Sesión 2: Modelo relacional en Power Bi + Operador Sweep y Descomposición matricial

Aplicaciones en Computación Estadística

Natalie Julian - www.nataliejulian.com

Estadística UC y Data Scientist en Zippedi Inc.

Ejercicio 1

Una empresa posee alrededor de 4000 empleados. Sin embargo, cada año, aproximadamente el 15% de sus empleados abandonan la empresa, número no menor que dificulta muchas veces la entrega de ciertos proyectos en el tiempo dado. Una teoría es que el porcentaje incremental del sueldo de un año a otro no es muy atractivo para permanecer en la empresa, por lo que, recursos humanos le pide a usted que realice análisis de este aumento de sueldo porcentual y que determine cuáles son las variables que se relacionan con un mayor aumento de sueldo de un año a otro.

Ejercicio 1

La información se encuentra en las bases de datos `employee_survey`, `manager_survey` y `general`. La información de cada variable se encuentra en `data_dictionary`.

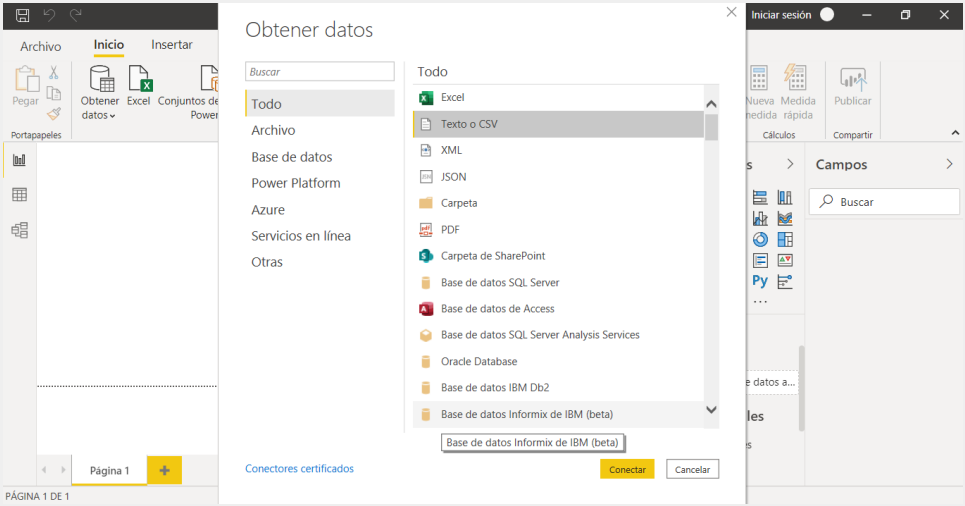
Ejercicio 1

a) Realice la importación de las bases de datos `employee_survey`, `manager_survey` y `general` y obtenga el cruce correspondiente para unificar la información.

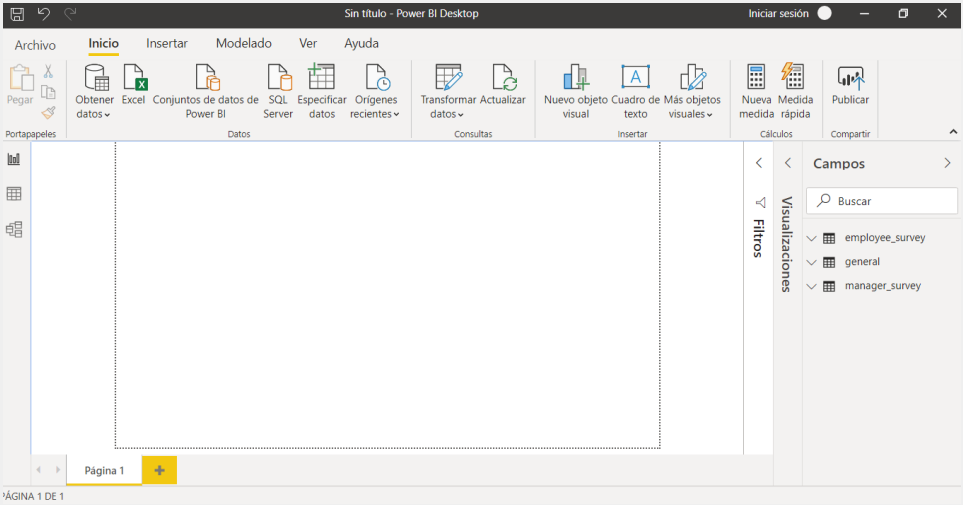
Modelo relacional

Las bases de datos se encuentran relacionadas y es necesario determinar cuál(es) es(son) la(s) llave(s) entre cada par de bases de datos.

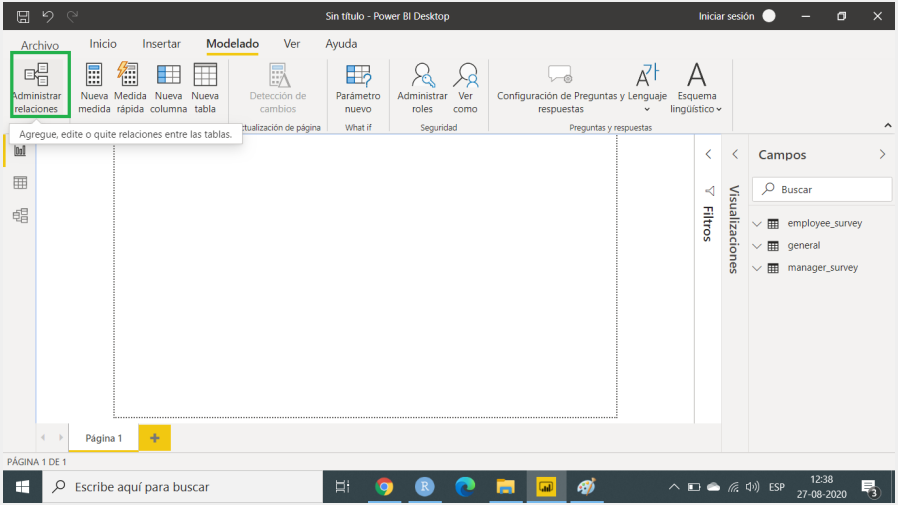
Modelo relacional en PowerBI



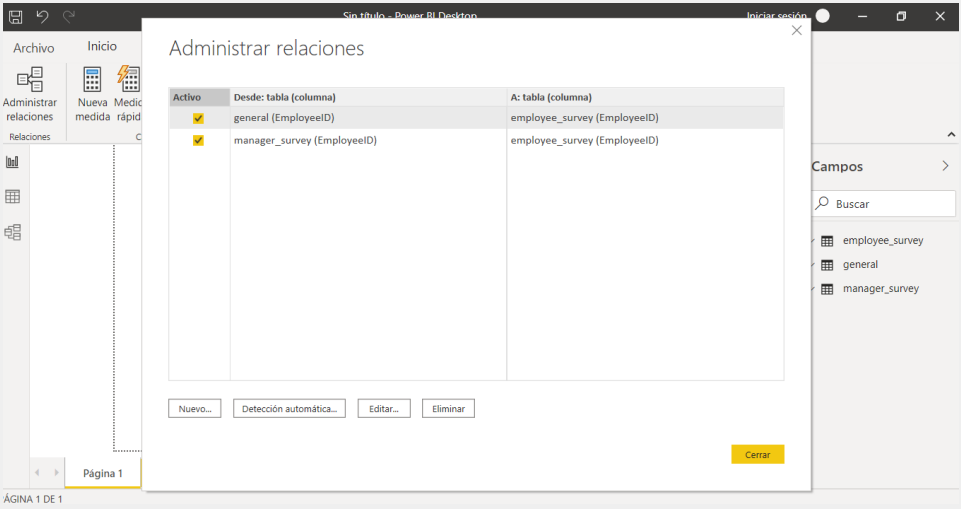
Modelo relacional en PowerBI



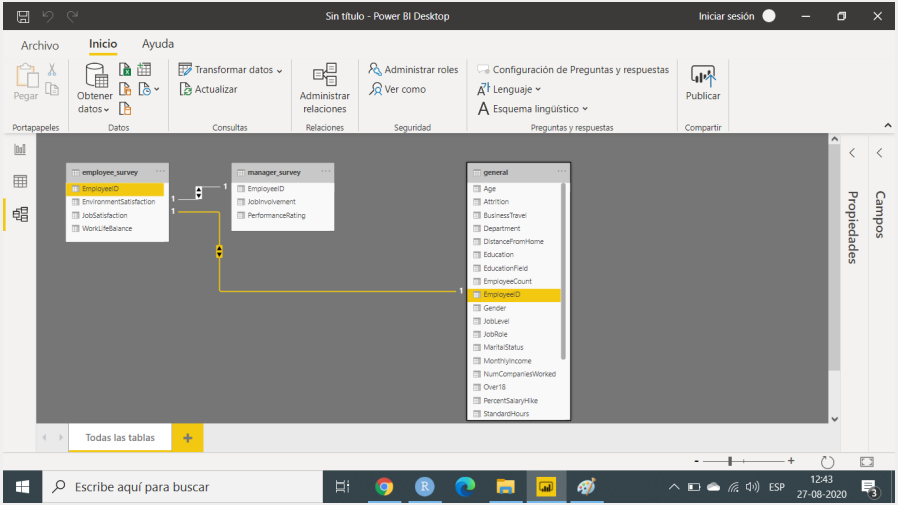
Modelo relacional en PowerBI



Modelo relacional en PowerBI



Modelo relacional en PowerBI



Llave primaria

Podemos observar que `EmployeeID` es la llave primaria, por lo tanto al realizar el cruce entre las tablas debemos utilizarla como conexión entre ellas.

Una llave primaria es la variable que logra conectar la información de todo el modelo, usualmente se denomina ID.

Subir múltiples bases de datos simultáneamente en R

En este caso tenemos tres bases de datos y todas en formato csv, es posible subirlas simultáneamente con la siguiente sintaxis:

```
#Defino el directorio a trabajar:
setwd("C:/Users/HP/Desktop/Trabajo/2020-2/Introducción a la Computación/Ayudantía 1")


#Obtiene un vector con el nombre de los archivos csv en el directorio:
nombres<- list.files(path = getwd(),
                     pattern = "\\.csv$",
                     full.names = FALSE)

library(readr)

#Carga todos los archivos csv en una lista:
Datas<-lapply(nombres,"read_csv")

#Añade nombre respectivo a cada data:
names(Datas)<-substr(nombres,1,nchar(nombres)-4)

class(Datas)
[1] "list"
```

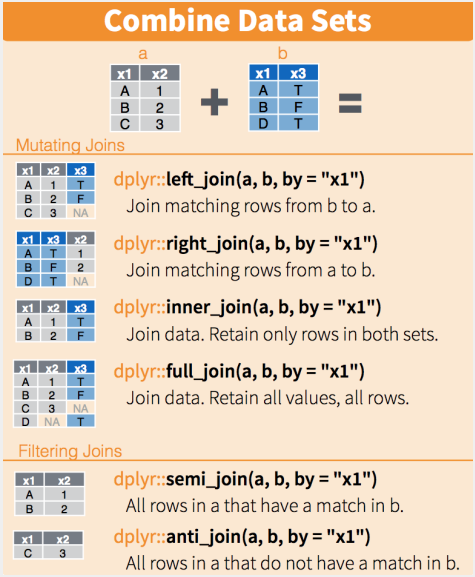
Aspectos importantes al realizar un cruce

Siempre al realizar un cruce es necesario determinar las dimensiones de cada base de datos. Muchas veces existen bases de datos más específicas que otras y por lo tanto, más pequeñas, por lo cual, en esos casos, empezar el cruce por una u otra base de datos entregaría resultados diferentes.

```
dim(Datas$employee_survey)
[1] 4410    4

dim(Datas$general)
[1] 4410   24

dim(Datas$manager_survey)
[1] 4410    3

intersect(names(Datas$general), names(Datas$employee_survey))
[1] "EmployeeID"

intersect(names(Datas$general), names(Datas$manager_survey))
[1] "EmployeeID"

intersect(names(Datas$employee_survey), names(Datas$manager_survey))
[1] "EmployeeID"
```

Como la única llave es `EmployeeID`, la base de datos resultante tendrá $24 + (4 - 1) + (3 - 1) = 29$ columnas.

Cruce de las tablas con dplyr



Full join

```
cruce<-full_join(Datas$general,Datas$employee_survey,by="EmployeeID") %>%
   full_join(., Datas$manager_survey, by="EmployeeID")

dim(cruce)
[1] 4410   29

names(cruce)
 [1] "Age"                       "Attrition"                 "BusinessTravel"
 [4] "Department"                "DistanceFromHome"          "Education"
 [7] "EducationField"            "EmployeeCount"             "EmployeeID"
[10] "Gender"                    "JobLevel"                  "JobRole"
[13] "MaritalStatus"             "MonthlyIncome"             "NumCompaniesWorked"
[16] "Over18"                    "PercentSalaryHike"         "StandardHours"
[19] "StockOptionLevel"          "TotalWorkingYears"         "TrainingTimesLastYear"
[22] "YearsAtCompany"            "YearsSinceLastPromotion"   "YearsWithCurrManager"
[25] "EnvironmentSatisfaction"   "JobSatisfaction"           "WorkLifeBalance"
[28] "JobInvolvement"            "PerformanceRating"
```

La variable de interés modelar es la diferencia porcentual en el salario respecto al año pasado. Esta información se encuentra en la variable `PercentSalaryHike`.

Ejercicio 1

b) Considere un modelo de regresión lineal con el aumento porcentual del salario respecto al año pasado de los empleados como variable respuesta y considere las siguientes variables predictoras:

1. `PerformanceRating:` Indicador del desempeño del empleado (de 1 a 4, de peor a mejor)
2. `YearsAtCompany:` Años del empleado en la compañía
3. `Age:` Edad del empleado
4. `Gender:` Sexo del empleado

Obtenga $\hat{\beta}$ utilizando la descomposición de Cholesky, QR, LU y descomposición espectral. Compare las estimaciones y también el costo computacional de cada descomposición.

Regresión Lineal

En una regresión lineal, el problema se reduce a resolver el siguiente sistema:

$$\left(X^\top X\right)\beta = X^\top Y$$

Equivalente a resolver un sistema $Ax = b$. Resulta de utilidad, descomponer $\left(X^\top X\right)$ de manera que obtener su inversa (cálculo costoso) sea menos doloroso.

SWEEP

c) Utilice el operador SWEEP y obtenga los estadísticos y valores-p asociados al test t de significancia. Comente.

Operador SWEEP

Definición Sea A una matriz cuadrada. El algoritmo para sweppear A en el k-ésimo elemento de su diagonal (asumiendo que ningun término de su diagonal es nulo) se denota por Z=Sweep(A,k) y se describe como sigue:

$$Z_{kk} = -\frac{1}{A_{kk}}$$

$$Z_{ik} = \frac{A_{ik}}{A_{kk}} \qquad i \neq k, \qquad i - esima \quad columna$$

$$Z_{kj} = \frac{A_{kj}}{A_{kk}} \qquad j \neq k \qquad j - esima \quad fila$$

$$Z_{ij} = A_{ij} - \frac{A_{ik}A_{kj}}{A_{kk}} \qquad i \neq k, j \neq k$$

Operador SWEEP

Después de sweepear la matriz en todos los elementos de su diagonal obtenemos:

$$\begin{pmatrix} -(X^TX)^{-1} & (X^TX)^{-1}X^Ty \\ y^TX(X^TX)^{-1} & y^Ty - y^TX(X^TX)^{-1}X^Ty \end{pmatrix} = \begin{pmatrix} \frac{-1}{\sigma^2}Var(\hat{\beta}) & \hat{\beta} \\ \hat{\beta}^T & ||y - \hat{y}|| \end{pmatrix}$$

Recordatorios

$$SCT = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$SCM = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

$$SCE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$R^2 = 1 - \frac{SCE}{SCT}$$

$$CME = \frac{SCE}{n - p}$$

$$CMM = \frac{SCM}{p - 1}$$

El estadístico del test T para cada coeficiente es:

$$T_j = \frac{\beta_j}{\sigma / S_{jj}} \quad \text{con} \quad S_{jj} = \sqrt{(X^T X)_{jj}}$$

Práctico

Ejercicio práctico 1

Considere la base de datos `Spotify.RData` la cual contiene información sobre canciones en la plataforma Spotify. A continuación se presenta la información:

- `acousticness`: Medida de la acústica en la pista, valores cercanos a 1 indican que gran parte de la canción es acústica.
- `danceability`: Medida de bailabilidad de la pista, valores cercanos a 1 indican que es muy bailable.
- `energy`: Medida que indica qué tanta energía se desprende de una canción, valores cercanos a 1 indican que induce mucha energía.
- `happyness`: Corresponde a una medida de la alegría transmitida en la canción, valores cercanos a 1 indican que la pista sería positiva, alegre o eufórica.
- `key`: El tono en el que está la canción usando notación entera.
- `liveness`: Detecta la presencia de audiencia en la pista, valores cercanos a 1 indican que la pista estaría grabada en vivo.
- `loudness`: Corresponde al volumen promedio de la pista en decibeles.
- `speechiness`: Detecta la presencia de monólogos en una pista (hablado), valores cercanos a 1 indicarían que prácticamente toda la pista es hablada.
- `instrumentalness`: Se refiere a si la pista carece de letra, valores cercanos a uno indican que la pista casi no tiene letra.

Ejercicio práctico 1

Para cargar las variables utilice:

```
load(file.choose())
```

Ejercicio práctico 1

a) Cree la matriz de covariables. Utilice alguna medida para cuantificar la multicolinealidad de la base de datos. Comente.

Ejercicio práctico 1

b) Obtenga la descomposición espectral y la descomposición de Cholesky para encontrar $\hat{\beta}$ con mínimos cuadrados (considere intercepto). ¿Qué ocurre con $\hat{\beta}$ al realizar ambas descomposiciones?, ¿son iguales las estimaciones? ¿en qué casos pudiera ocurrir que las estimaciones no coincidan? Comente. Compare los tiempos computacionales de ambos procesos con , ¿qué observa? ¿tiene sentido lo obtenido?

Ejercicio práctico 1

c) Cree una función que utilice SWEEP y que entregue los coeficientes estimados de una regresión lineal, el valor estimado de R^2, errores estándar, estadísticos y valores p del test t para cada coeficiente. (*Ejercicio propuesto*)

d) Compare con la función `lm()`. ¿Cuáles son las variables más significativas? ¿cómo es el ajuste del modelo? Comente.(*Ejercicio propuesto*)