Métodos de calibración multivariada.

Cristóbal Honores

Resumen clase anterior

1. Test de hipótesis.
2. Intervalos de confianza.
3. Comparación de dos muestras.
4. Comparación de varias muestras.

Test de hipótesis

Se plantean dos hipótesis:

$$H_0 \leftarrow \text{nula}$$
$$H_1 \leftarrow \text{hipótesis alternativa}$$

- **t.test():** Test para media con varianza desconocida.
- **z.test():** Test para media con varianza conocida.
- **sigma.test():** Test para varianzas.
- **prop.test():** Test para proporciones.

Intervalos de confianza

Del test de hipótesis que se realizo previamente:

- Si se selecciona la variable conf.int entrega el valor del intervalo de confianza con el nivel de confianza deseado.

```
as.numeric(TEST$conf.int)
```

Ojo!!! El IC solo se obtiene cuando el test de hipótesis se realiza “two.sided”

Comparación de dos muestras

1. Realizar test de varianzas:

   var.test()

2. Test para medias:
   1. Varianzas igueles:

      t.test(....., var.equal=TRUE)

   2. Varianzas distintas:

      t.test(...., var.equal=FALSE)

   3. Medias pareadas:

      t.test(....,paired=TRUE)

3. Test para proporciones:

   prop.test(x=c(...,...),n=c(....,...),....)

Test de normalidad

Con el fin de comprobar la normalidad de los datos se realizan los siguientes test:

1. Q-q plot

   qqnorm(datos) ; qqline(datos)

2. Kolmogorov-Smirnov

   ks.test(datos)

3. Shapiro-Wilk

   shapiro.test(datos)

OJO! Shapiro-Wilk solo puede realizarse con tamaños de muestra menores a 5000.

Comparación varias muestras

Esta prueba estadística es de las más utilizadas para poder comparar más de dos muestras poblacionales

$$H_0: \mu_A = \mu_B = \mu_c = \mu_D = \cdots$$
$$H_1: A\ lo\ menos\ una\ distinta$$

- ANOVA:

aov(valores ~ categoría)

- Kruskal-Wallis

kruskal.test(valores ~ categoría, data = ...)

Propósito de la clase

Conocer diferentes modelos de calibración multivariada y lograr hacer una comparación entre los modelos obtenidos.

Métodos de calibración multivariada

1. Calibración univariada.
2. Método de Stepwise.
3. Validación cruzada.
4. PCR.
5. PLS.
6. Comparación de modelos.

Calibración univariada

- Tiene como objetivo comparar los valores obtenidos por un instrumento de medición con la medida correspondiente de un patrón de referencia (o estándar).

- Es el tipo de calibración mas sencilla y mas utilizada.

- Regresión lineal simple.

Regresión lineal simple

- Modelo matemático:

$$Y = \beta_0 + \beta_1 X$$

Donde:

Y es mi variable respuesta (medición).

X es mi variable independiente o regresor.

β_0 es el intercepto de la ecuación.

β_1 es la pendiente de la ecuación (coeficiente de sensibilidad).

Forma matricial

$$Y = X\beta + \varepsilon$$

Donde:

$$\beta = (X^T X)^{-1} X^T Y$$

Concepto bajo el cual trabaja

- Utiliza una variable independiente para modelar una variable respuesta.

- Ejemplo: Una concentración que explica la absorbancia en una longitud de onda.

- Luego de modelar mi variable respuesta, esta ecuación matemática encontrada se utiliza para predecir otros valores de X.

De manera inversa ....

- También es correcto la manera inversa, predecir una concentración atreves de una medición.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Es (matemáticamente) equivalente a:

$$X = \beta_0 + \beta_1 Y + \varepsilon$$

Ejemplo en R

- Se utiliza atreves del comando lm().

```
reg1 <- lm(base$variable_x ~ base$variable_y)
```

- Para obtener los coeficientes, valor de los residuos, etc, se usa el comando summary().

```
summary(reg1)
```

Regresión lineal múltiple

- Modelo matemático:

$$Y = \beta_0 + \beta_1 X_2 + \beta_1 X_2 + \cdots$$

Donde:

Y es mi variable respuesta ().

X son las variables independientes o regresores (mediciones).

β_0 es el intercepto de la ecuación.

β_1 son los coeficientes de sensibilidad para cada medición.

Concepto bajo el cual trabaja

- Utiliza varias mediciones para modelar y predecir una concentración.

- No es muy utilizado ya que cada medición trae en si un error asociado la incertidumbre y precisión del instrumento con el cual se midió.

- Aumentando el error de ajuste del modelo.

¿Para que es útil entonces?

- Es muy útil para instrumentos capaces de realizar mas de una observación por cada experimento.

- **Cromatografia**
- **Espectroscopia (NIR o UV-Vis)**

- Esto permite obtener el espectro de un compuesto en todo el rango de longitudes de onda.

Ventajas y desventajas

- Al utilizar todo el rango de longitudes de onda, es posible obtener mas información y realizar un mejor ajuste y tener un modelo mas robusto.

- El problema es, seleccionar las longitudes de a las cuales se utilizaran para el modelo. (seleccionar los regresores)

- Si bien un regresor adecuado me ayuda a predecir mi Y, un mal regresor solo aporta mayor incertidumbre al modelo.

Método de Stepwise

- Selecciona cuales son los mejores regresores y la cantidad de estos para obtener el mejor modelo posible.

- Trabaja atreves de criterios de selección de variables.

- Un buen regresor mejora el valor de este criterio, un mal regresor, empeora el valor de este.

Criterio de selección de variables

- Existen diferentes tipos de criterios matemáticos que ven el ajuste de un modelo de regresión.

1. Coeficiente de determinación R^2.

2. Coeficiente de determinación corregido $R^2 - ajust$.

3. Criterio de información de Akaike “AIC”.

4. Criterio de información Bayesiana “BIC”.

Ejemplo en R

- Se utiliza atreves del comando step() atreves de un modelo de regresión múltiple.

step(lm(variable_deseada ~ ., data=...)**)**

También se puede guardar la regresión como un objeto.

reg2 <- lm(variable_deseada ~ ., data=...)

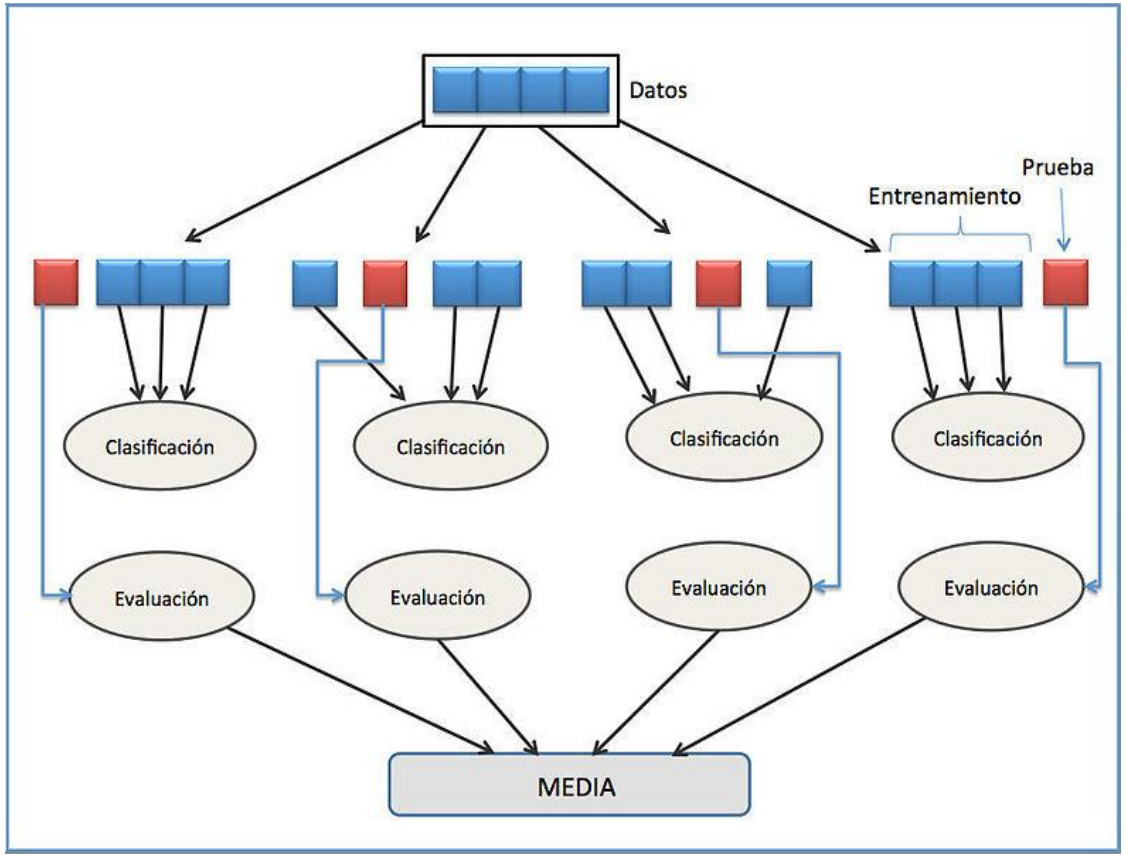
step(reg2**)**

Validación cruzada (CV)

- Una vez seleccionados los predictores adecuados, generado el modelo y comprobado que se cumplen las condiciones necesarias del método de ajuste empleado, el siguiente paso es evaluar la capacidad de dicho modelo para predecir la variable respuesta.

¿En que consiste?

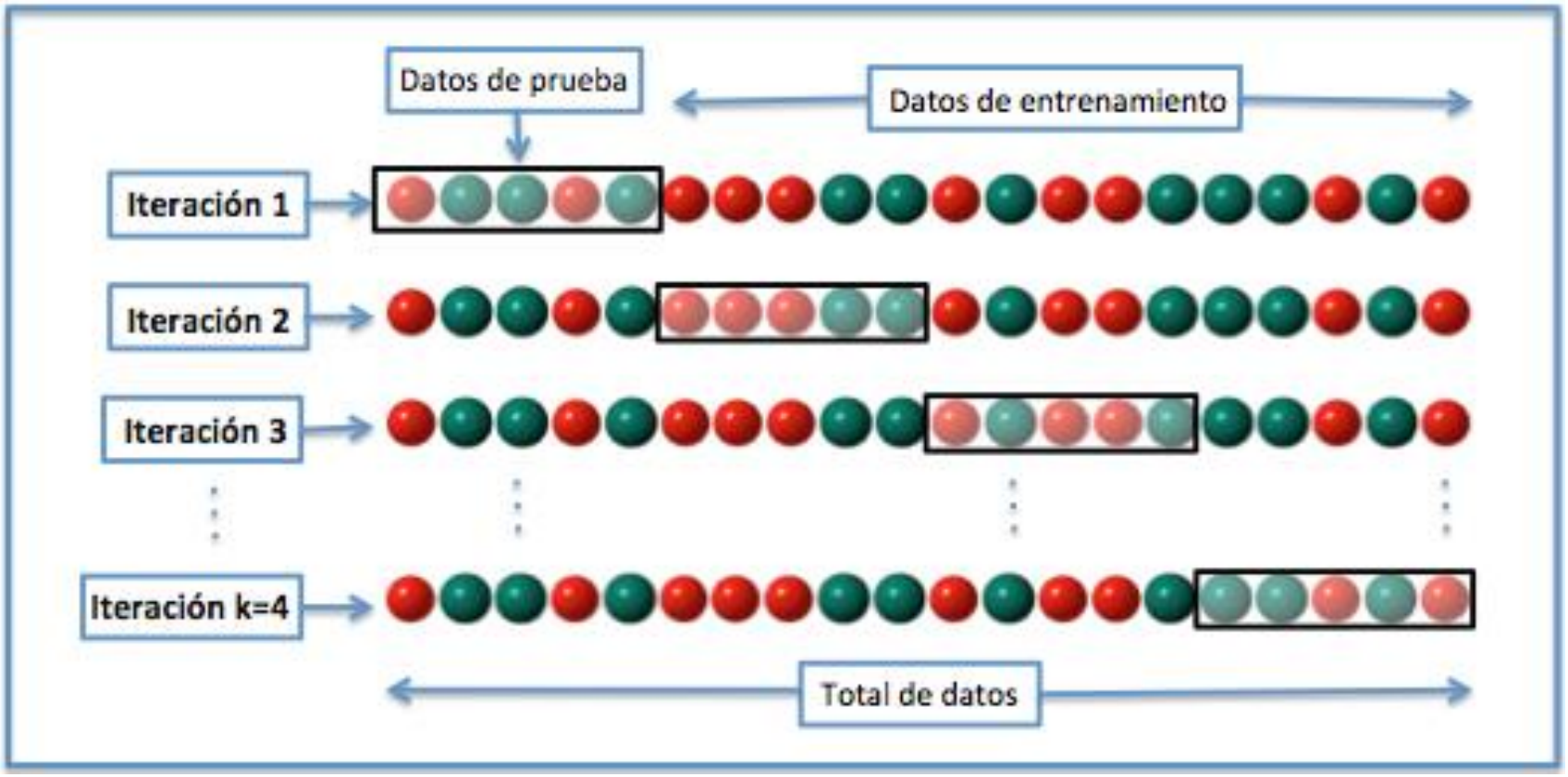
- Consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones.



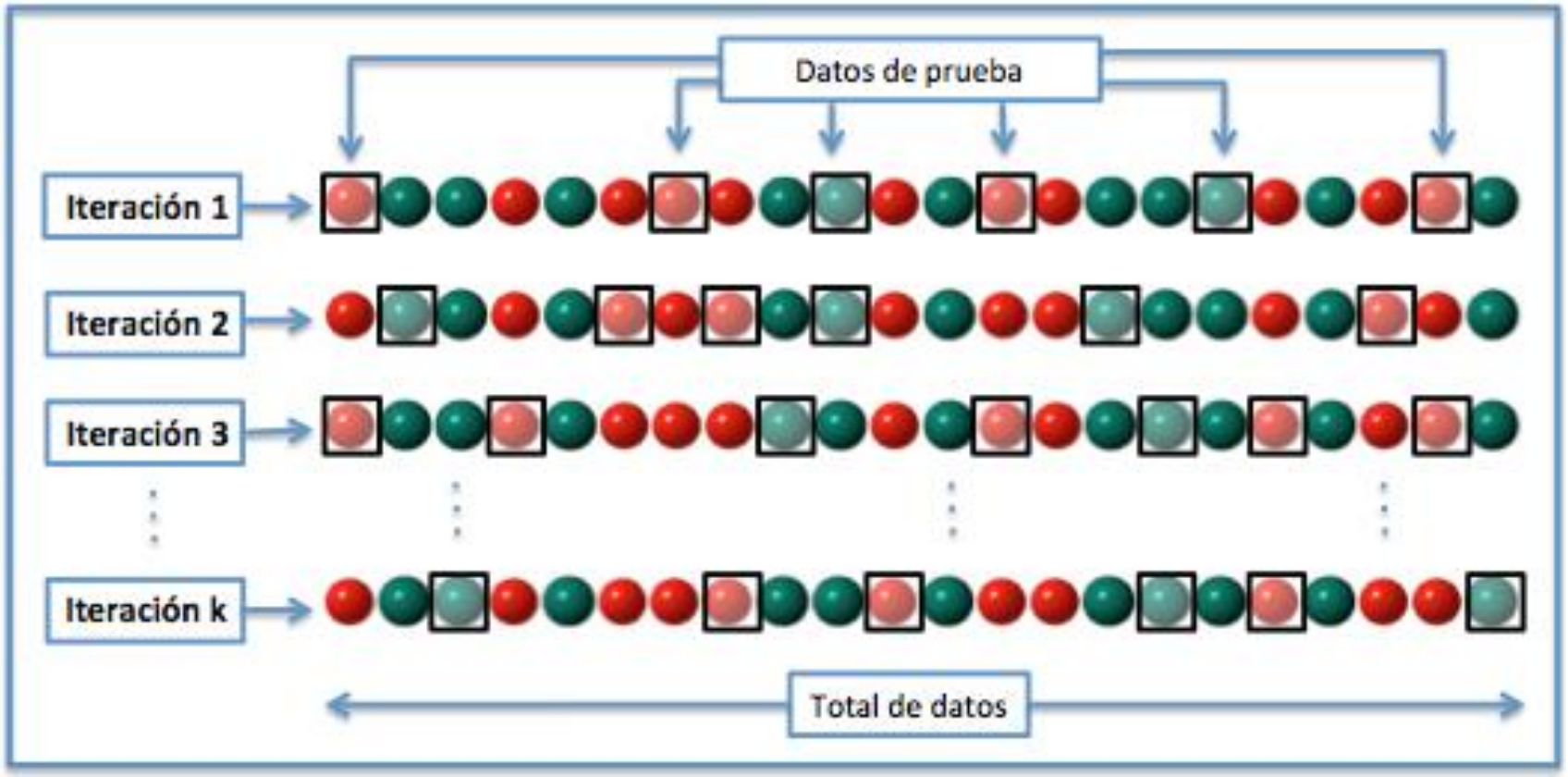
Tipos de cross-validataion

1. Validación cruzada de K iteraciones.
2. Validación cruzada aleatoria.
3. Validación cruzada dejando uno fuera.

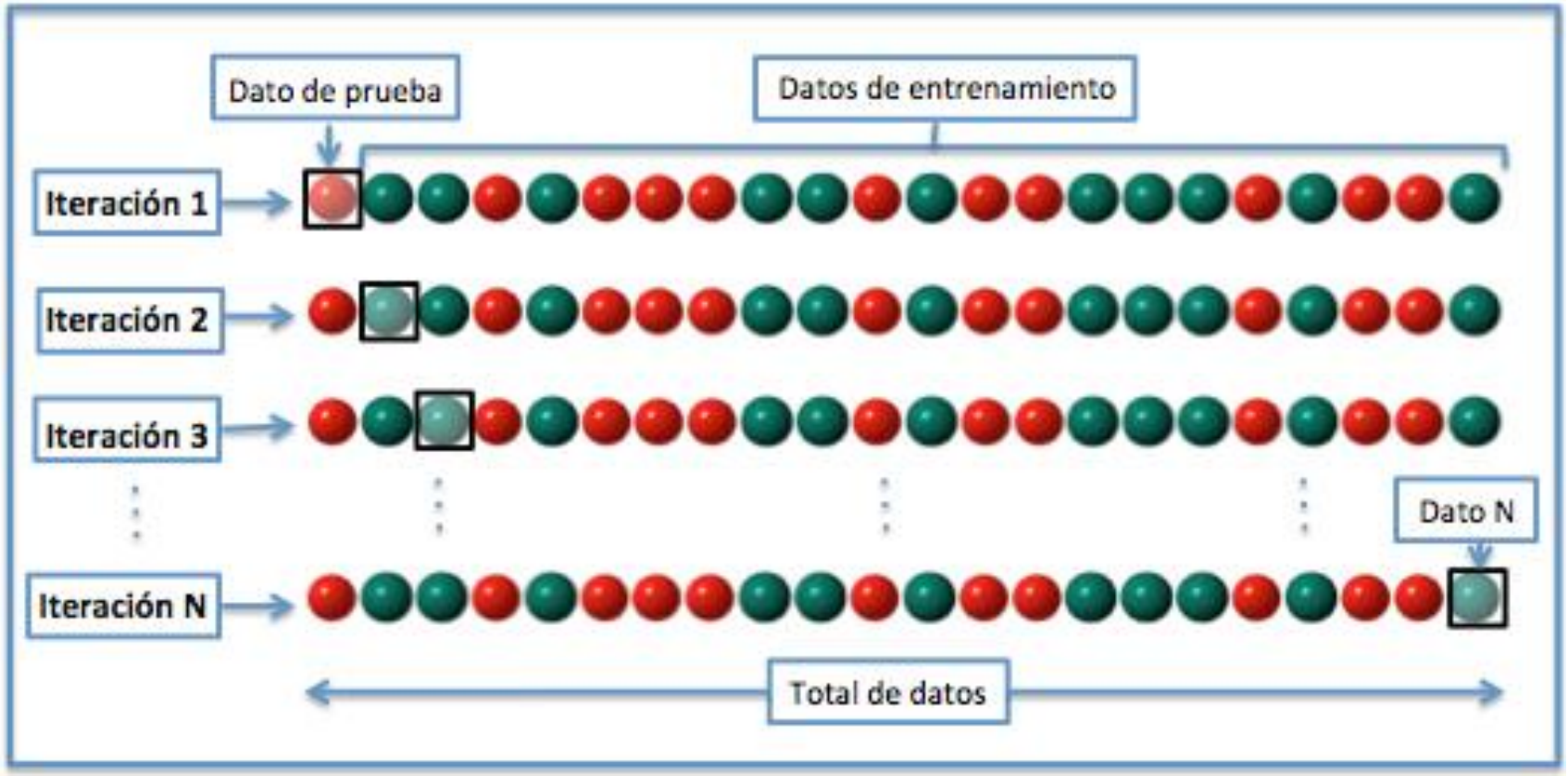
Validación cruzada de K iteraciones



Validación cruzada aleatoria



Validación cruzada dejando uno fuera



Regresión sobre componentes principales (PCR)

La regresión de componentes principales (PCR) aplica el Análisis de componentes principales en el conjunto de datos para resumir las variables predictoras originales en pocas variables nuevas también conocidas como componentes principales (PC), que son una combinación lineal de los datos originales.

Forma matricial de un PC

Regresión lineal simple:

$$Y = X\beta + \varepsilon$$

Se reemplaza X por:

$$X = TP$$

¿Qué se logra con esto?

- Se cambia la matriz X por las matrices TP.
  - T corresponde a los scores ortogonales.
  - P corresponde a los loadings.

- Al reemplazar X, logramos disminuir la dimensionalidad de la matriz ("menor cantidad de variables").

Diferencia entre PCA y PCR

- La gran diferencia matemática entre estos métodos es en la formación de las matrices T y P.

- La diferencia conceptual entre estos:

  - PCA: Maximiza la varianza retenida.

  - PCR: Minimiza el error cuadrático medio en la predicción (MSEP).

Aplicación en R

1. Paquete: pls

2. Se utiliza el comando pcr().

pcr(X ~ Y, ncomp=... , data... , validation = ...**)**

Selección del numero de componentes

- Seleccionar el numero de componentes:

Para esto se realiza un grafico del error vs numero de componentes.

plot(RMSEP(pcr), legendpos = ...**)**

Regresión de mínimos cuadrados parciales. (PLSR)

La regresión sobre mínimos cuadrados parciales identifica nuevos componentes principales que no solo resumen los predictores originales, sino que también están relacionados con el resultado.

En comparación con la PCR, PLS utiliza una estrategia de reducción de dimensiones supervisada por el resultado.

Diferencia entre PCR y PLSR

- Ambos modelos consisten en descomponer la matriz X en dos matrices T y P.

- PCR toma información de solo de X para formar los scores y los loadings.
  - Se enfoca en la varianza de X.

- PLSR toma en cuenta información de la matriz X e Y.
  - Puede explicar la covarianza entre X e Y.

Aplicación en R

1. Paquete: pls

2. Se utiliza el comando plsr().

pslr(X ~ Y, ncomp=... , data... , validation = ...**)**

Selección del numero de componentes

- Seleccionar el numero de componentes:

Para esto se realiza un grafico del error vs numero de componentes.

plot(RMSEP(pcr), legendpos = ...**)**

Predicción con modelos de regresión

- Para poder predecir los valores deseados de un muestra desconocida es necesario seguir los siguientes pasos:

1. Elegir el modelo de predicción deseado.

2. Guardar los resultados de las mediciones como una nueva data.

3. Utilizar la función predict() en R.

Predicción en R

Para predecir utilizando un modelo ya generado, se utiliza el comando predict().

```
predict(object = modelo, newdata = datos_a_predecir)
```

Comparación de modelos

Para comparar dos modelos se calcula atreves del error cuadrático medio (MSE).

$$MSE = (Datos\ de\ test - Predicciones)^2$$

En R:

```
mean((BaseTest$Variable - predicción)^2)
```

Métodos de calibración multivariada.

Cristóbal Honores