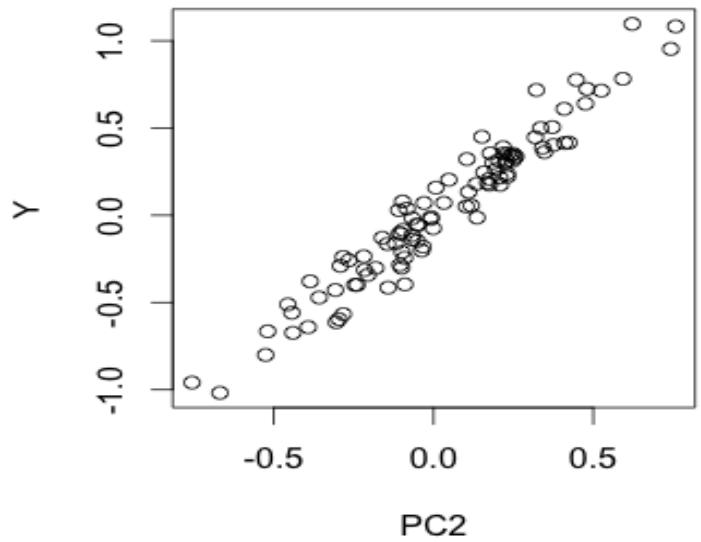
- Pareciera que no hay ninguna relación. Grafico vs PC_2.

Ejemplo de juguete

$\mathbf{X} = (X_1, X_2)$, $Y = f(X_1, X_2) + \epsilon$

- Calculo la primera PC y gráfico Y vs PC_1





- ¿Porqué?: en este ejemplo $Y = X_1 - X_2 + \epsilon$,

  PC_1 teórica: $X_1 + X_2$ y PC_2 teórica: $X_1 - X_2$.

- $PC_1 :\approx 95\%$ variabilidad $PC_2 :\approx 5\%$ variabilidad

¿Qué pasó con la tetera

¿Qué pasó con la tetera

Y, queriamos estudiar la tapa

Resumen

- Mayor problema con el uso de PC es que no hay conexión entre PCs y Y.

Resumen

- Mayor problema con el uso de PC es que no hay conexión entre PCs y Y.
- Están ordenados de acuerdo a su importancia con respecto a $\mathbf{X}$ y no con respecto a Y.

Resumen

- Mayor problema con el uso de PC es que no hay conexión entre PCs y Y.
- Están ordenados de acuerdo a su importancia con respecto a $\mathbf{X}$ y no con respecto a Y.
- Positivo: se calculan antes de modelar $Y|\mathbf{X}$ y si funcionan pueden servir de mucho (esperar que Y se mueva en la misma dirección que los primeros k PCs).

Reducción suficiente de dimensiones

Definición

Una reducción $R : \mathbb{R}^p \to \mathbb{R}^d$, con $d \leq p$ es suficiente para la regresión de $Y|\mathbf{X}$ si satisface la siguiente condición

$$Y|\mathbf{X} \sim Y|R(\mathbf{X})$$

Reducción suficiente de dimensiones

Definición

Una reducción $R : \mathbb{R}^p \to \mathbb{R}^d$, con $d \leq p$ es suficiente para la regresión de $Y|\mathbf{X}$ si satisface la siguiente condición

$$Y|\mathbf{X} \sim Y|R(\mathbf{X})$$

¿Cual es la diferencia con $R(\mathbf{X})$ via componentes principales?

Reducción suficiente de dimensiones

¿Cómo encontramos $R(\mathbf{X})$?

Respuesta: Enfoque de reducción inversa

Reducción suficiente de dimensiones

¿Cómo encontramos $R(\mathbf{X})$?

Respuesta: Enfoque de reducción inversa

Objetivo: $Y|\mathbf{X}$ pero estudiaremos $\mathbf{X}|Y$

- $Y|\mathbf{X} \Rightarrow$ Regresión de $\mathbb{R}$ en $\mathbb{R}^p$, estudiar una función $Y : \mathbb{R}^p \to \mathbb{R}$

Reducción suficiente de dimensiones

¿Cómo encontramos $R(\mathbf{X})$?

Respuesta: Enfoque de reducción inversa

Objetivo: $Y|\mathbf{X}$ pero estudiaremos $\mathbf{X}|Y$

- $Y|\mathbf{X} \Rightarrow$ Regresión de $\mathbb{R}$ en $\mathbb{R}^p$, estudiar una función $Y : \mathbb{R}^p \to \mathbb{R}$
- $\mathbf{X}|Y \Rightarrow p$ regresiones univariadas, $\mathbb{R} \to \mathbb{R}^p$,

$$X_1|Y$$
$$X_2|Y$$
$$\vdots$$
$$X_{p-1}|Y$$
$$X_p|Y$$

Más fáciles de estudiar!

¿Cómo encontramos $R(\mathbf{X})$ cuando tenemos datos?

Métodos basados en modelos para la regresión inversa para $\mathbf{X}|Y$

Cook (2007)

¿Cómo encontramos $R(\mathbf{X})$ cuando tenemos datos?

Métodos basados en modelos para la regresión inversa para $\mathbf{X}|Y$

Cook (2007)

Ventajas

- Estimadores ML
- Identificación exhaustiva de $R(\mathbf{X})$

PFC - Principal Fitted Component

Método PFC:

$$\mathbf{X}_y \doteq \mathbf{X}|(Y = y) \sim \mathcal{N}(\mu_y, \Delta) \tag{2}$$

PFC - Principal Fitted Component

Método PFC:

$$\mathbf{X}_y \doteq \mathbf{X}|(Y = y) \sim \mathcal{N}(\mu_y, \Delta) \tag{2}$$

- $R(\mathbf{X}) = \alpha^T \mathbf{X}$ es una reducción suficiente minimal si

$$\mathcal{S}_\alpha = \Delta^{-1}\text{span}\{\mu_y - \mu, y \in \Omega_Y\}$$

PFC - Principal Fitted Component

Método PFC:

$$\mathbf{X}_y \doteq \mathbf{X}|(Y = y) \sim \mathcal{N}(\mu_y, \Delta) \tag{2}$$

- $R(\mathbf{X}) = \alpha^T \mathbf{X}$ es una reducción suficiente minimal si

$$\mathcal{S}_\alpha = \Delta^{-1}\text{span}\{\mu_y - \mu, y \in \Omega_Y\}$$

- ¿En los datos? para el modelo (2) se encuentran los estimadores ML y $\hat{R}(\mathbf{X}) = \hat{\alpha}^T \mathbf{X}$, con

$$\hat{\alpha} = \hat{\Delta}^{-1}\widehat{\text{span}}\{\mu_y - \mu, y \in \Omega_Y\}$$

PFC - Principal Fitted Component

Método PFC:

$$\mathbf{X}_y \doteq \mathbf{X}|(Y = y) \sim \mathcal{N}(\mu_y, \Delta) \tag{2}$$

- $R(\mathbf{X}) = \alpha^T \mathbf{X}$ es una reducción suficiente minimal si

$$\mathcal{S}_\alpha = \Delta^{-1}\text{span}\{\mu_y - \mu, y \in \Omega_Y\}$$

- ¿En los datos? para el modelo (2) se encuentran los estimadores ML y $\hat{R}(\mathbf{X}) = \hat{\alpha}^T \mathbf{X}$, con

$$\hat{\alpha} = \hat{\Delta}^{-1}\widehat{\text{span}}\{\mu_y - \mu, y \in \Omega_Y\}$$

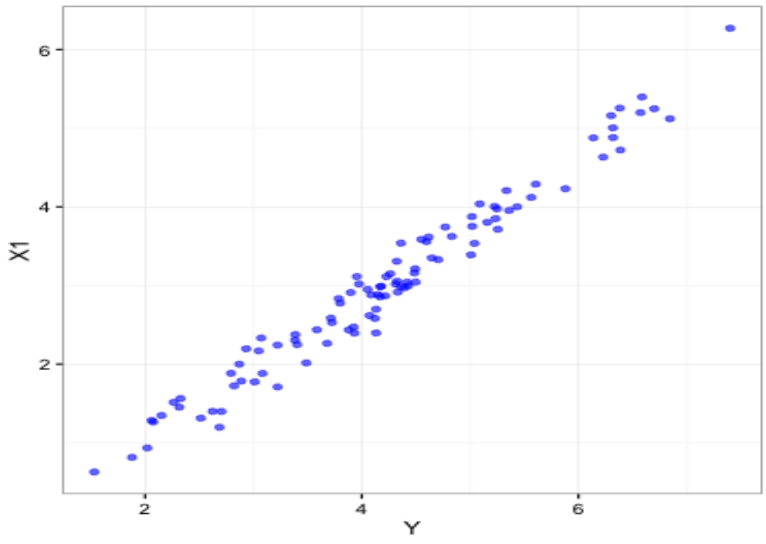
- Extensión al caso Δ_y (Método LAD)

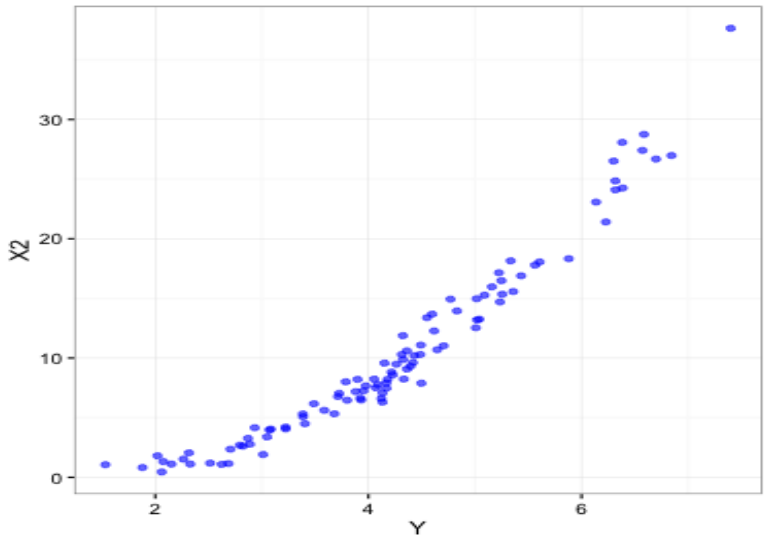
Regresión inversa

Regresión inversa.¿Ventajas? p regresiones uni-dimensionales

Regresión inversa

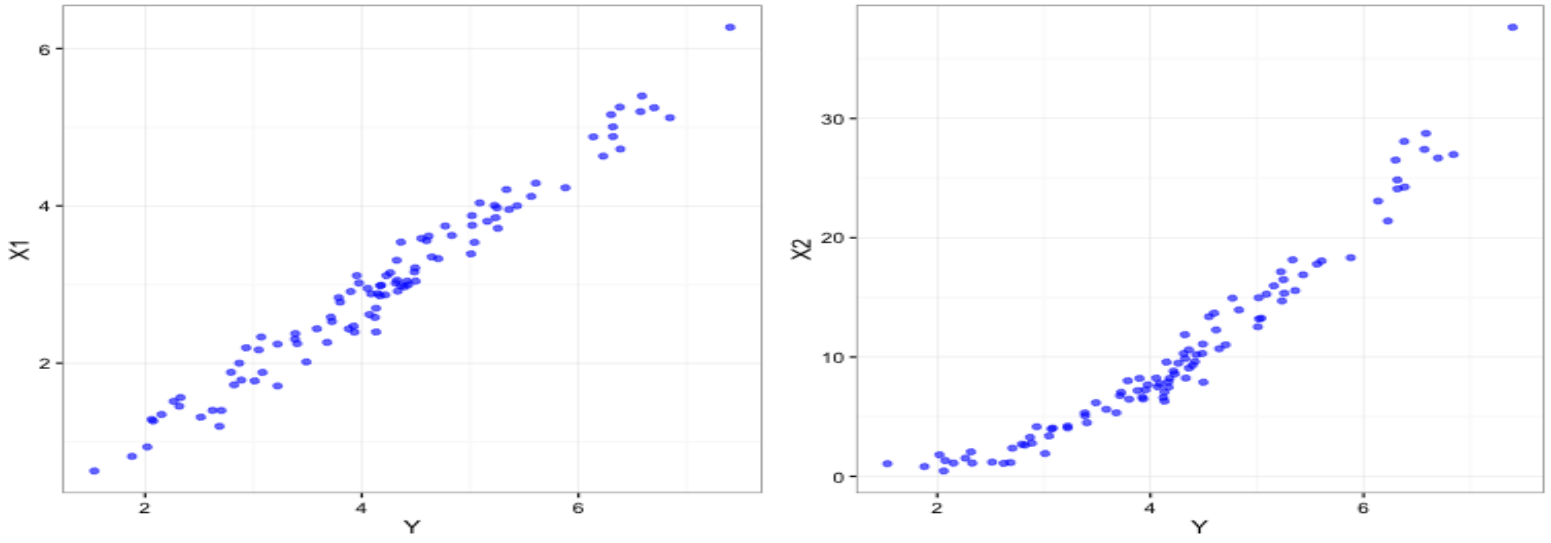
Regresión inversa.¿Ventajas? p regresiones uni-dimensionales





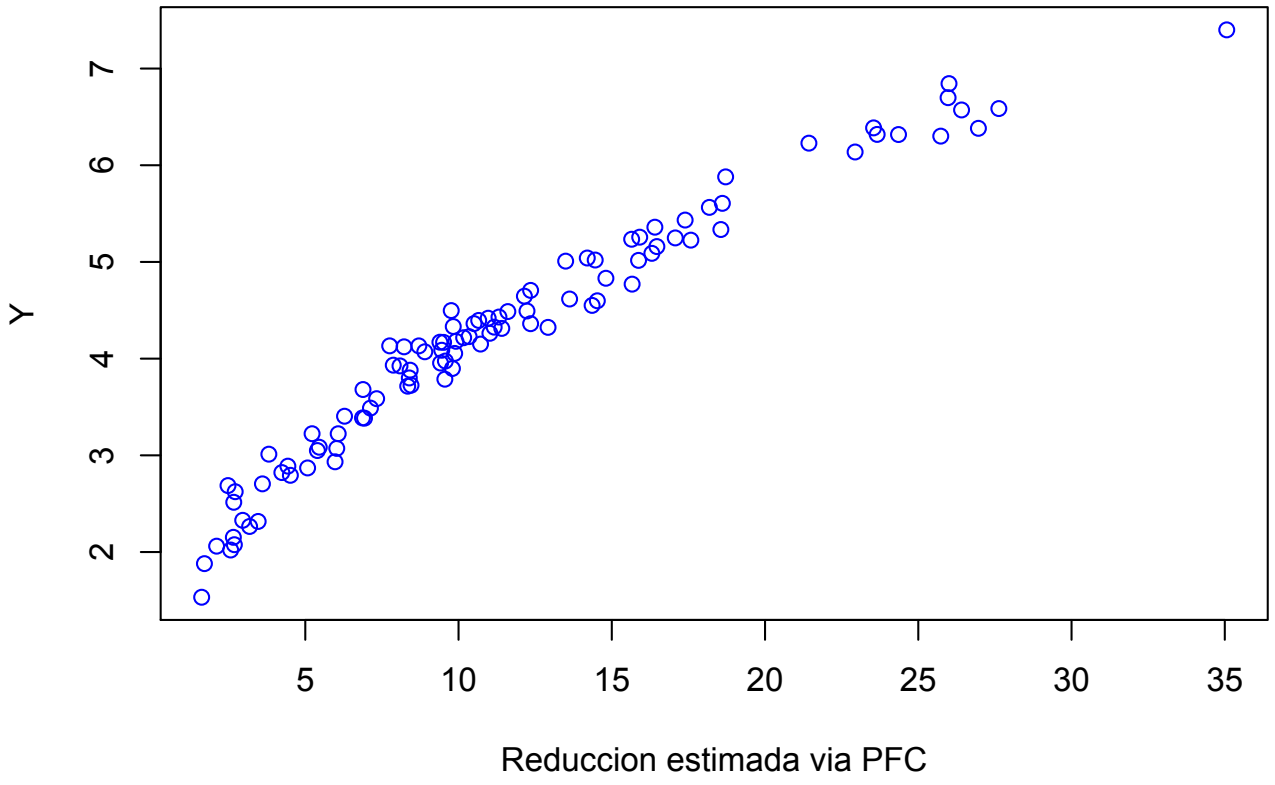
Regresión inversa

Regresión inversa.¿Ventajas? p regresiones uni-dimensionales



- Modelar $X_1|Y$, $X_2|Y$ o $(X_1, X_2)|Y$
- Encontrar R tal que $(X_1, X_2)|(R(X_1, X_2)^T, Y)$ no dependa de Y. Via PFC o LAD

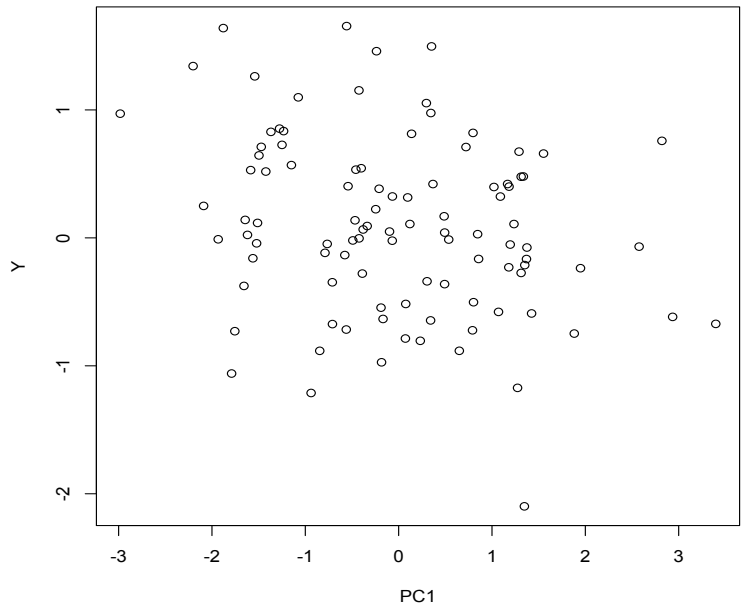
Solución



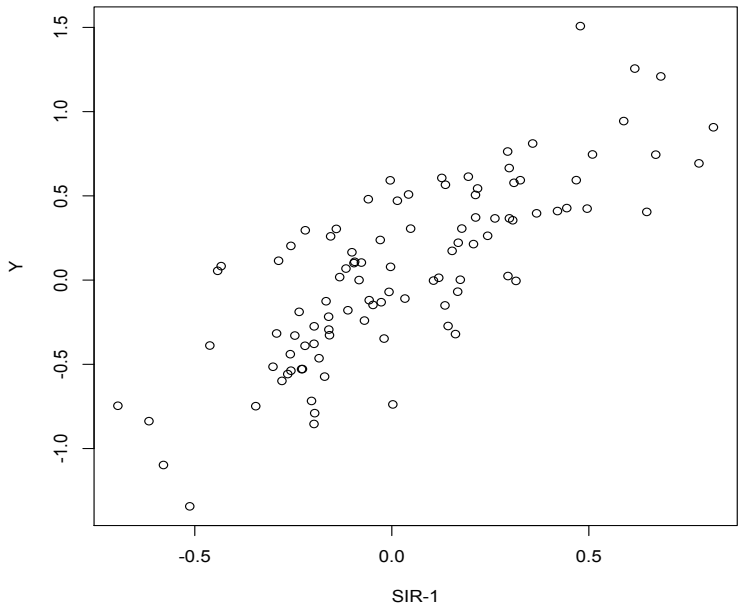
Y de ahí modelamos Y en función de esta combinación lineal.

Primera PC en el ejemplo de juguete

Primera PC

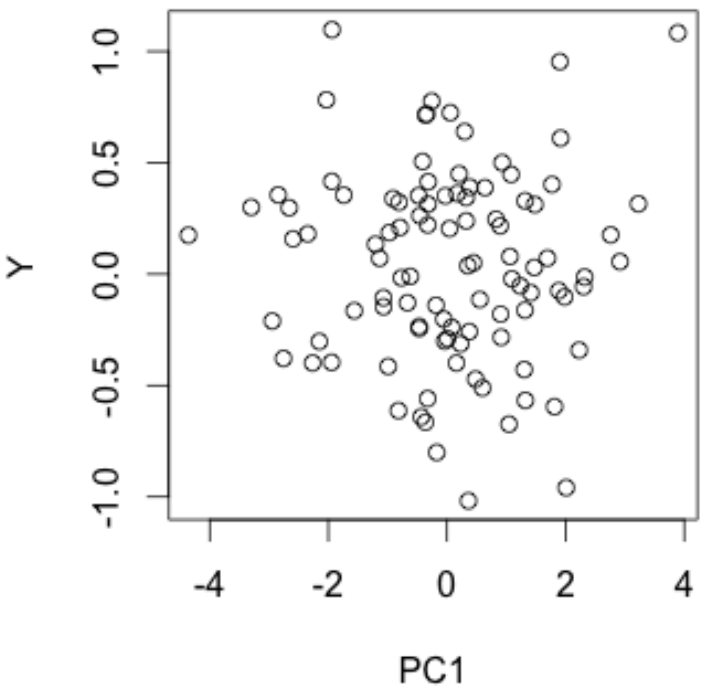


Primera PFC

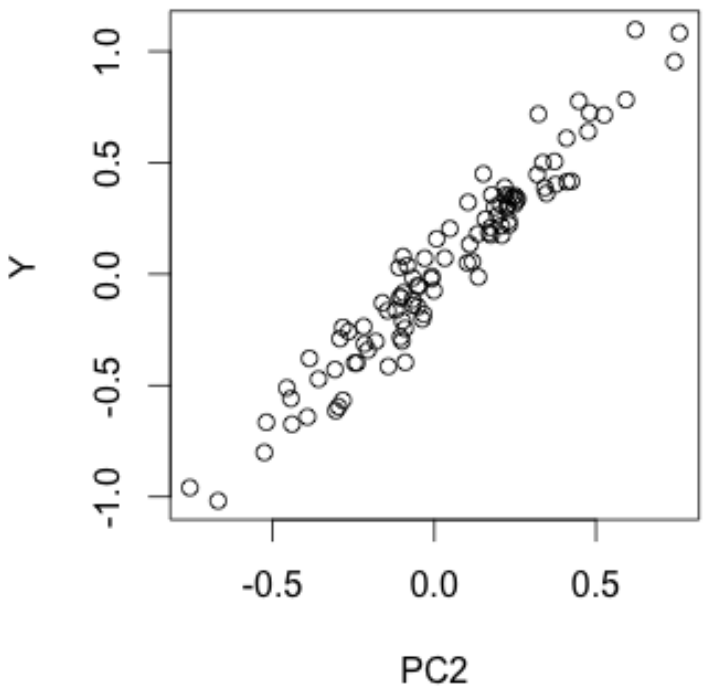


Primera PC en el ejemplo de juguete

Primera PC



Segunda PC



Y discreta: discriminación. Cáncer de pulmón

- Los datos de entrenamiento: $\mathbf{X} \in \mathbb{R}^p$ en poblaciones diferentes $Y = 1, \ldots, H$ para n individuos.

  $\mathbf{X}$: 24 biomarcadores para cada uno de los 174 (individuos con cáncer) + 59 (individuos sin cáncer)

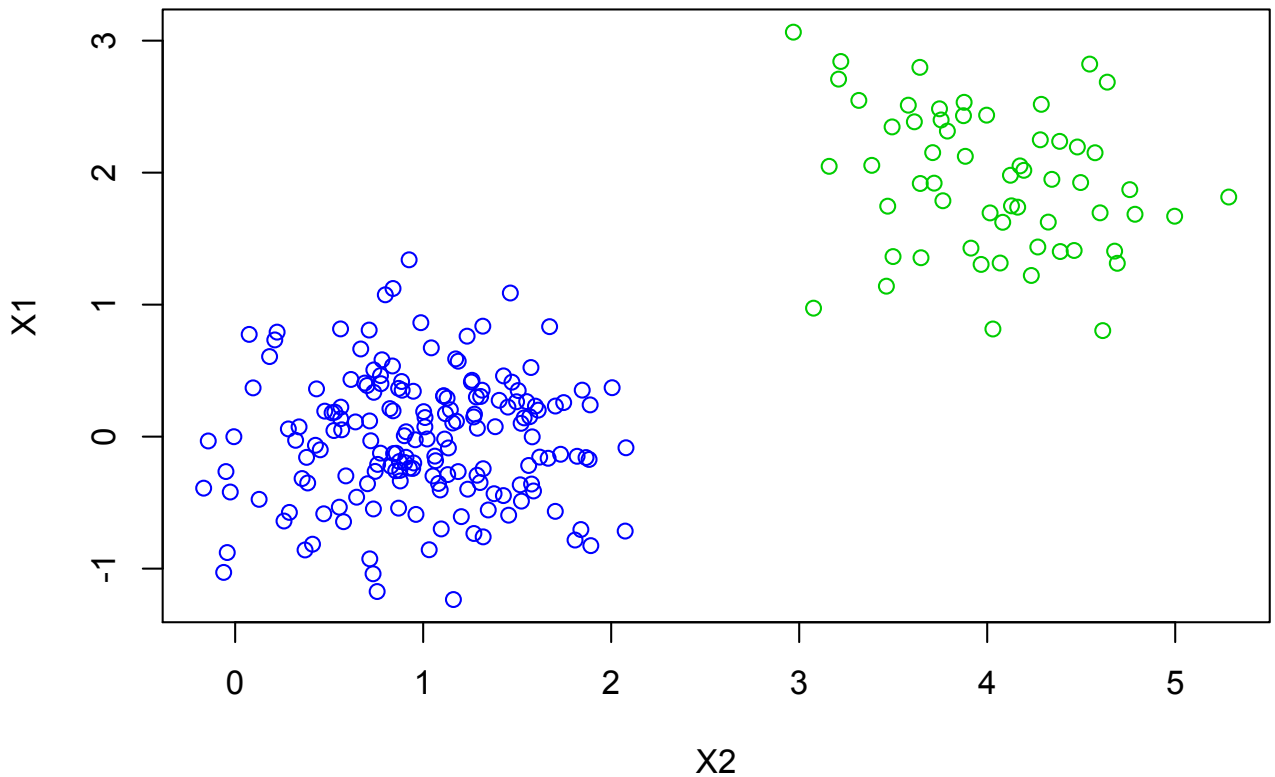
Y discreta: discriminación. Cáncer de pulmón

- Los datos de entrenamiento: $\mathbf{X} \in \mathbb{R}^p$ en poblaciones diferentes $Y = 1, \ldots, H$ para n individuos.

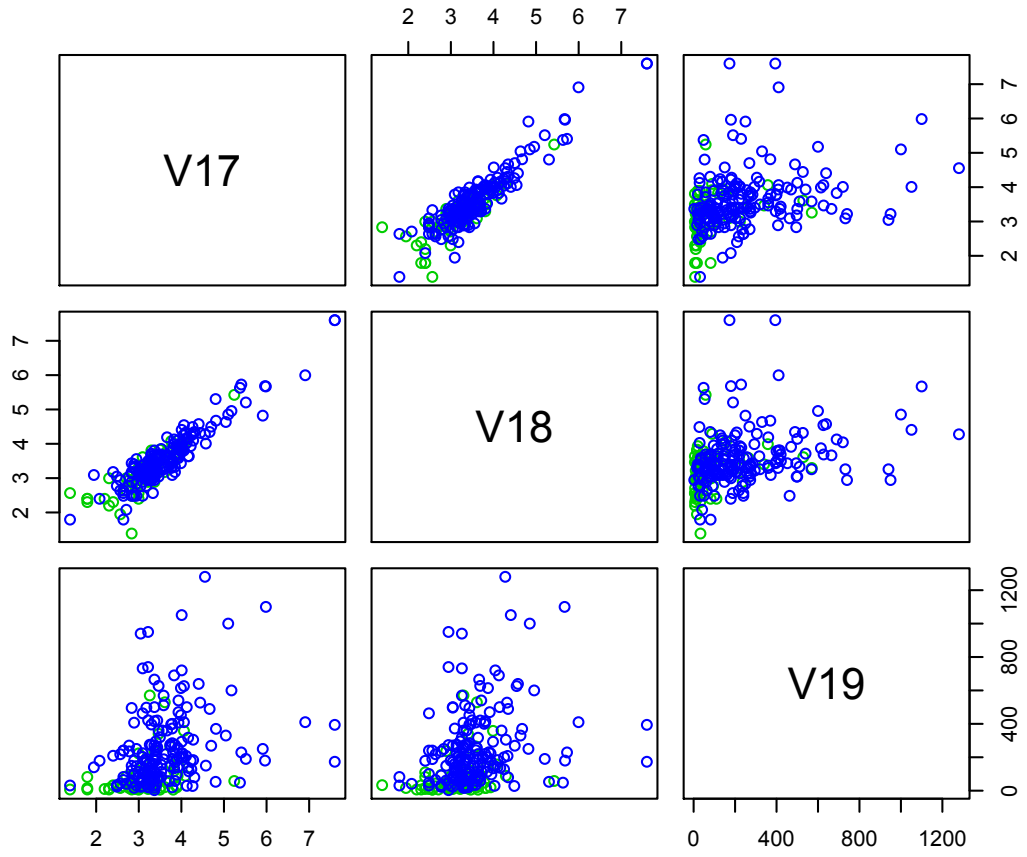
  $\mathbf{X}$: 24 biomarcadores para cada uno de los 174 (individuos con cáncer) + 59 (individuos sin cáncer)

  ¿Los biomarcadores dicen algo de la enfermedad?

Gráficos



Gráficos



Y discreta: discriminación. Cáncer de pulmón

- Los datos de entrenamiento: $\mathbf{X} \in \mathbb{R}^p$ en poblaciones diferentes $Y = 1, \ldots, H$ para n individuos.

  $\mathbf{X}$: 24 biomarcadores para cada uno de los 174 (individuos con cáncer) + 59 (individuos sin cáncer)
  ¿Los biomarcadores dicen algo de la enfermedad?

- Objetivo: encontrar $\widehat{\alpha}$ such that $\widehat{\alpha}^T \mathbf{X}$ tenga la misma información que $\mathbf{X}$ de Y (reducción suficiente de dimensiones) y que además ayude visualmente.

Y discreta: discriminación. Cáncer de pulmón

- Los datos de entrenamiento: $\mathbf{X} \in \mathbb{R}^p$ en poblaciones diferentes $Y = 1, \ldots, H$ para n individuos.

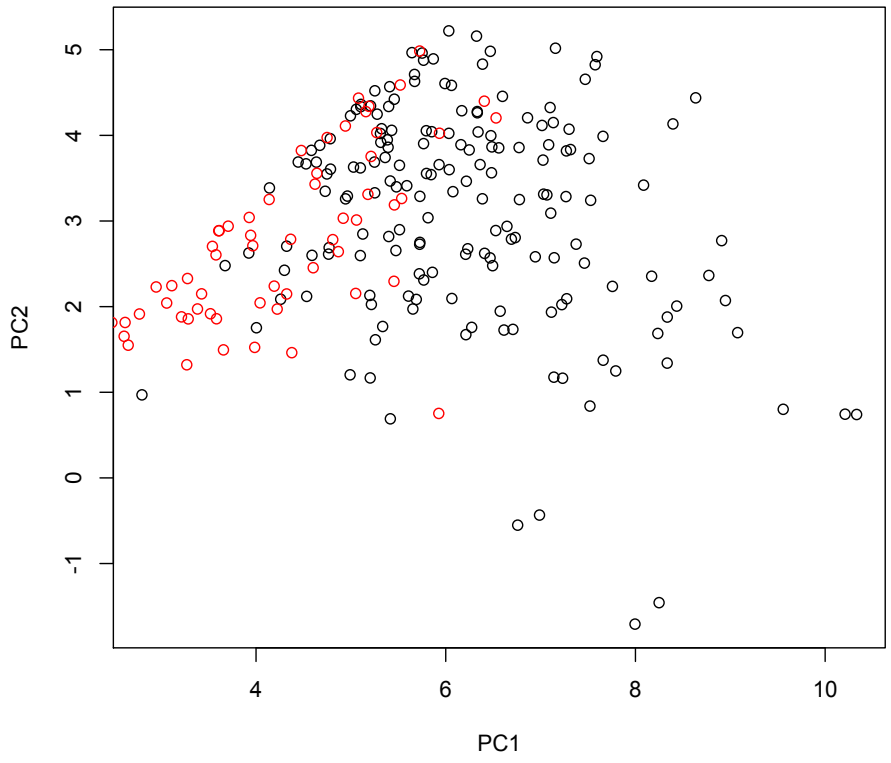
  $\mathbf{X}$: 24 biomarcadores para cada uno de los 174 (individuos con cáncer) + 59 (individuos sin cáncer)
  ¿Los biomarcadores dicen algo de la enfermedad?
- Objetivo: encontrar $\widehat{\alpha}$ such that $\widehat{\alpha}^T \mathbf{X}$ tenga la misma información que $\mathbf{X}$ de Y (reducción suficiente de dimensiones) y que además ayude visualmente.
- Consideramos los predictores $\widehat{\alpha}^T \mathbf{X}$ en lugar de $\mathbf{X}$

Y discreta: discriminación. Cáncer de pulmón

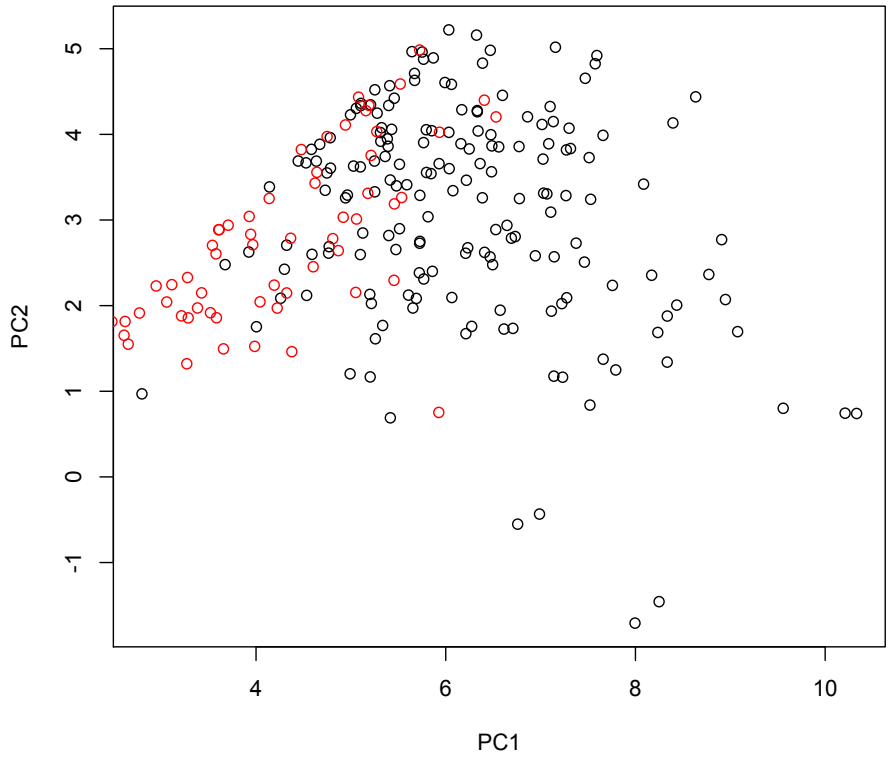
- Los datos de entrenamiento: $\mathbf{X} \in \mathbb{R}^p$ en poblaciones diferentes $Y = 1, \ldots, H$ para n individuos.

  $\mathbf{X}$: 24 biomarcadores para cada uno de los 174 (individuos con cáncer) + 59 (individuos sin cáncer)
  ¿Los biomarcadores dicen algo de la enfermedad?
- Objetivo: encontrar $\widehat{\alpha}$ such that $\widehat{\alpha}^T \mathbf{X}$ tenga la misma información que $\mathbf{X}$ de Y (reducción suficiente de dimensiones) y que además ayude visualmente.
- Consideramos los predictores $\widehat{\alpha}^T \mathbf{X}$ en lugar de $\mathbf{X}$
- Hacemos gráficos (para la visualización) y usamos estos nuevos predictores para predecir Y, cáncer o no

Cánder de pulmón data - PC

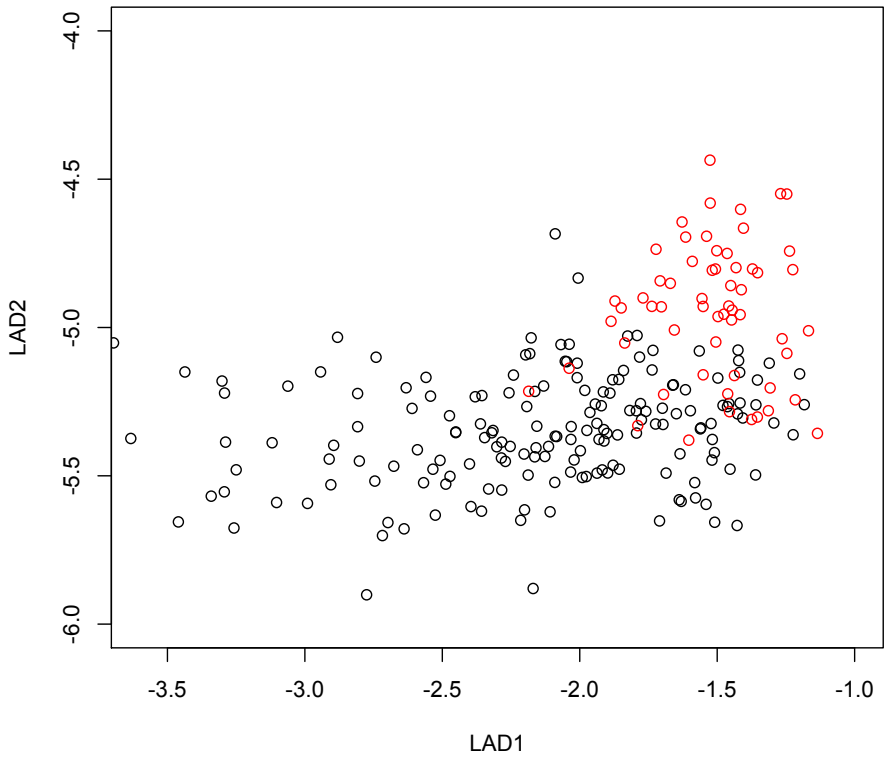


Cánder de pulmón data - PC

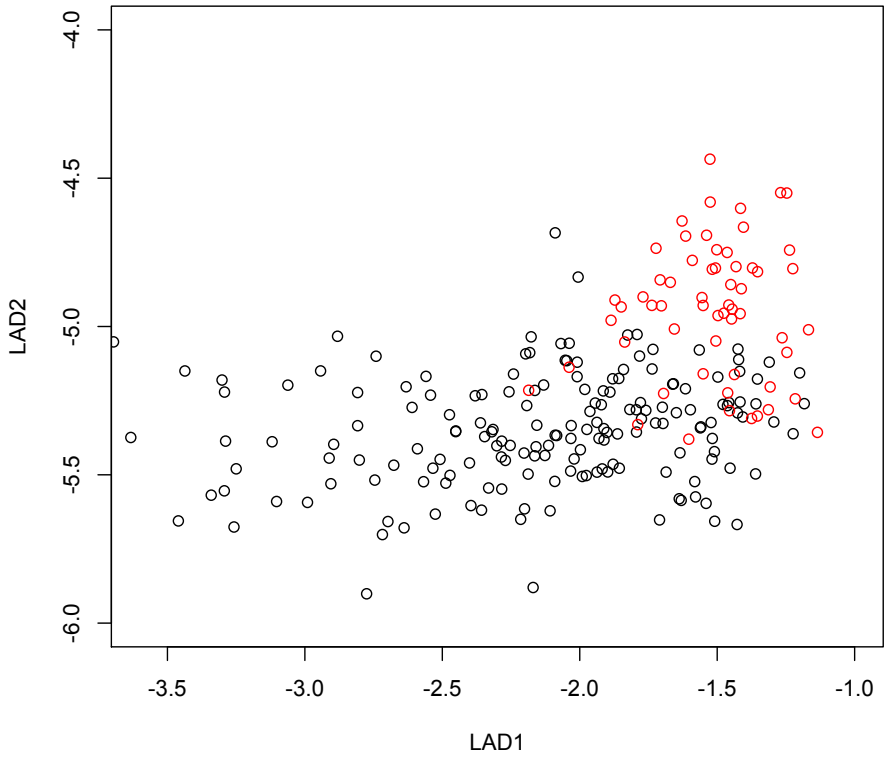


Porcentaje de bien clasificado: 75 %.

Cánder de pulmón data - LAD

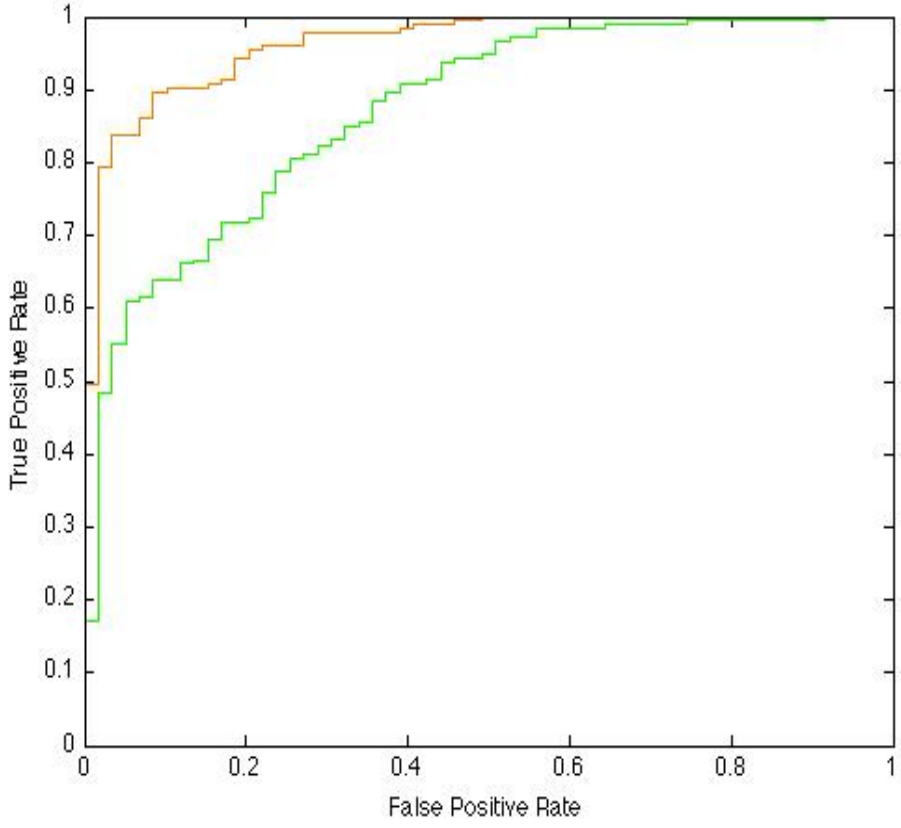


Cánder de pulmón data - LAD



Porcentaje de bien clasificado: 83 %.

curva ROC para datos de cáncer



PC, AUC=.8710 and LAD, AUC= 0.9628

Limitaciones. Datos: Atletas de Autralia

- Objetivo: investigar la relación entre el índice de masa corporal (Y) y varios predictores e identificar los factores que están asociados al sobrepeso
- Variables predictoras: altura, peso, cant. de glóbulos rojos, cant. de glóbulos blancos, hemoglobina $(\mathbf{X})$ y sexo (W)
- $n = 102$

Limitaciones: PFC y LAD es para predictores continuos

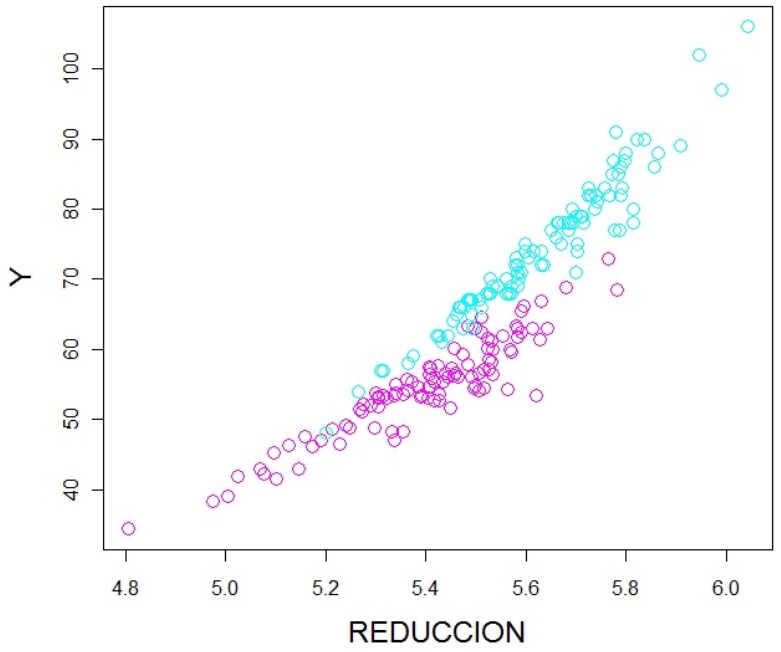
Solución: trabajar con cada nivel de la variable no continua: y considerar

Reducción suficiente: $R((\mathbf{X}, W)) = (\alpha^T \mathbf{X}, W)$ con $\alpha : 5 \times 1$

Limitaciones: PFC y LAD es para predictores continuos

Solución: trabajar con cada nivel de la variable no continua: y considerar

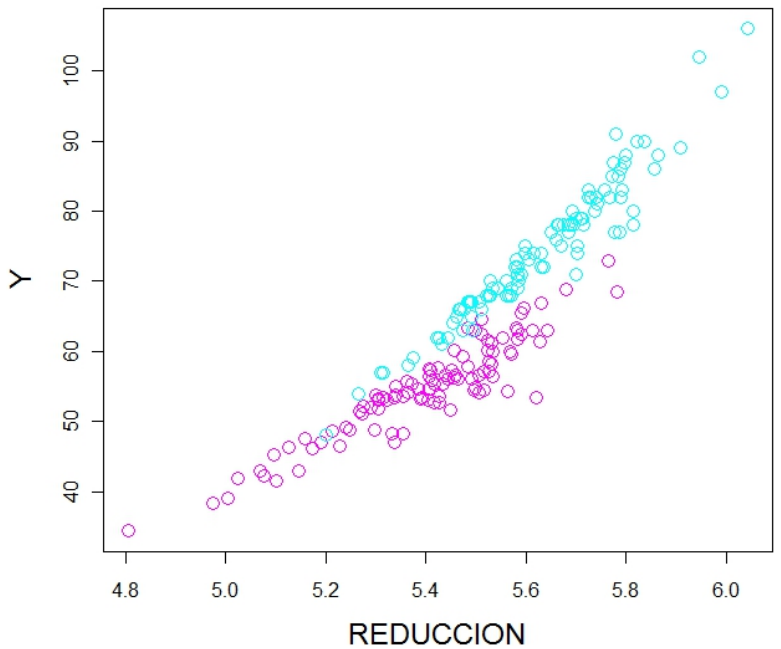
Reducción suficiente: $R((\mathbf{X}, W)) = (\alpha^T \mathbf{X}, W)$ con $\alpha : 5 \times 1$



Limitaciones: PFC y LAD es para predictores continuos

Solución: trabajar con cada nivel de la variable no continua: y considerar

Reducción suficiente: $R((\mathbf{X}, W)) = (\alpha^T \mathbf{X}, W)$ con $\alpha : 5 \times 1$



Modelamos $Y|(R(\mathbf{X}), W)$

Datos: Atletas de Autralia - EFDR

Solución via modelos inversos

- Modelo de regresión inversa: $f_{\mathbf{X},W|Y} = f_{\mathbf{X}|W,Y} \cdot f_{W|Y}$

Datos: Atletas de Autralia - EFDR

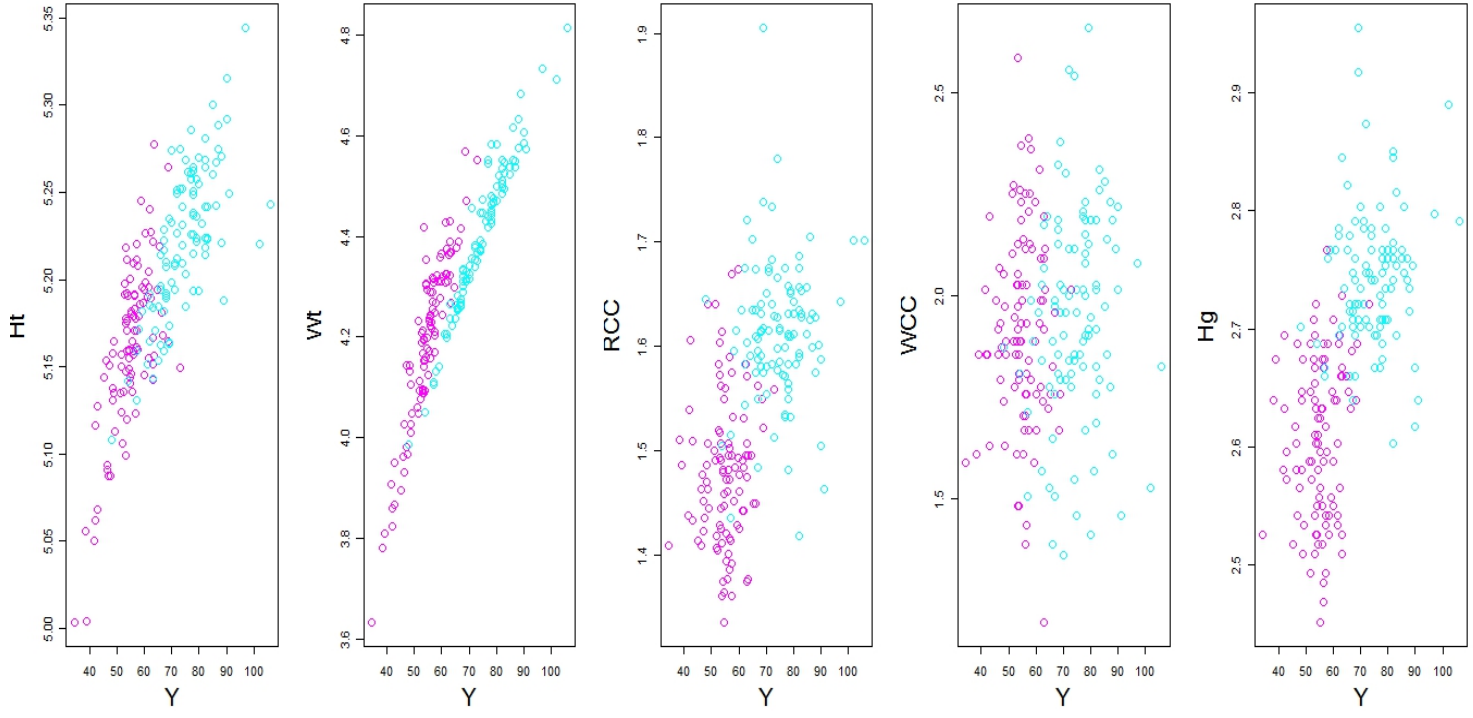
Solución via modelos inversos

- Modelo de regresión inversa: $f_{\mathbf{X},W|Y} = f_{\mathbf{X}|W,Y} \cdot f_{W|Y}$
  - $W|Y \sim$ Bernoulli(p_y)

Datos: Atletas de Autralia - EFDR

Solución via modelos inversos

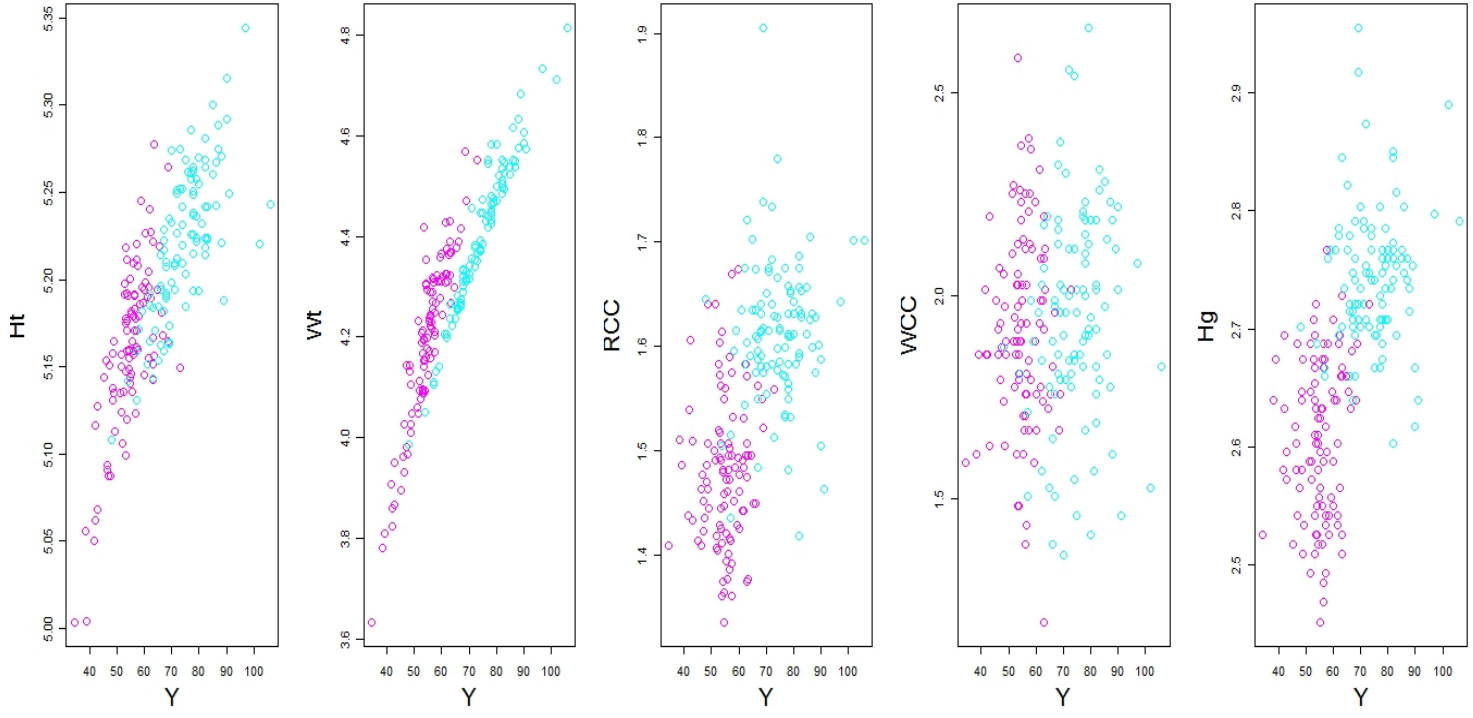
- Modelo de regresión inversa: $f_{\mathbf{X},W|Y} = f_{\mathbf{X}|W,Y} \cdot f_{W|Y}$
  - $W|Y \sim$ Bernoulli(p_y)



Datos: Atletas de Autralia - EFDR

Solución via modelos inversos

- Modelo de regresión inversa: $f_{\mathbf{X},W|Y} = f_{\mathbf{X}|W,Y} \cdot f_{W|Y}$
  - $W|Y \sim$ Bernoulli(p_y)
  - $\mathbf{X}|(Y,W) \sim \mathcal{N}(\mu_{y,w}, \Delta)$ con
    $\mu_{y,w} = \mu_{\mathbf{X}} + b_1(y - \mu_Y) + b_2(w - \mu_W)$

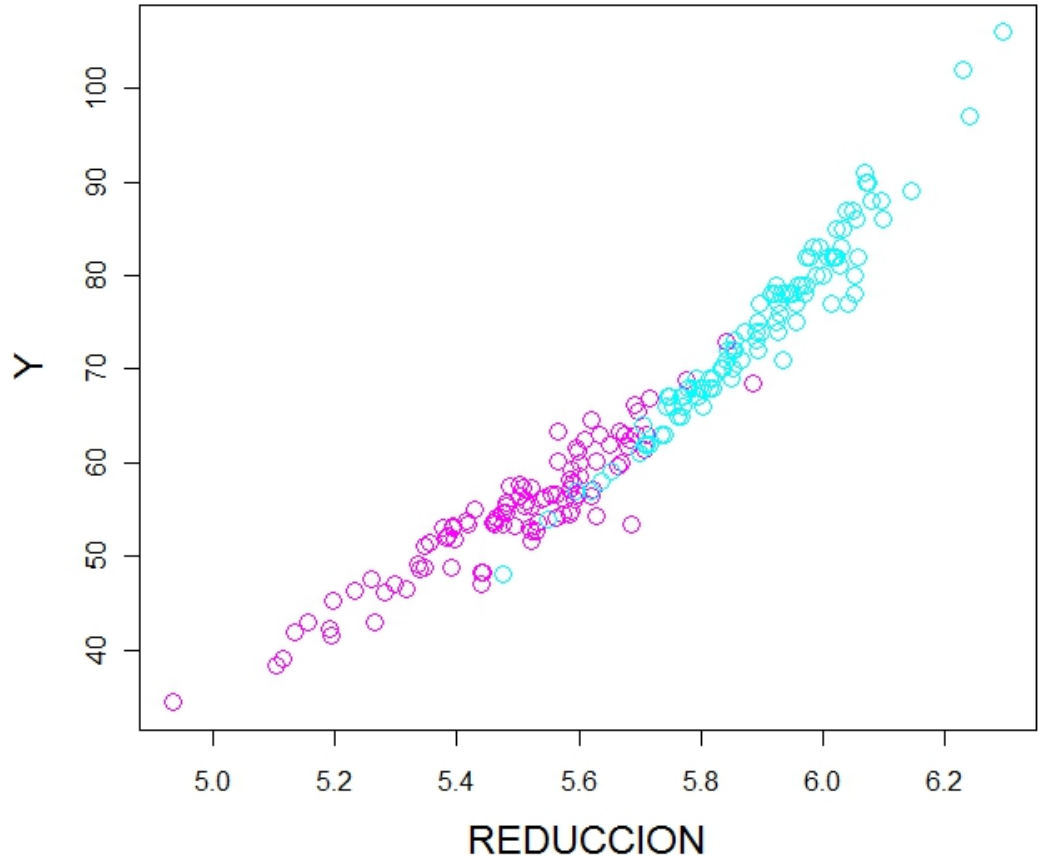


Datos: Atletas de Autralia - EFDR

- Reducción suficiente: $R((\mathbf{X}, W)) = \alpha^T(\mathbf{X}, W)$ con $\alpha : 6 \times 1$.

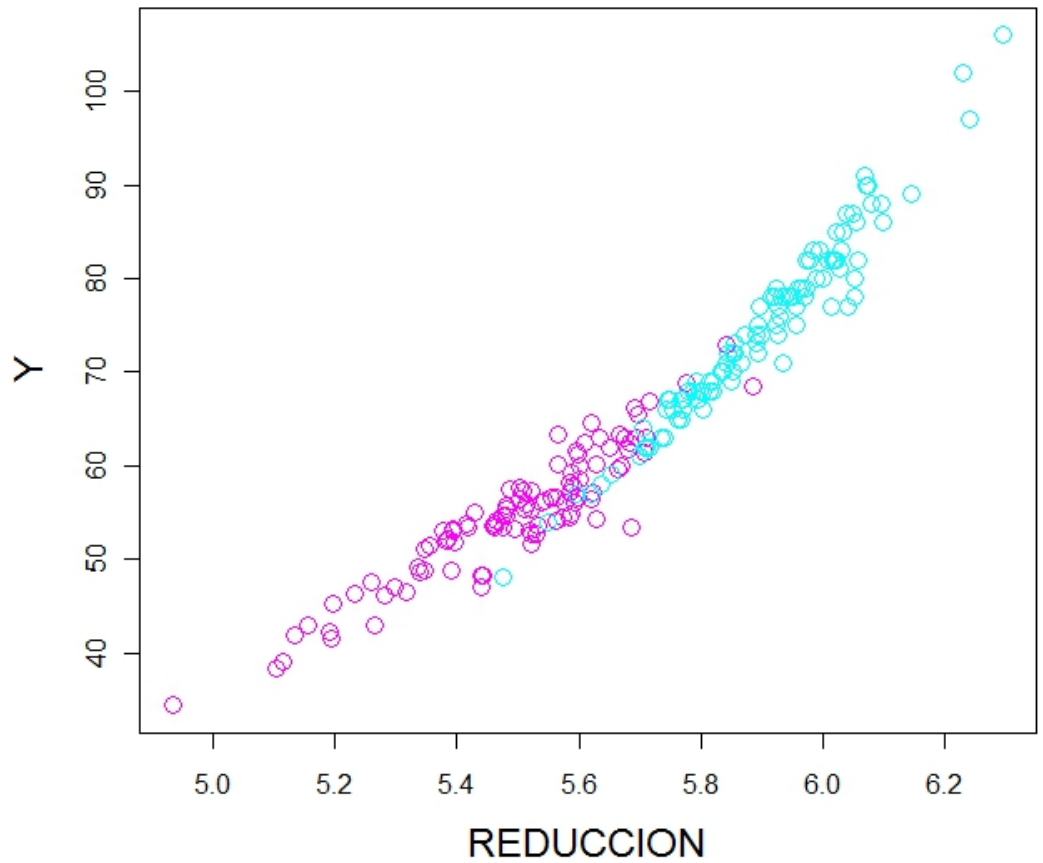
Datos: Atletas de Autralia - EFDR

- Reducción suficiente: $R((\mathbf{X}, W)) = \alpha^T(\mathbf{X}, W)$ con $\alpha : 6 \times 1$.



Datos: Atletas de Autralia - EFDR

- Reducción suficiente: $R((\mathbf{X}, W)) = \alpha^T(\mathbf{X}, W)$ con $\alpha : 6 \times 1$.



- Regresión directa:

$$E(Y|\mathbf{X}, W) = E(Y|R(\mathbf{X}, W)) = \gamma_1 + \gamma_2 R(\mathbf{X}, W) + \gamma_3 R^2(\mathbf{X}, W)$$

Nota sobre EFDR

Las reducciones no son más necesariamente lineales en $\mathbf{X}$

Ejemplo: biomarcadores de cáncer de pulmón

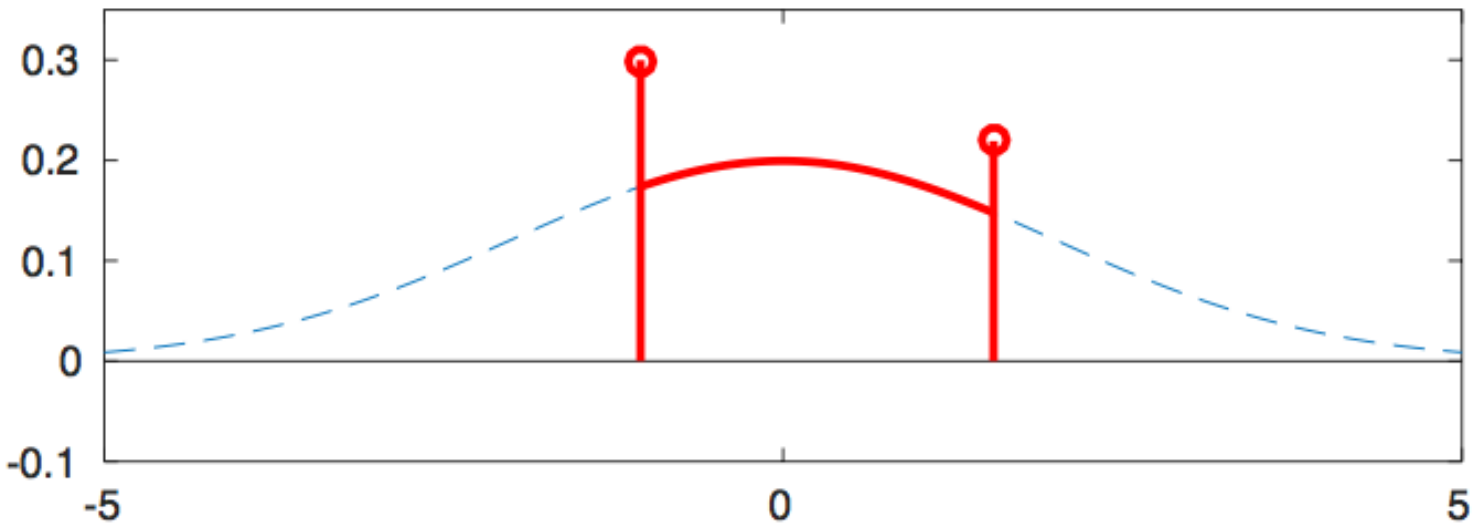
Descripción de los datos:

- 509 casos positivos ($Y = 1$) y 606 casos de control ($Y = 0$).
- 47 biomarcadores altamente correlacionados.
- Mediciones sujetas a límites de detección de las técnicas analticas.

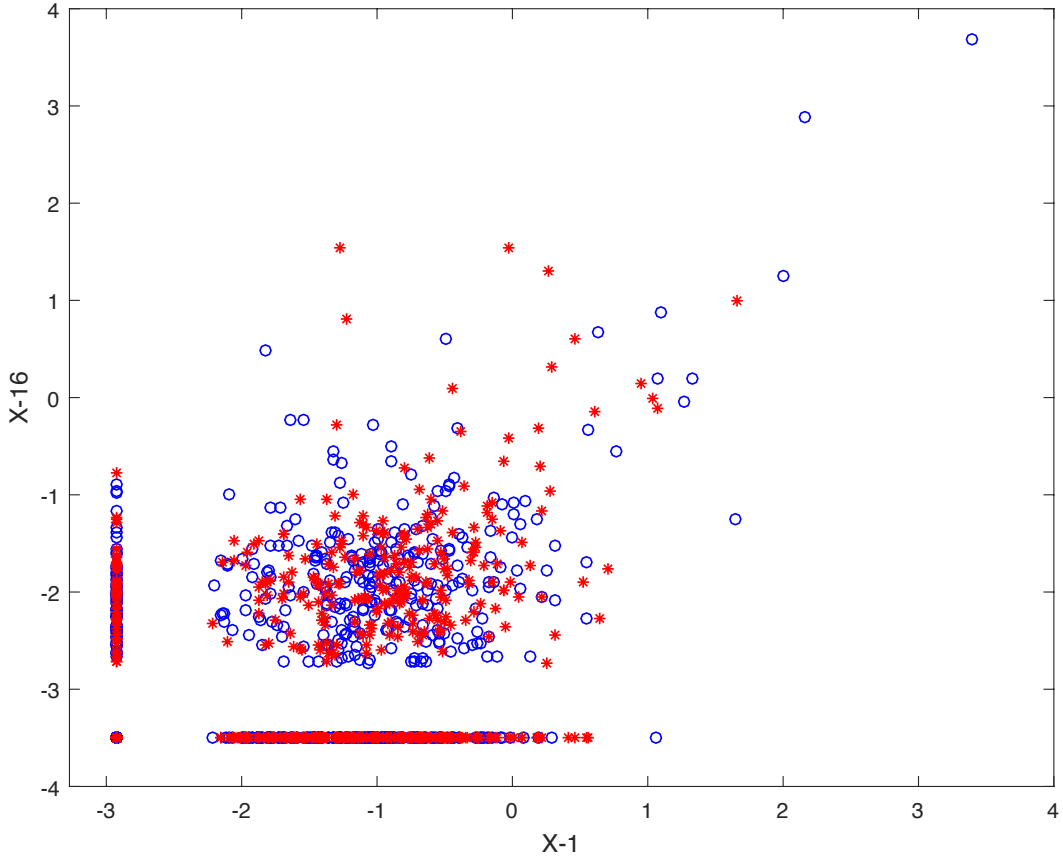
Ejemplo: biomarcadores de cáncer de pulmón

Descripción de los datos:

- 509 casos positivos ($Y = 1$) y 606 casos de control ($Y = 0$).
- 47 biomarcadores altamente correlacionados.
- Mediciones sujetas a límites de detección de las técnicas analticas.



dos biomarcadores vs Y



Ejemplo: biomarcadores de cáncer de pulmón

Descripción de los datos:

- 509 casos positivos ($Y = 1$) y 606 casos de control ($Y = 0$).
- 47 biomarcadores altamente correlacionados.
- Mediciones sujetas a lmites de detección de las técnicas analticas.

Objetivos del análisis:

- Combinar biomarcadores en unas pocas nuevas variables que faciliten el modelado y clasificación.
- Identificar los biomarcadores realmente relacionados con cáncer de pulmón y remover el resto.

Resultados usando las técnicas para variables continuas

Resultados de predicción (20-fold CV) Logística:

PREDICTORES	ERROR DE PREDICCIÓN		
	$\hat{d} = 1$	$\hat{d} = 6$	$p = 47$
Sin reducción			
$\mathbf{X}$			.46
MI-$\mathbf{X}$ (con imputación)			.45
Con reducción convencional			
PFC	.43		
LAD	.48	.40	

¿Qué pasa si tenemos en cuenta la presencia de límites de detección

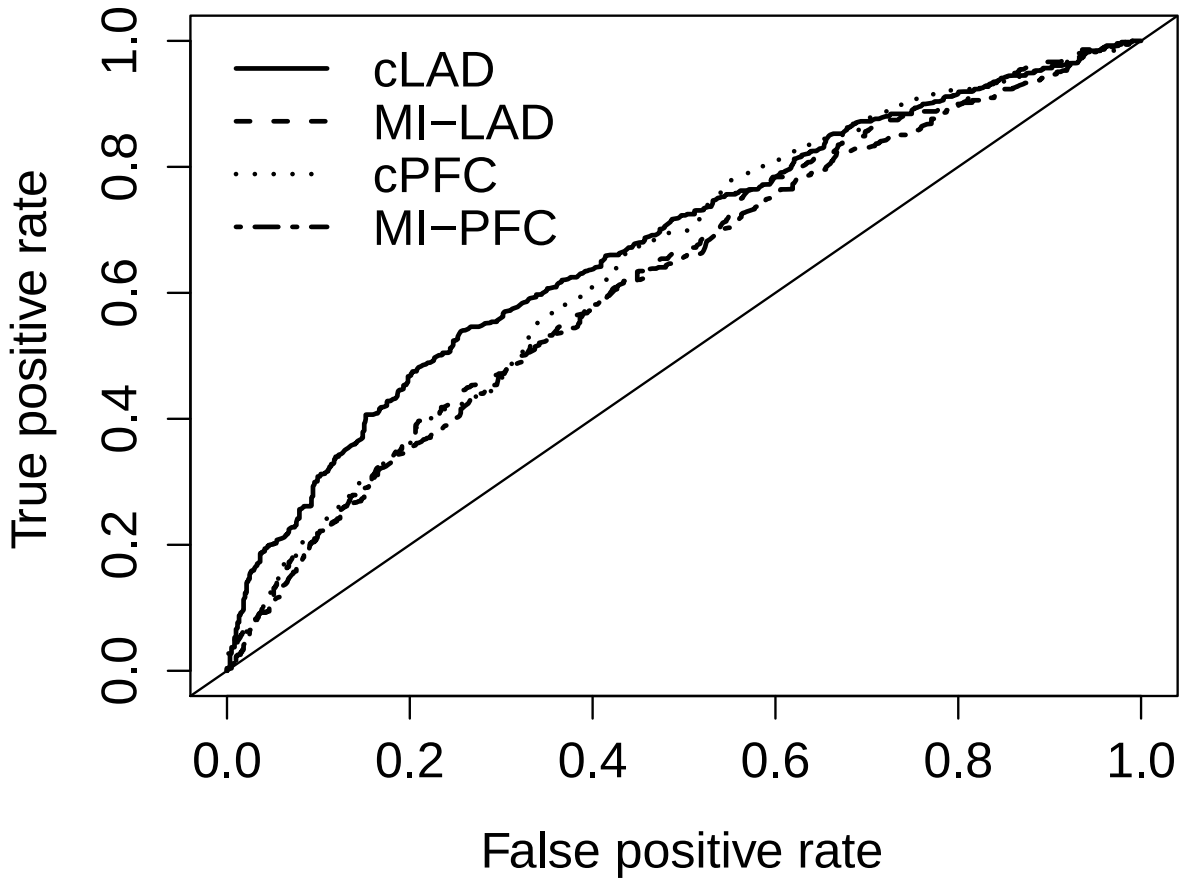
- Tener en cuenta la presencia de límites de detección. Extensión de los métodos para datos continuos
- ¿Ganamos algo?

Ejemplo: biomarcadores de cáncer de pulmón

Resultados de predicción (20-fold CV) con Logística:

PREDICTORES	ERROR DE PREDICCIÓN		
	$\hat{d} = 1$	$\hat{d} = 6$	$p = 47$
Sin reducción			
$\mathbf{X}$			.46
MI-$\mathbf{X}$ (con imputación)			.45
Con reducción convencional			
PFC	.43		
LAD	.48	.40	
Con reducción adaptada			
cPFC	**.39**		
cLAD	.42	**.35**	

Ejemplo: biomarcadores de cáncer de pulmón



Identificación de biomarcadores relevantes

- En una reducción convencional, todos los predictores originales forman parte de cada una de las nuevas direcciones.
- Cuando nos interesa **explicar** la relación entre predictores y respuesta, es útil seleccionar variables relevantes.
- **Importante:** bajo el enfoque de suficiencia podemos seleccionar variables sin depender de un modelo predictivo (como LASSO y sus parientes), pero con mayor estabilidad que los métodos paso-a-paso.

Identificación de biomarcadores relevantes

Table: Marcadores seleccionados basado en máxima verosimilitud penalizada. Los marcadores resaltados en rojo son los identificados experimentalmente (*Shiels et al. 2013*). Nuestra mayor contribución

Marker	**case-control status (binary Y)**				
	cLAD	LAD	MI-LAD	cPFC	stepwise LR
X6CKINE	•	•	•		
CRP	•			•	•
CTACK	•	•	•		
CXCL11ITAC					
CXCL9MIG	•	•	•	•	
EGFR					•
FGF2	•				•
GRO		•	•		
IL1RA	•				
MIP1B		•	•		
SEGFR	•	•	•		•
SVEGFR2	•	•	•		
TARC	•				•
TNFB	•				
VEGF					•
VEGFR2					•

Ejemplo. Los famosos índices sociales

Asignar una ayuda económica a hogares o individuos que viven en situación de pobreza

Ejemplo. Los famosos índices sociales

Asignar una ayuda económica a hogares o individuos que viven en situación de pobreza

Políticas o Programas Focalizados

(Ejemplos reales: CAS in Chile, Sisben en Colombia, SISFOH en Perú, Tekoporá en Paraguay, SIERP en Honduras, PANES en Uruguay, entre otros)

Ejemplo. Los famosos índices sociales

Asignar una ayuda económica a hogares o individuos que viven en situación de pobreza

Políticas o Programas Focalizados

(Ejemplos reales: CAS in Chile, Sisben en Colombia, SISFOH en Perú, Tekoporá en Paraguay, SIERP en Honduras, PANES en Uruguay, entre otros)

Aquí *pobreza* es *pobreza monetaria*

Un hogar j es pobre si su ingreso monetario Y_j no alcanza para cubrir las necesidades básicas, i.e. $Y_j \leq LP$ donde LP es el ingreso que determina la *línea de pobreza*.

Solución con Situación Ideal

- Conocemos el valor exacto de ingreso monetario de cada hogar $\rightsquigarrow$ Si $Y_j \leq LP$ le asigno la ayuda prevista.

Solución con Situación Ideal

- Conocemos el valor exacto de ingreso monetario de cada hogar $\rightsquigarrow$ Si $Y_j \leq LP$ le asigno la ayuda prevista.

Pero...

Solución con Situación Ideal

- Conocemos el valor exacto de ingreso monetario de cada hogar $\rightsquigarrow$ Si $Y_j \leq LP$ le asigno la ayuda prevista.

Pero...

- Existen varios problemas en la captación del ingreso como medida creíble o fiable (*reporting biases*):
  - Incentivos a revelar el valor verdadero.
  - Economía informal u oculta (ej. pagos en especie(trueque), auto-provisión).
  - Estacionalidad (ej. changas, empleos rurales).

¿Qué hacer?

Construir un índice ($I \in \mathbb{R}$) como *proxy* de ingreso o riqueza

Si $I \leq LP^*$ asigno la ayuda (para un LP^* determinado)

El *cómo*

Miramos otras variables del hogar, de observación más directa y en general más fáciles de recolectar, que en conjunto son *proxy* del bienestar económico del mismo. Ej:

El *cómo*

Miramos otras variables del hogar, de observación más directa y en general más fáciles de recolectar, que en conjunto son *proxy* del bienestar económico del mismo. Ej:

- Vivienda (materiales del techo, materiales del suelo, forma de acceso al agua potable, etc.).
- Activos físicos (tiene radio?, tiene TV?, tiene internet? tiene moto? tienen auto?)
- Otras socio-demográficas (cant. de miembros, escolaridad, situación ocupacional, etc.)

El *cómo*

Miramos otras variables del hogar, de observación más directa y en general más fáciles de recolectar, que en conjunto son *proxy* del bienestar económico del mismo. Ej:

- Vivienda (materiales del techo, materiales del suelo, forma de acceso al agua potable, etc.).
- Activos físicos (tiene radio?, tiene TV?, tiene internet? tiene moto? tienen auto?)
- Otras socio-demográficas (cant. de miembros, escolaridad, situación ocupacional, etc.)

Definición del índice (I)

Sean $X_1, \ldots, X_p$ diferentes variables que caracterizan al hogar en términos económicos y sociales, se define por índice de estatus socieconómico (ESE), a la combinación lineal de dichas variables para un determinado vector de pesos fijos $(a_1, \ldots, a_p)$, i.e.

$$I = a_1 X_1 + \cdots + a_p X_p$$

Solución con Situación Casi Ideal

- Un Ser omnisciente me dicta los verdaderos valores de los pesos: $a_1^*, a_2^*, \ldots, a_p^*$.

Solución con Situación Casi Ideal

- Un Ser omnisciente me dicta los verdaderos valores de los pesos: $a_1^*, a_2^*, \ldots, a_p^*$.

- Para un cierto hogar j miro sus $X's$, i.e. si tiene auto, internet, TV, materiales del techo, de los pisos, etc.

Solución con Situación Casi Ideal

- Un Ser omnisciente me dicta los verdaderos valores de los pesos: $a_1^*, a_2^*, \ldots, a_p^*$.
- Para un cierto hogar j miro sus $X's$, i.e. si tiene auto, internet, TV, materiales del techo, de los pisos, etc.
- Luego computo el índice $a_1^* auto + a_2^* internet + a_3^* TV + a_4^* techo + a_5^* pisos + \ldots$ y ya puedo clasificar.

Solución con Situación Casi Ideal

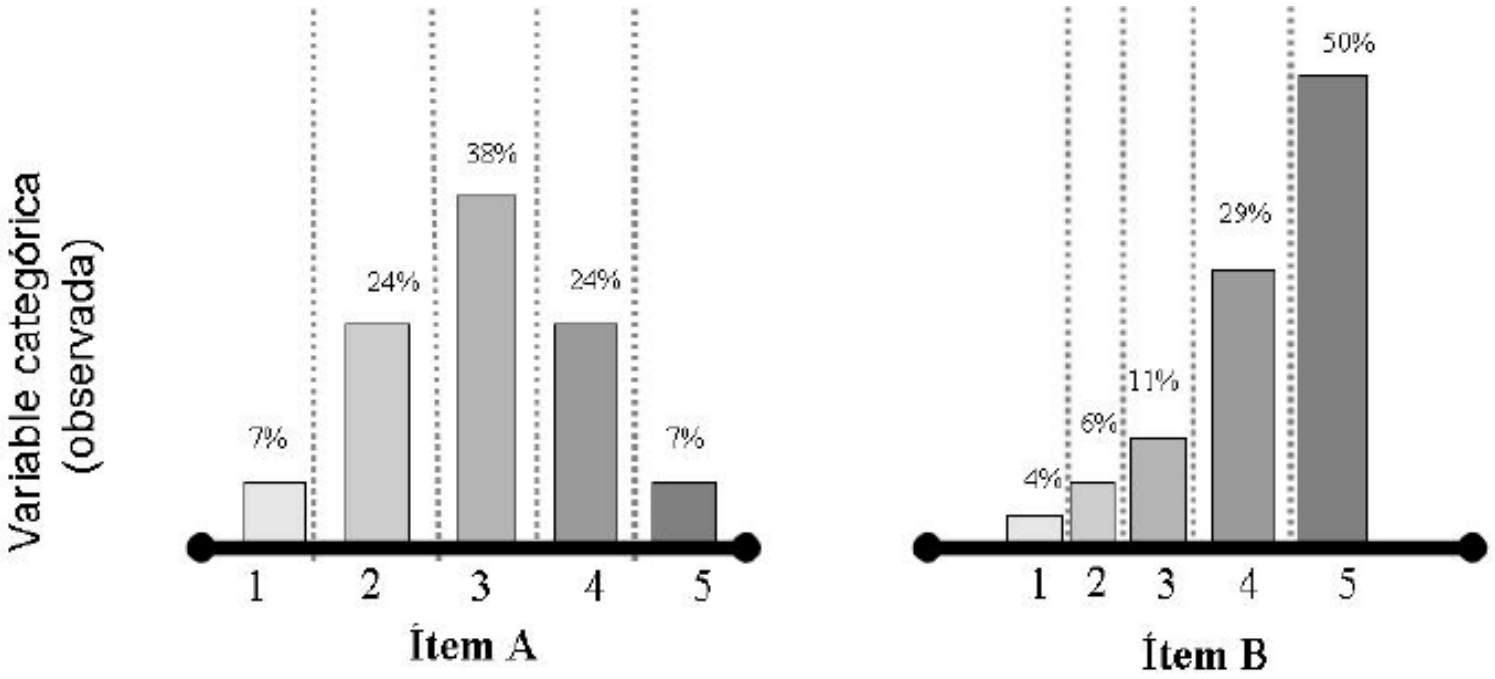
- Un Ser omnisciente me dicta los verdaderos valores de los pesos: $a_1^*, a_2^*, \ldots, a_p^*$.
- Para un cierto hogar j miro sus $X's$, i.e. si tiene auto, internet, TV, materiales del techo, de los pisos, etc.
- Luego computo el índice $a_1^* auto + a_2^* internet + a_3^* TV + a_4^* techo + a_5^* pisos + \ldots$ y ya puedo clasificar.
- Dura realidad: Con una muestra (denominada de entrenamiento) debo estimar $(a_1, \ldots, a_p)$ y que me de un buen índice para implementar de mejor manera el programa focalizado (i.e. evitar **falsos-positivos o falsos-negativos**)

Características de las variables socio-económicas

Ejemplos:

- *pisos*= 1 (tierra/ladrillos sueltos), 2 (cemento o ladrillo fijo), 3 (mosaicos o baldosas), 4(madera/cerámica y alfombra).
- *techo*= 1 (caña o paja), 2 (chapa de cartón), 3 (chapa de fibrocemento o de metal),...
- *agua*= 1 (perforación con bomba manual), 2 (perforación con bomba a motor), 3(red pública).
- *baño*= 1 (letrina), 2 (inodoro sin botón o cadena -balde), 3 (inodoro botón/cadena/mochila).
- *TV-CPU-auto-moto*= 1 (no tiene), 2 (si tiene).
- *escolaridad*=: 1 (sin inst.), 2(primaria incompleta), 3 (primaria completa), 4 (secundaria incomp.),..

¿Cómo son estas variables?



Se llaman *variables ordinales*

Solución Real Convencional para conseguir un índice

- **Componentes principales adaptado a que las variables son ordinales, ej. Kolenikov & Angeles (2009)**

Solución Real Convencional para conseguir un índice

- **Componentes principales adaptado a que las variables son ordinales, ej. Kolenikov & Angeles (2009)**

- ¿Qué ocurre si en la muestra de entrenamiento tenemos información de la variable respuesta que buscamos predecir (e.j. ingreso o pobreza)? $\rightsquigarrow$ PCA no usa tal información!

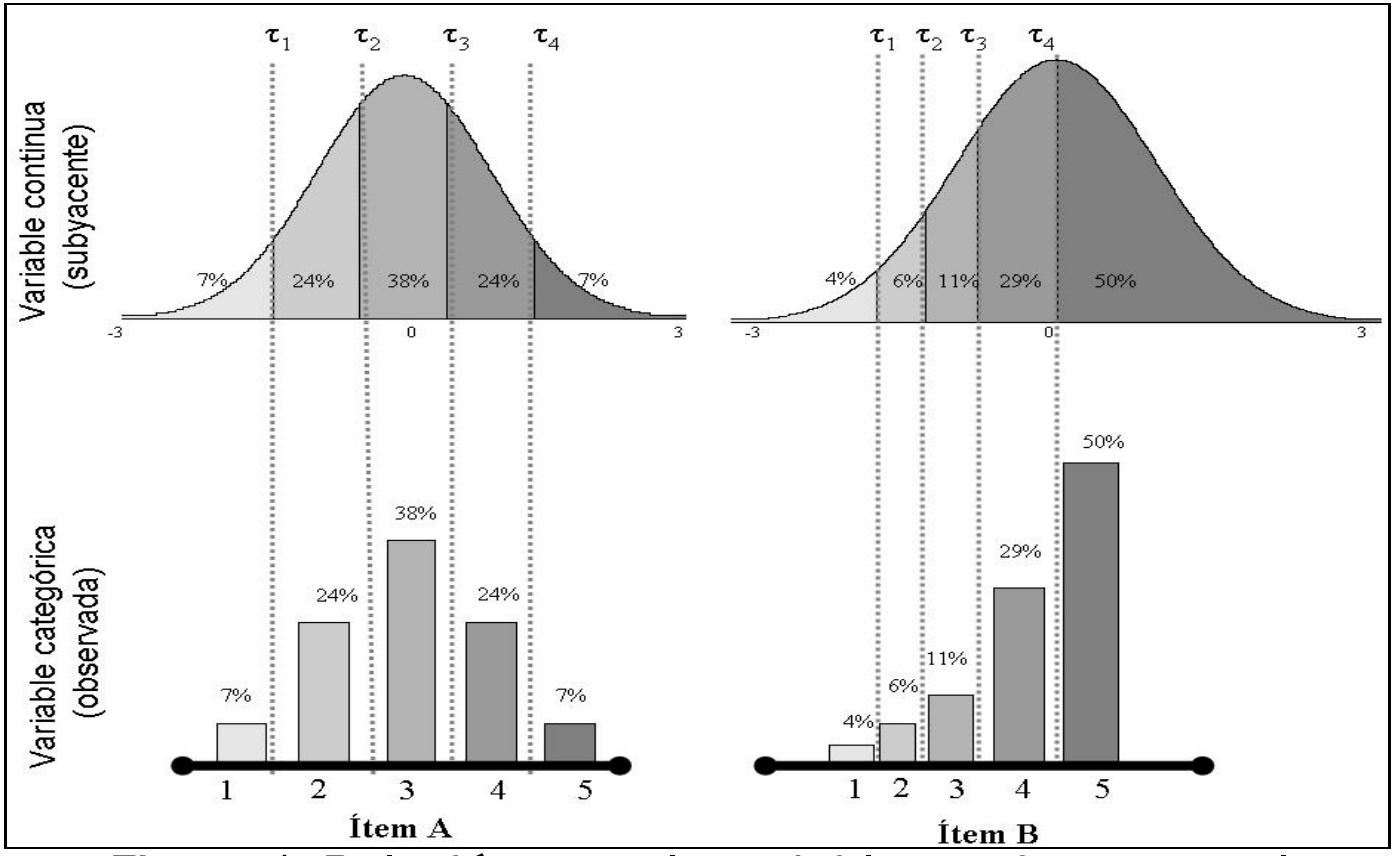
Problema Estadístico

- Usar los métodos de reduccion suficiente para variables continuas

Problema Estadístico

- Usar los métodos de reduccion suficiente para variables continuas
- Desarrollar un método de reducción suficiente para $\mathbf{X}$ ordinales

¿Cómo podemos extender el método a variables ordinales?



RSD para Categorías Ordinales

- p variables aleatorias ordinales $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$, con $X_j \in \{1, 2, \ldots, K_j\}$;

RSD para Categorías Ordinales

- p variables aleatorias ordinales $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$, con $X_j \in \{1, 2, \ldots, K_j\}$;
- Respuesta $Y \in \mathbb{R}$;

RSD para Categorías Ordinales

- p variables aleatorias ordinales $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$, con $X_j \in \{1, 2, \ldots, K_j\}$;
- Respuesta $Y \in \mathbb{R}$;
- p variables latentes $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)^T$;

RSD para Categorías Ordinales

- p variables aleatorias ordinales $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$, con $X_j \in \{1, 2, \ldots, K_j\}$;
- Respuesta $Y \in \mathbb{R}$;
- p variables latentes $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)^T$;
- Conjunto de umbrales $\Theta = \{\theta_0^{(j)}, \theta_1^{(j)}, \ldots, \theta_{K_j}^{(j)}\}$, donde $-\infty = \theta_0^{(j)} < \theta_1^{(j)} < \cdots < \theta_{K_j-1}^{(j)} < \theta_{K_j}^{(j)} = +\infty$, $j = 1, 2, \ldots, p$;

RSD para Categorías Ordinales

- p variables aleatorias ordinales $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$, con $X_j \in \{1, 2, \ldots, K_j\}$;
- Respuesta $Y \in \mathbb{R}$;
- p variables latentes $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)^T$;
- Conjunto de umbrales $\Theta = \{\theta_0^{(j)}, \theta_1^{(j)}, \ldots, \theta_{K_j}^{(j)}\}$, donde $-\infty = \theta_0^{(j)} < \theta_1^{(j)} < \cdots < \theta_{K_j-1}^{(j)} < \theta_{K_j}^{(j)} = +\infty$, $j = 1, 2, \ldots, p$;
- Las variables ordinales $\mathbf{X}$ y las variables latentes $\mathbf{Z}$ están relacionadas mediante:

$$\Pr(X_j = k|Y) = \Pr(\theta_{k-1}^{(j)} \leq Z_j < \theta_k^{(j)}|Y);$$

RSD para Categorías Ordinales

- p variables aleatorias ordinales $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$, con $X_j \in \{1, 2, \ldots, K_j\}$;
- Respuesta $Y \in \mathbb{R}$;
- p variables latentes $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)^T$;
- Conjunto de umbrales $\Theta = \{\theta_0^{(j)}, \theta_1^{(j)}, \ldots, \theta_{K_j}^{(j)}\}$, donde $-\infty = \theta_0^{(j)} < \theta_1^{(j)} < \cdots < \theta_{K_j-1}^{(j)} < \theta_{K_j}^{(j)} = +\infty$, $j = 1, 2, \ldots, p$;
- Las variables ordinales $\mathbf{X}$ y las variables latentes $\mathbf{Z}$ están relacionadas mediante:

$$\Pr(X_j = k|Y) = \Pr(\theta_{k-1}^{(j)} \leq Z_j < \theta_k^{(j)}|Y);$$

- Se asume $\mathbf{Z}|Y \sim N(\beta f_Y, \Delta)$

RSD para Categorías Ordinales

- p variables aleatorias ordinales $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$, con $X_j \in \{1, 2, \ldots, K_j\}$;
- Respuesta $Y \in \mathbb{R}$;
- p variables latentes $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)^T$;
- Conjunto de umbrales $\Theta = \{\theta_0^{(j)}, \theta_1^{(j)}, \ldots, \theta_{K_j}^{(j)}\}$, donde $-\infty = \theta_0^{(j)} < \theta_1^{(j)} < \cdots < \theta_{K_j-1}^{(j)} < \theta_{K_j}^{(j)} = +\infty$, $j = 1, 2, \ldots, p$;
- Las variables ordinales $\mathbf{X}$ y las variables latentes $\mathbf{Z}$ están relacionadas mediante:

$$\Pr(X_j = k|Y) = \Pr(\theta_{k-1}^{(j)} \leq Z_j < \theta_k^{(j)}|Y);$$

- Se asume $\mathbf{Z}|Y \sim N(\beta f_Y, \Delta)$
- Con la ayuda de la teoría de normalidad encontramos la reducción suficiente para las observadas $\mathbf{X}$.

Aplicación usando EPH-Argentina

- EPH 3er. trimestre de 2013 (INDEC).

Aplicación usando EPH-Argentina

- EPH 3er. trimestre de 2013 (INDEC).
- 9 predictores ordinales.

Aplicación usando EPH-Argentina

- EPH 3er. trimestre de 2013 (INDEC).
- 9 predictores ordinales.
- Heterogeneidad regional $\rightsquigarrow$ Estimación de diferentes I para 5 regiones (GBA, Pampeana, NOA, NEA y Patagonia).

Aplicación usando EPH-Argentina

- EPH 3er. trimestre de 2013 (INDEC).
- 9 predictores ordinales.
- Heterogeneidad regional $\rightsquigarrow$ Estimación de diferentes I para 5 regiones (GBA, Pampeana, NOA, NEA y Patagonia).
- 2 tipos de respuesta (Y): una continua (ingreso per cápita del hogar) y una dicotómica (pobreza).

Aplicación usando EPH-Argentina

- EPH 3er. trimestre de 2013 (INDEC).
- 9 predictores ordinales.
- Heterogeneidad regional $\rightsquigarrow$ Estimación de diferentes I para 5 regiones (GBA, Pampeana, NOA, NEA y Patagonia).
- 2 tipos de respuesta (Y): una continua (ingreso per cápita del hogar) y una dicotómica (pobreza).
- Estimación: EM + selección de variables .

Aplicación usando EPH-Argentina

- EPH 3er. trimestre de 2013 (INDEC).
- 9 predictores ordinales.
- Heterogeneidad regional $\rightsquigarrow$ Estimación de diferentes I para 5 regiones (GBA, Pampeana, NOA, NEA y Patagonia).
- 2 tipos de respuesta (Y): una continua (ingreso per cápita del hogar) y una dicotómica (pobreza).
- Estimación: EM + selección de variables .
- Evaluación: El método aquí propuesto PFCORD *vs.* 2 variantes del PCA que contemplan predictores categóricos: no lineal (NLPCA) y con polychoric correlations (PCAPOLY)

Performance del método

Table: 10-fold MSE para el índice I

Response	Method	Prediction Errors -MSE *Buenos Aires*	*Humid Pampas*	*Northwest*	*Northeast*	*Patagonia*
Per capita Income (continuous)	ORIG-I	7.22	4.69	4.68	3.32	13.14
		(3.50)	(0.88)	(2.47)	(0.93)	(3.34)
	PCAPOLY	7.60	5.10	5.07	3.68	14.7
		(2.45)	(0.90)	(1.77)	(0.90)	(4.01)
	NLPCA	7.38	4.95	4.89	3.52	13.67
		(2.29)	(0.61)	(1.48)	(0.65)	(3.71)
	PFCORD	7.26	4.73	4.71	3.30	13.20
		(2.75)	(1.66)	(2.75)	(1.35)	(3.34)
Poverty (discrete)	ORIG-I	0.202	0.162	0.274	0.287	0.129
		(0.021)	(0.008)	(0.026)	(0.036)	(0.020)
	PCAPOLY	0.229	0.204	0.366	0.390	0.132
		(0.028)	(0.018)	(0.023)	(0.063)	(0.025)
	NLPCA	0.228	0.186	0.302	0.290	0.161
		(0.021)	(0.019)	(0.019)	(0.018)	(0.016)
	PFCORD	0.208	0.167	0.279	0.287	0.129
		(0.015)	(0.011)	(0.028)	(0.028)	(0.022)

Note: standard deviations in parentheses. Database: EPH (2013)

Performance del método

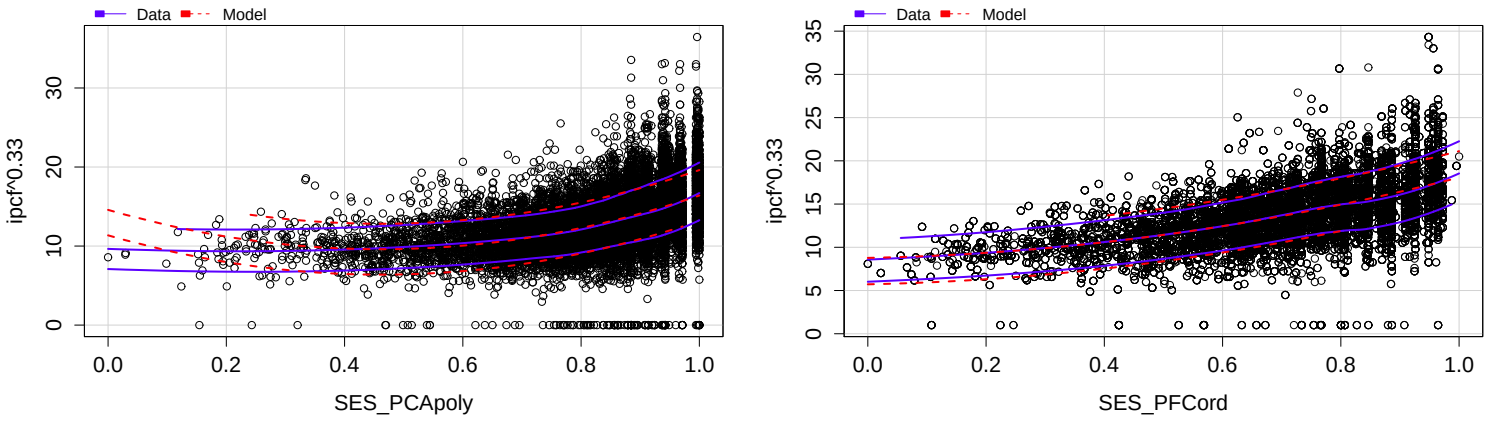


Figure: Comparación de Ajuste de modelo lineal ($Ingreso \sim I$) PCA (izquierda) vs. PFCORD (derecha)

Comparación de los ponderadores (**a**) de I

1. La importancia relativa que tiene cada variable (i.e. predictores) es diferente

Table: Comparación de índices para predecir ingreso per cápita en el hogar.

Variables	Buenos Aires		
	PFCORD	PCAPOLY	NLPCA
housing location	0	-0.1690	-0.0943
housing quality	-0.2455	-0.3768	-0.2199
sources of cooking fuel	-0.4637	-0.3788	-0.2080
overcrowding	-0.7222	-0.2888	-0.1788
schooling	-0.2860	-0.2275	-0.1474
toilet drainage	-0.1175	-0.3381	-0.2047
toilet facility	0	-0.4061	-0.2246
toilet sharing	0	-0.2759	-0.1186
water location	0	-0.3918	-0.1790
water source	0	-0.2023	-0.1033
working hours	-0.3278	0	0

Comparación de los ponderadores (**a**) de I

1. La importancia relativa que tiene cada variable (i.e. predictores) es diferente

Table: Comparación de índices para predecir ingreso per cápita en el hogar.

Variables	Buenos Aires		
	PFCORD	PCAPOLY	NLPCA
housing location	0	-0.1690	-0.0943
housing quality	-0.2455	-0.3768	-0.2199
sources of cooking fuel	-0.4637	-0.3788	-0.2080
overcrowding	-0.7222	-0.2888	-0.1788
schooling	-0.2860	-0.2275	-0.1474
toilet drainage	-0.1175	-0.3381	-0.2047
toilet facility	0	-0.4061	-0.2246
toilet sharing	0	-0.2759	-0.1186
water location	0	-0.3918	-0.1790
water source	0	-0.2023	-0.1033
working hours	-0.3278	0	0

Comparación de los ponderadores (**a**) de I

1. La importancia relativa que tiene cada variable (i.e. predictores) es diferente

Table: Comparación de índices para predecir ingreso per cápita en el hogar.

Variables	Buenos Aires		
	PFCORD	PCAPOLY	NLPCA
housing location	0	-0.1690	-0.0943
housing quality	-0.2455	-0.3768	-0.2199
sources of cooking fuel	-0.4637	-0.3788	-0.2080
overcrowding	-0.7222	-0.2888	-0.1788
schooling	-0.2860	-0.2275	-0.1474
toilet drainage	-0.1175	-0.3381	-0.2047
toilet facility	0	-0.4061	-0.2246
toilet sharing	0	-0.2759	-0.1186
water location	0	-0.3918	-0.1790
water source	0	-0.2023	-0.1033
working hours	-0.3278	0	0

Comparación de los ponderadores (**a**) de I

1. La importancia relativa que tiene cada variable (i.e. predictores) es diferente

Table: Comparación de índices para predecir ingreso per cápita en el hogar.

Variables	Buenos Aires		
	PFCORD	PCAPOLY	NLPCA
housing location	0	-0.1690	-0.0943
housing quality	-0.2455	-0.3768	-0.2199
sources of cooking fuel	-0.4637	-0.3788	-0.2080
overcrowding	-0.7222	-0.2888	-0.1788
schooling	-0.2860	-0.2275	-0.1474
toilet drainage	-0.1175	-0.3381	-0.2047
toilet facility	0	-0.4061	-0.2246
toilet sharing	0	-0.2759	-0.1186
water location	0	-0.3918	-0.1790
water source	0	-0.2023	-0.1033
working hours	-0.3278	0	0

Comparación de los ponderadores ($\mathbf{a}$) de I

2. Mientras que PCA da pesos muy similares para las diferentes regiones, el índice PFCORD logra captar las divergencias regionales

Comparación de los ponderadores (**a**) de I

2. Mientras que PCA da pesos muy similares para las diferentes regiones, el índice PFCORD logra captar las divergencias regionales

Ejemplo: PCA

Table: Indice PCAPOLY por regiones. Respuesta Ingreso per cápita

Variables	GBA	Pampeana	NOA	NEA	Patag.
	PCA	PCA	PCA	PCA	PCA
housing location	-0.1690	-0.1903	-0.1068	-0.1809	-0.1437
housing quality	-0.3768	-0.3557	-0.3278	-0.3727	-0.3258
sources of cooking fuel	-0.3788	-0.3609	-0.3287	-0.1648	-0.4026
overcrowding	-0.2888	-0.2351	-0.1991	-0.2025	-0.2207
schooling	-0.2275	-0.2075	-0.2197	-0.1869	-0.1284
toilet drainage	-0.3381	-0.3519	-0.3623	-0.3572	-0.4122
toilet facility	-0.4061	-0.4105	-0.4217	-0.4383	-0.4376
toilet sharing	-0.2759	-0.3176	-0.2579	-0.2921	-0.2937
water location	-0.3918	-0.3933	-0.4202	-0.4227	-0.4169
water source	-0.2023	-0.2461	-0.3646	-0.3733	-0.1525
working hours	0	0	0	0	0

Comparación de los ponderadores (**a**) de I

2. Mientras que PCA da pesos muy similares para las diferentes regiones, el índice PFCORD logra captar las divergencias regionales

Ejemplo: PCA

Table: Indice PCAPOLY por regiones. Respuesta Ingreso per cápita

Variables	GBA	Pampeana	NOA	NEA	Patag.
	PCA	PCA	PCA	PCA	PCA
housing location	-0.1690	-0.1903	-0.1068	-0.1809	-0.1437
housing quality	-0.3768	-0.3557	-0.3278	-0.3727	-0.3258
sources of cooking fuel	-0.3788	-0.3609	-0.3287	-0.1648	-0.4026
overcrowding	-0.2888	-0.2351	-0.1991	-0.2025	-0.2207
schooling	-0.2275	-0.2075	-0.2197	-0.1869	-0.1284
toilet drainage	**-0.3381**	**-0.3519**	**-0.3623**	**-0.3572**	**-0.4122**
toilet facility	**-0.4061**	**-0.4105**	**-0.4217**	**-0.4383**	**-0.4376**
toilet sharing	**-0.2759**	**-0.3176**	**-0.2579**	**-0.2921**	**-0.2937**
water location	**-0.3918**	**-0.3933**	**-0.4202**	**-0.4227**	**-0.4169**
water source	**-0.2023**	**-0.2461**	**-0.3646**	**-0.3733**	**-0.1525**
working hours	0	0	0	0	0

Comparación de los ponderadores (**a**) de I

Para nuestro método propuesto:

Table: Indice con PFCORD por regiones. Respuesta Ingreso per cápita

Variables	GBA	Pampeana	NOA	NEA	Patag.
	PFC	PFC	PFC	PFC	PFC
housing location	0	0	-0.1412	-0.1677	-0.1105
housing quality	-0.2455	-0.3077	-0.1195	-0.0858	-0.3341
sources of cooking fuel	-0.4637	-0.3735	-0.1340	-0.0920	-0.1926
overcrowding	-0.7222	-0.8086	-0.8556	-0.8367	-0.7447
schooling	-0.2860	-0.2703	-0.3556	-0.3364	-0.3807
toilet drainage	-0.1175	0	-0.1153	0	-0.1406
toilet facility	0	-0.1214	-0.0927	-0.0743	0
toilet sharing	0	0	0	-0.2369	-0.1242
water location	0	0	-0.2119	-0.1058	0
water source	0	0	-0.0787	-0.2423	0.1956
working hours	-0.3278	-0.1555	-0.1279	-0.1054	-0.2570

Comparación de los ponderadores (**a**) de I

Para nuestro método propuesto:

Table: Indice con PFCORD por regiones. Respuesta Ingreso per cápita

Variables	GBA	Pampeana	NOA	NEA	Patag.
	PFC	PFC	PFC	PFC	PFC
housing location	0	0	-0.1412	-0.1677	-0.1105
housing quality	-0.2455	-0.3077	-0.1195	-0.0858	-0.3341
sources of cooking fuel	-0.4637	-0.3735	-0.1340	-0.0920	-0.1926
overcrowding	-0.7222	-0.8086	-0.8556	-0.8367	-0.7447
schooling	-0.2860	-0.2703	-0.3556	-0.3364	-0.3807
toilet drainage	-0.1175	0	-0.1153	0	-0.1406
toilet facility	0	-0.1214	-0.0927	-0.0743	0
toilet sharing	0	0	0	-0.2369	-0.1242
water location	0	0	-0.2119	-0.1058	0
water source	0	0	-0.0787	-0.2423	0.1956
working hours	-0.3278	-0.1555	-0.1279	-0.1054	-0.2570

Comparación de los ponderadores (**a**) de I

3. Impacto de la variable respuesta

Table: PFCORD ordinal según Variable respuesta.

Respuesta:	Ingreso (pc)		Pobreza	
	GBA	PAMPEANA	GBA	PAMPEANA
housing location	0	0	0	0
housing quality	-0.2455	-0.3077	-0.3940	-0.3561
sources of cooking fuel	-0.4637	-0.3735	-0.4629	-0.3490
overcrowding	-0.7222	-0.8086	-0.6865	-0.7133
schooling	-0.2860	-0.2703	0	0
toilet drainage	-0.1175	0	0	0
toilet facility	0	-0.1214	-0.2426	-0.3538
toilet sharing	0	0	-0.1729	-0.1350
water location	0	0	-0.2231	-0.2612
water source	0	0	0	0
working hours	-0.3278	-0.1555	-0.2378	-0.1557

Comparación de los ponderadores (**a**) de I

3. Impacto de la variable respuesta

Table: PFCORD ordinal según Variable respuesta.

Respuesta:	Ingreso (pc)		Pobreza	
	GBA	PAMPEANA	GBA	PAMPEANA
housing location	0	0	0	0
housing quality	-0.2455	-0.3077	-0.3940	-0.3561
sources of cooking fuel	-0.4637	-0.3735	-0.4629	-0.3490
overcrowding	-0.7222	-0.8086	-0.6865	-0.7133
schooling	-0.2860	-0.2703	0	0
toilet drainage	-0.1175	0	0	0
toilet facility	0	-0.1214	-0.2426	-0.3538
toilet sharing	0	0	-0.1729	-0.1350
water location	0	0	-0.2231	-0.2612
water source	0	0	0	0
working hours	-0.3278	-0.1555	-0.2378	-0.1557

Comparación de los ponderadores (**a**) de I

3. Impacto de la variable respuesta

Table: PFCORD ordinal según Variable respuesta.

Respuesta:	Ingreso (pc)		Pobreza	
	GBA	PAMPEANA	GBA	PAMPEANA
housing location	0	0	0	0
housing quality	-0.2455	-0.3077	-0.3940	-0.3561
sources of cooking fuel	-0.4637	-0.3735	-0.4629	-0.3490
overcrowding	-0.7222	-0.8086	-0.6865	-0.7133
schooling	-0.2860	-0.2703	0	0
toilet drainage	-0.1175	0	0	0
toilet facility	0	-0.1214	-0.2426	-0.3538
toilet sharing	0	0	-0.1729	-0.1350
water location	0	0	-0.2231	-0.2612
water source	0	0	0	0
working hours	-0.3278	-0.1555	-0.2378	-0.1557

Conclusiones

- Creciente interés en el análisis cuantitativo de políticas y programas sociales.
- Indices de estatus socio-económico $\rightsquigarrow$ predominancia de PCA.
- El enfoque de RSD permite usar la información de la variable respuesta:
  - Mejor predicción de la respuesta de interés.
  - Diferentes ponderaciones respecto de PCA.
  - Capta diferencias regionales.
  - Sensibilidad ante la variable respuesta que caracteriza el fenómeno social de interés.

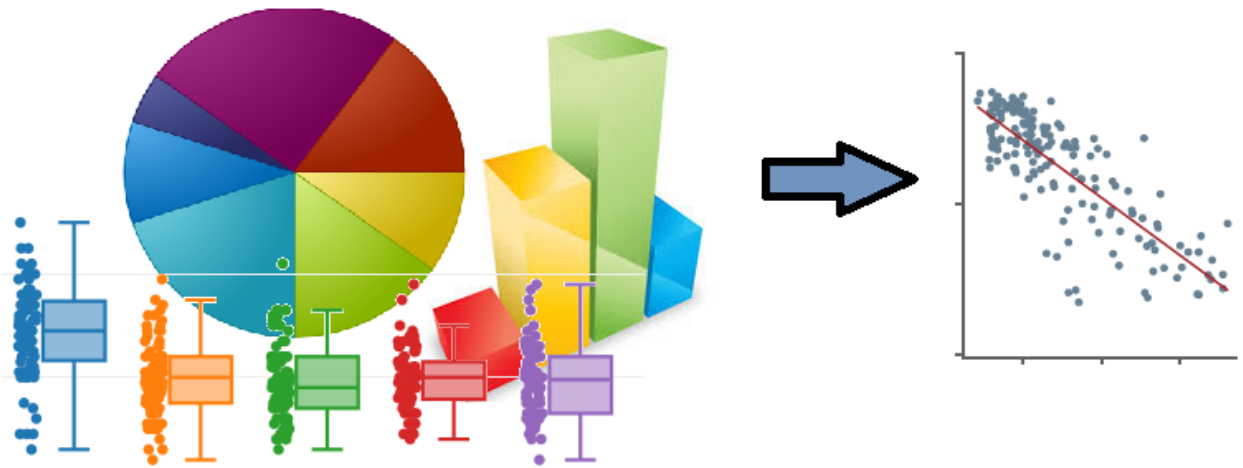
Extensiones

- Adaptaciones a predictores de distinta naturaleza.

Extensiones

- Adaptaciones a predictores de distinta naturaleza. Con el ejemplo de datos de cancer de pulmón vimos que se obtienen mejores resultados adaptando la metodologa de reduccin suficiente a la naturaleza de los predictores.

  Podemos hacer lo mismo cuando tenemos mezclas de variables continuas, categóricas, ordinales, dicotómicas, ...



Extensiones

- Adaptaciones a predictores de distinta naturaleza.
- Descubrir asociaciones entre variables en grandes bases de datos.

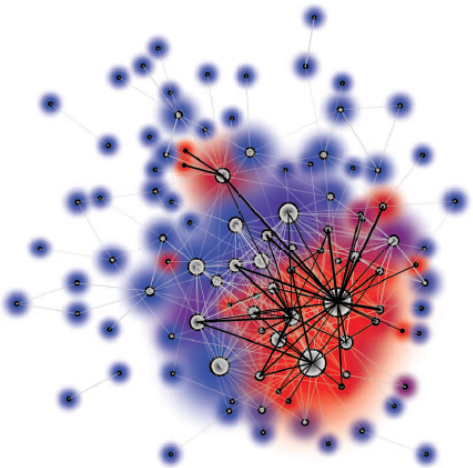
Extensiones

- Adaptaciones a predictores de distinta naturaleza.
- Descubrir asociaciones entre variables en grandes bases de datos.

  Objetivo: detectar interacciones y relaciones no lineales usando diferentes medidas de dependencia estadística propuestas recientemente y/o modelándolo como regresión multivariada

Science
AAAS

Detecting Novel Associations in Large Data Sets
David N. Reshef, *et al.*
Science **334**, 1518 (2011);
DOI: 10.1126/science.1205438

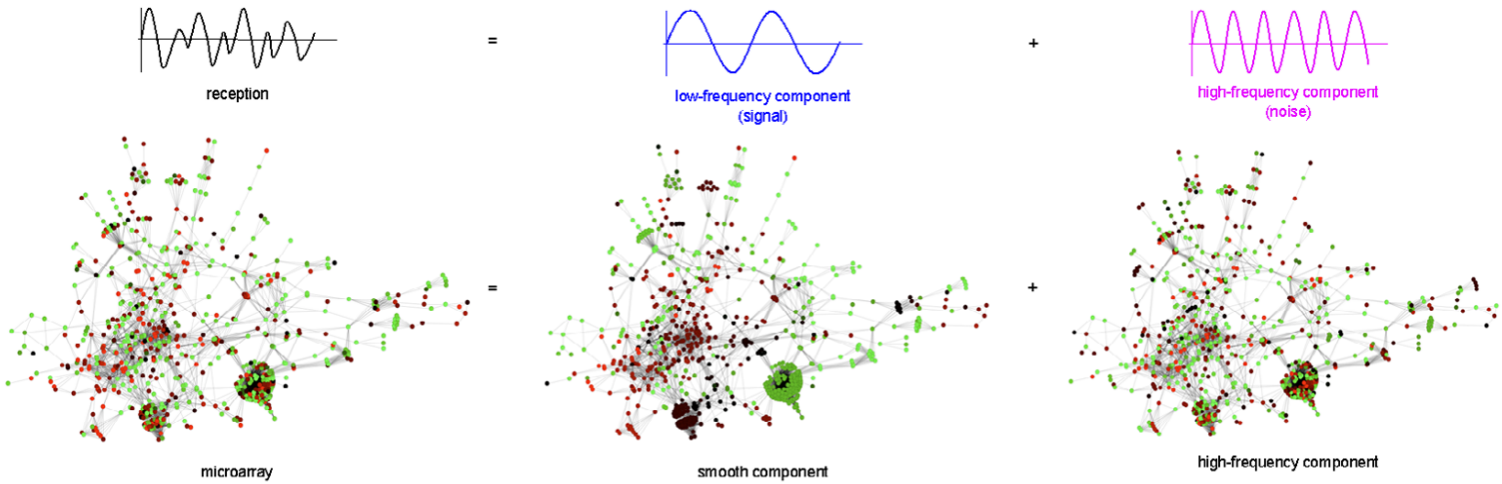


Extensiones

- Adaptaciones a predictores de distinta naturaleza.
- Descubrir asociaciones entre variables en grandes bases de datos.
- Desarrollar métodos especialmente adaptados para redes biológicas.

Extensiones

- Adaptaciones a predictores de distinta naturaleza.
- Descubrir asociaciones entre variables en grandes bases de datos.
- Desarrollar métodos especialmente adaptados para redes biológicas. Funciones biológicas básicas (metabólicas, regulatorias) se relacionan con *pathways*, no con genes aislados. Podemos usar conceptos de análisis armónico en grafos y semigrupos para capturar la estructura de los datos.

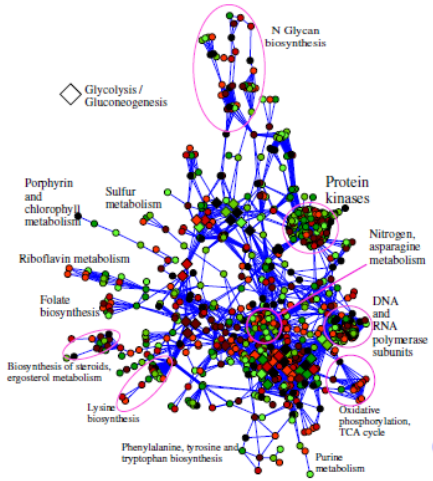


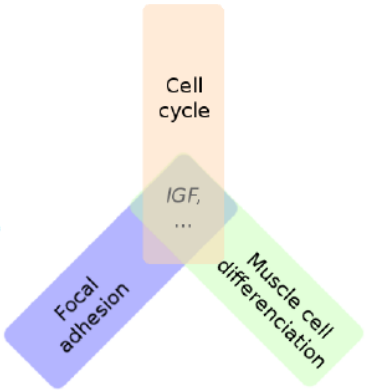
Extensiones

- Adaptaciones a predictores de distinta naturaleza.
- Descubrir asociaciones entre variables en grandes bases de datos.
- Desarrollar métodos especialmente adaptados para redes biológicas.
- Regularización estructurada para selección eficiente de variables.

Extensiones

- Adaptaciones a predictores de distinta naturaleza.
- Descubrir asociaciones entre variables en grandes bases de datos.
- Desarrollar métodos especialmente adaptados para redes biológicas.
- Regularización estructurada para selección eficiente de variables. *grupos* de variables que influyen de manera coordinada sobre la respuesta.





Extensiones

- Adaptaciones a predictores de distinta naturaleza.
- Descubrir asociaciones entre variables en grandes bases de datos.
- Desarrollar métodos especialmente adaptados para redes biológicas.
- Regularización estructurada para selección eficiente de variables.

Extensiones

- Adaptaciones a predictores de distinta naturaleza.
- Descubrir asociaciones entre variables en grandes bases de datos.
- Desarrollar métodos especialmente adaptados para redes biológicas.
- Regularización estructurada para selección eficiente de variables.

Extensiones

- Adaptaciones a predictores de distinta naturaleza.
- Descubrir asociaciones entre variables en grandes bases de datos.
- Desarrollar métodos especialmente adaptados para redes biológicas.
- Regularización estructurada para selección eficiente de variables.

  ...

Teoría

Usando modelos para la reducción inversa encontramos $R(\mathbf{X})$ que no pierda información que $\mathbf{X}$ tiene sobre Y

Teoría

Usando modelos para la reducción inversa encontramos $R(\mathbf{X})$ que no pierda información que $\mathbf{X}$ tiene sobre Y

Teoría

Usando modelos para la reducción inversa encontramos $R(\mathbf{X})$ que no pierda información que $\mathbf{X}$ tiene sobre Y

- La teoría se puede dividir en las siguientes contribuciones
  - Encontrar la RS teórica para las diferentes naturaleza de las variables

Teoría

Usando modelos para la reducción inversa encontramos $R(\mathbf{X})$ que no pierda información que $\mathbf{X}$ tiene sobre Y

- La teoría se puede dividir en las siguientes contribuciones
  - Encontrar la RS teórica para las diferentes naturaleza de las variables
  - Encontrar estimadores los más eficientes posibles

Teoría

Usando modelos para la reducción inversa encontramos $R(\mathbf{X})$ que no pierda información que $\mathbf{X}$ tiene sobre Y

- La teoría se puede dividir en las siguientes contribuciones
  - Encontrar la RS teórica para las diferentes naturaleza de las variables
  - Encontrar estimadores los más eficientes posibles
  - Dar un error a las estimaciones (importantísimo)

Teoría

Usando modelos para la reducción inversa encontramos $R(\mathbf{X})$ que no pierda información que $\mathbf{X}$ tiene sobre Y

- La teoría se puede dividir en las siguientes contribuciones
  - Encontrar la RS teórica para las diferentes naturaleza de las variables
  - Encontrar estimadores los más eficientes posibles
  - Dar un error a las estimaciones (importantísimo)
  - Probar que se puede substituir $\mathbf{X}$ por $\hat{R}(\mathbf{X})$ para predecir $Y|R(\mathbf{X})$ o para modelar $Y|R(\mathbf{X})$.

Teoría

Usando modelos para la reducción inversa encontramos $R(\mathbf{X})$ que no pierda información que $\mathbf{X}$ tiene sobre Y

- La teoría se puede dividir en las siguientes contribuciones
  - Encontrar la RS teórica para las diferentes naturaleza de las variables
  - Encontrar estimadores los más eficientes posibles
  - Dar un error a las estimaciones (importantísimo)
  - Probar que se puede substituir $\mathbf{X}$ por $\hat{R}(\mathbf{X})$ para predecir $Y|R(\mathbf{X})$ o para modelar $Y|R(\mathbf{X})$.
- Trabajo actual: $p > n$. Avances en PFC escencialmente y algunos métodos de regresión directa. En la práctica y teoría: aún cuando no se consiguen buenos estimadores de la reducción suficiente, en predicción (cuando realmente incorporo información al hacer crecer p) tengo buenos resultados

GRACIAS!

FIQ UNL

UNIVERSIDAD NACIONAL DEL LITORAL
FACULTAD DE INGENIERÍA QUÍMICA

FIQUNL
www.fiq.unl.edu.ar