Laboratorio 6 - Asociación, tablas, loops y limpieza de datos

Métodos Cuantitativos

Carlos Eduardo Molina Berumen

El Colegio de México

14 de noviembre de 2025

Repaso

Ya sabemos como:

- Hacer una regresión lineal
- Gráficar una regresión lineal
- Buscar la covarianza de dos variables.

En este lab aprenderemos:

- Más medidas de asociación
- Cómo hacer tablas de contingencia
- Cómo hacer loops
- Cómo limpiar bases de datos

Explicación conceptual

Covanrianza: cómo dos variables numéricas varían juntas. No estandarizada.

Coeficiente de determinación R^2: qué proporción de la variación de la VD es explicada por la VI. Nos la da la regresión lineal.

Coeficiente de correlación de Pearson: es una versión estandarizada de la covarianza que va de -1 a 1.

Chi cuadrada χ^2: se usa para ver si dos variables categóricas están relacionadas o son independientes.

Coeficiente de Pearson:

```
library(palmerpenguins)
library(stargazer)
penguins <- na.omit(penguins)
```

```r
cor.test(penguins$bill_length_mm, penguins$body_mass_g,
         method = "pearson")
```

```
    Pearson's product-moment correlation

data:  penguins$bill_length_mm and penguins$body_mass_g
t = 13.276, df = 331, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5145745 0.6554058
sample estimates:
      cor
0.5894511
```

Hay una relación positiva moderada-fuerte y estadísticamente significativa.

También se puede usar cor(), peo cor.test() les da una relación más completa.

Chi cuadrada χ^2

Recordatorio: La chi cuadrada χ^2 mide la relación entre dos variables categóricas. No numéricas.

Primer paso, hacer una tabla de contingencia.

Tabla de contingencia

Tabla en forma de matriz que organiza dos o más variables categóricas. Ella mide la **frecuencia** de la combinación de los cruces entre cada categoría.

```
tabla <- table(penguins$species, penguins$island)
tabla
```

```
           Biscoe Dream Torgersen
  Adelie       44    55        47
  Chinstrap     0    68         0
```

```r
objeto <- chisq.test(tabla)

stargazer(objeto, type = "text")
```

```
% Error: Unrecognized object type.
```

Hay relación estadísticamente significativa entre las variables.

Parte dos: loops

Los loops son una herramienta muy útil para evitar tareas repetitivas en r. En lugar de pedirle a r que imprima el cuadrado de cada entero del uno al cinco por separado, lo podemos hacer en un solo comando.

```
for (i in 1:5) {
  print(i**2)
}
```

```
[1] 1
[1] 4
[1] 9
[1] 16
[1] 25
```

Hagámos un loop más interesante:

```
library(fivethirtyeight)
data <- airline_safety

for (i in 1:nrow(data)) {
  aerolinea <- data$airline[i]
  fatalidades <- data$fatalities_00_14[i]

  if (fatalidades > 0) {
    cat("La aerolínea", aerolinea, "tuvo", fatalidades,
        "fatalidades entre el 2000 y el 2014.\n")
  }
}
```

```
La aerolínea Aeroflot tuvo 88 fatalidades entre el 2000 y el 2
La aerolínea Air France tuvo 337 fatalidades entre el 2000 y e
La aerolínea Air India tuvo 158 fatalidades entre el 2000 y el
La aerolínea Air New Zealand tuvo 7 fatalidades entre el 2000
```

Parte 3: limpiar datos.

```
library(ggplot2)
economics <- economics
```

Seleccionar columnas

```
library(dplyr)
unemployment <- select(economics, date, unemploy, psavert)
head(unemployment, n= 2)
```

```
# A tibble: 2 x 3
  date       unemploy psavert
  <date>        <dbl>   <dbl>
1 1967-07-01     2944    12.6
2 1967-08-01     2945    12.6
```

```r
df1 <- select(economics, date:pop)
head(df1, n = 2)
```

```
# A tibble: 2 x 3
  date         pce    pop
  <date>     <dbl>  <dbl>
1 1967-07-01  507. 198712
2 1967-08-01  510. 198911
```

```r
df2 <- select(economics, starts_with("p"))
head(df2, n = 2)
```

```
# A tibble: 2 x 3
    pce    pop psavert
  <dbl>  <dbl>   <dbl>
1  507. 198712    12.6
2  510. 198911    12.6
```

Quitar columnas

```
df3 <- select(economics, -c(pop, uempmed))
head(df3)
```

```
# A tibble: 6 x 4
  date         pce psavert unemploy
  <date>     <dbl>   <dbl>    <dbl>
1 1967-07-01  507.    12.6     2944
2 1967-08-01  510.    12.6     2945
3 1967-09-01  516.    11.9     2958
4 1967-10-01  512.    12.9     3143
5 1967-11-01  517.    12.8     3066
6 1967-12-01  525.    11.8     3018
```

Pivot longer (table ancha a tabla larga)

Sirve para crear una nueva variable “indicator” que contenga los datos de personal consumption expenditures, personal servings rate y number of unemployed in thousands.

```
#install.packages("tidyr")
library(tidyr)

df4<- pivot_longer(
  economics,
  cols = c(pce, psavert, unemploy),
  names_to = "indicator", values_to = "value")
```

```r
head(df4)
```

```
# A tibble: 6 x 5
  date          pop uempmed indicator  value
  <date>      <dbl>   <dbl> <chr>      <dbl>
1 1967-07-01 198712     4.5 pce         507.
2 1967-07-01 198712     4.5 psavert      12.6
3 1967-07-01 198712     4.5 unemploy  2944
4 1967-08-01 198911     4.7 pce         510.
5 1967-08-01 198911     4.7 psavert      12.6
6 1967-08-01 198911     4.7 unemploy  2945
```

Pivot wider (tabla larga a tabla ancha)

Proceso inverso a pivot longer. Sirve para que ahora nuestra columna "indicator" y "values" se conviertan en columnas "pce", "psavert", "unemploy".

```
df5 <- pivot_wider(
  df4,
  names_from = indicator,
  values_from = value
)
```

```r
head(df5)
```

```
# A tibble: 6 x 6
  date          pop uempmed   pce psavert unemploy
  <date>      <dbl>   <dbl> <dbl>   <dbl>    <dbl>
1 1967-07-01 198712     4.5  507.    12.6     2944
2 1967-08-01 198911     4.7  510.    12.6     2945
3 1967-09-01 199113     4.6  516.    11.9     2958
4 1967-10-01 199311     4.9  512.    12.9     3143
5 1967-11-01 199498     4.7  517.    12.8     3066
6 1967-12-01 199657     4.8  525.    11.8     3018
```

En suma

- **Extraer columnas:** select(col1, col2)
- **Quitar columnas:** select(-col)
- **De ancho a largo:** pivot_longer()
- **De largo a ancho:** pivot_wider()

Muchas gracias a todas!!! :)))

eduardo.molina@alumnos.cide.edu

carlos.e.molinaberumen@outlook.com