Pontificia Universidad JAVERIANA Bogotá

CORRELACIÓN Y ASOCIACIÓN DE VARIABLES

ING. HERIBERTO FELIZZOLA JIMENEZ

RUTA DE CIENCIA DE DATOS

Educación Continua

Generamos experiencias educativas

CONTENIDO

Pontificia Universidad JAVERIANA Bogotá

1. Motivación
2. Tipos de análisis
3. Gráficos de dispersión
4. Coeficiente de correlación
5. Gráficos comparativos y estadísticas por grupo
6. ANOVA

MOTIVACIÓN

Pontificia Universidad JAVERIANA Bogotá

- En el caso de análisis de los cargos médicos varios interrogantes pueden surgir al tratar de identificar cuales son factores que explican el comportamiento de los cargos.
- Ejemplo:
  - ¿Los pacientes con mayor edad tienen mayores cargos?
  - ¿Los pacientes con mayor IMC tienen mayores cargos?
  - ¿Hay diferencias en los cargos entre hombres y mujeres?
  - ¿La condición de fumador genera mayores cargos?
  - ¿Qué relación hay entre la cantidad de hijos y los cargos?
  - ¿Los cargos varían por región?
  - ¿Hay una relación entre la edad y el IMC?
- Todas estas preguntas conducen a evaluar la correlación o asociación entre variables.
- Cabe resaltar que la correlación o asociación no implica causalidad.

TIPOS DE ANÁLISIS

- Para analizar la asociación o correlación entre dos variables X, Y podemos utilizar diferentes herramientas gráficas y estadísticas.
- El tipo de herramienta depende del tipo de variable:

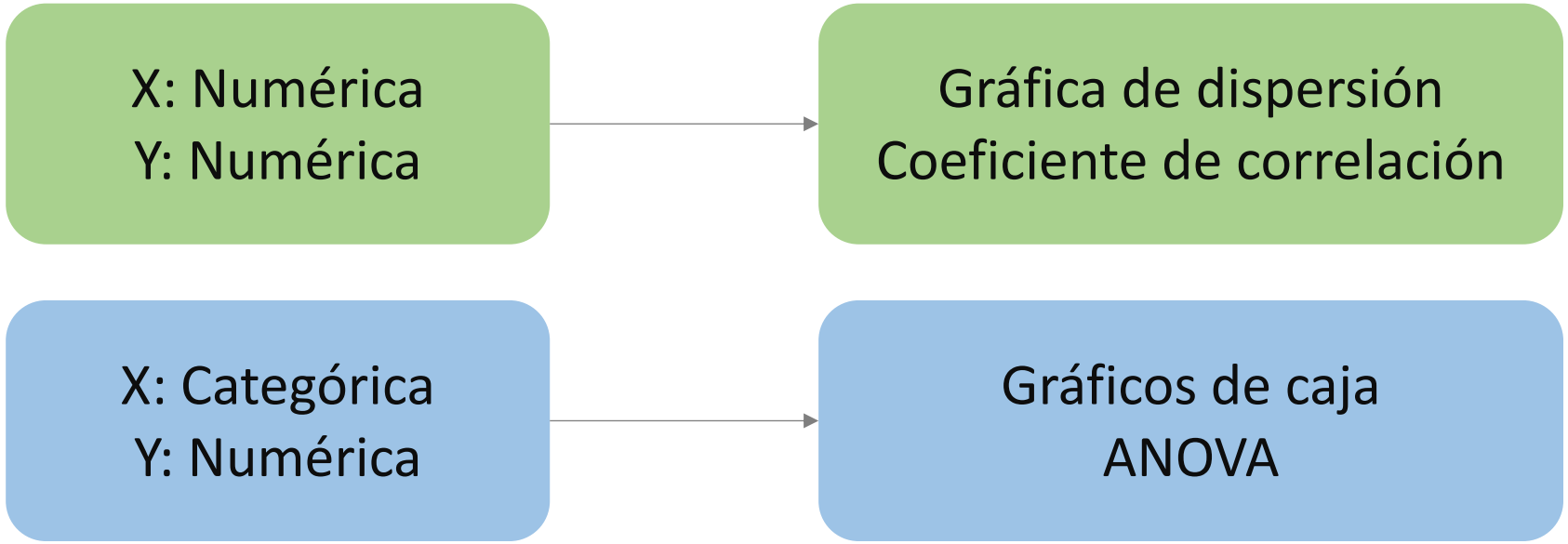
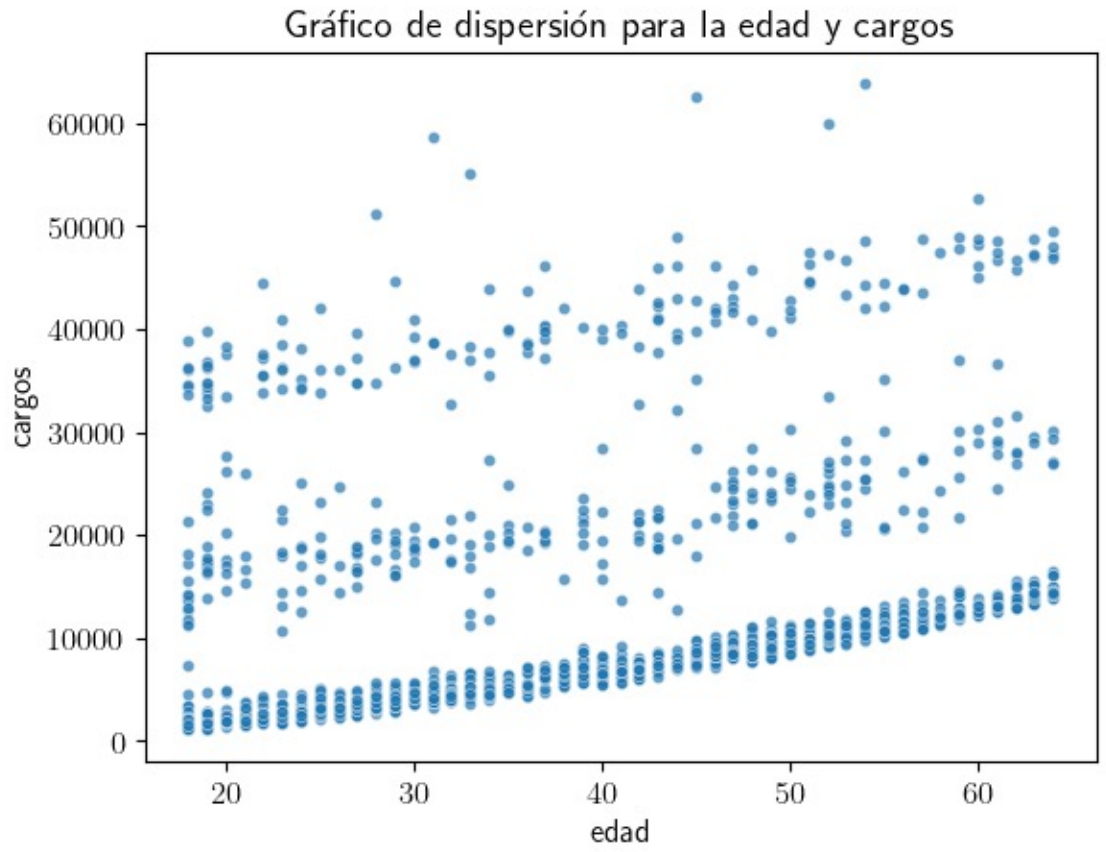


GRAFICO DE DISPERSIÓN

- El gráfico de dispersión se utiliza para analizar la asociación entre dos variables numéricas X-Y
- Cada punto en la gráfica son pares ordenados de las dos variables numéricas (x_1, y_1), (x_2, y_2), …,(x_n, y_n).

X	Y
edad	**cargos**
54	12105.32
31	6799.46
38	5488.26
63	13887.20
54	9850.43
...	...
27	4846.92
19	2719.28
19	2709.11
22	2643.27
51	9391.35

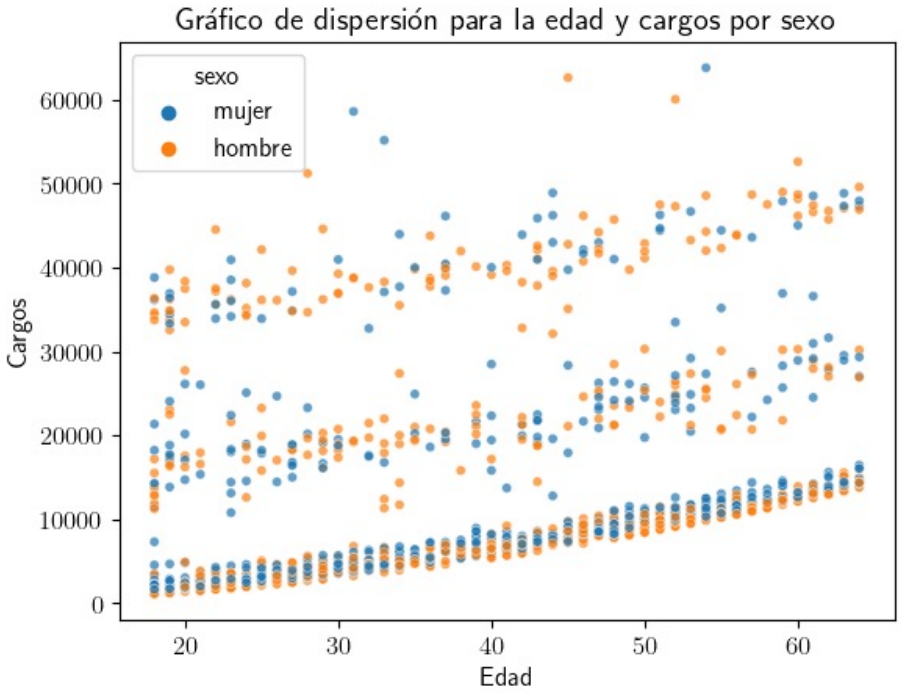


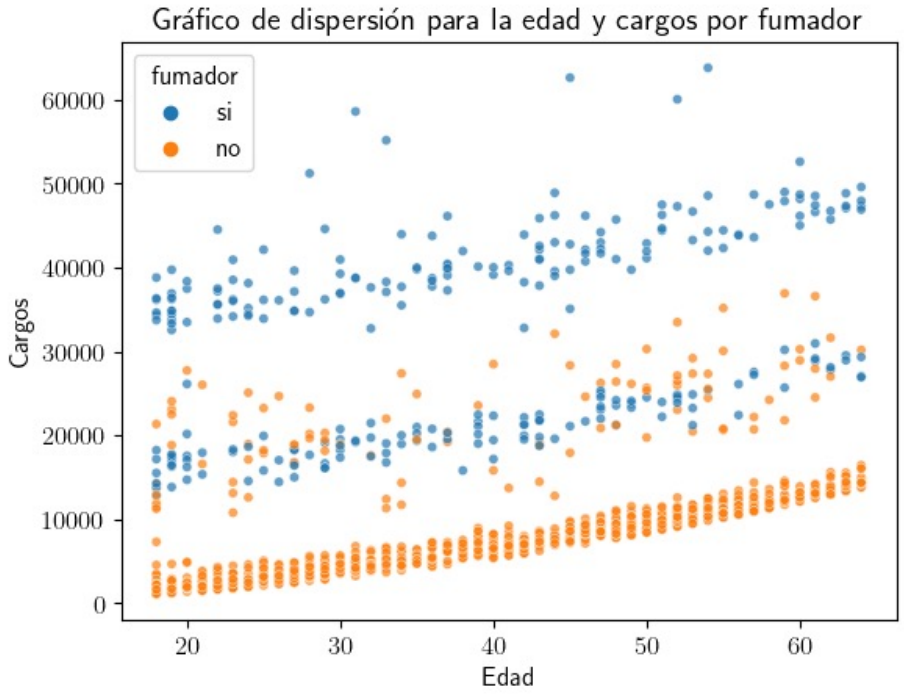
En la gráfica se observa que los pacientes de mayor edad tienden a tener mayores cargos.

¿Hay algún patrón particular que resalte en el gráfico?

INCLUSIÓN DE VARIABLES CATEGÓRICAS

- En la gráfica anterior se podía observar alguna relación entre la variable edad y cargos.
- Sin embargo, se observaron dos grupos poblacionales.
- Para identificar estos grupos se pueden incluir variables categóricas (sexo o fumador) al gráfico de dispersión.
- La variable fumador genera una mejor separación de los grupos observados en la gráfica anterior





COEFICIENTE DE CORRELACIÓN

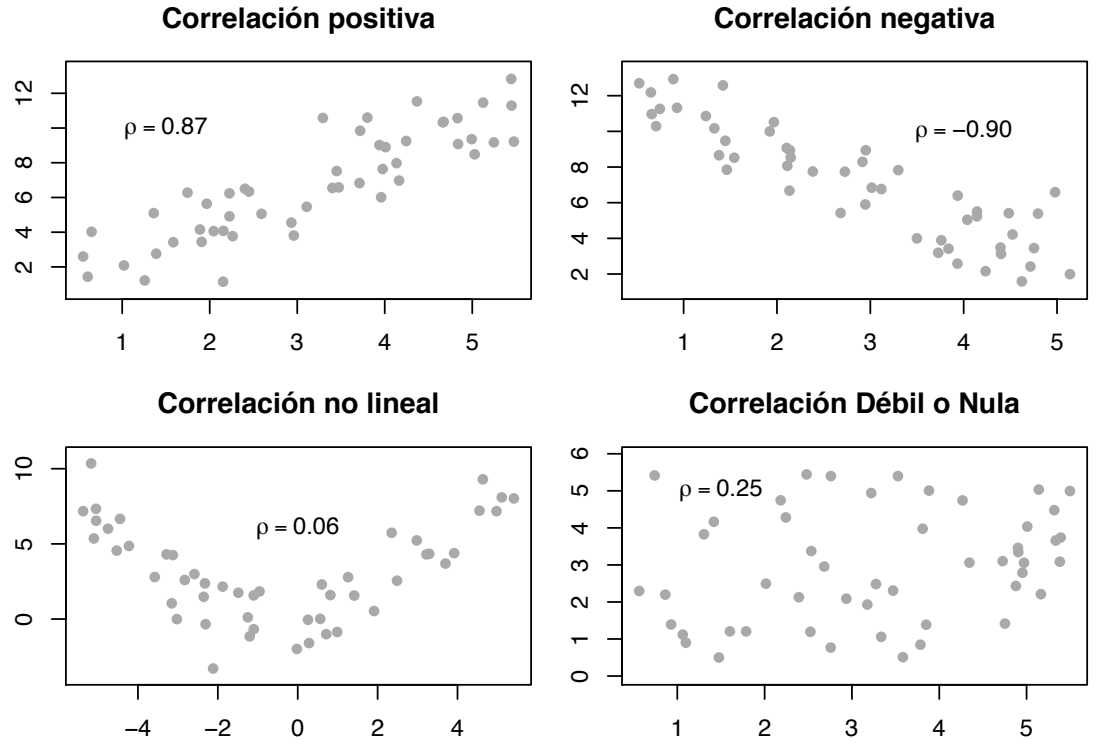
- El coeficiente de correlación permite medir la intensidad y el tipo de asociación lineal entre dos variables.
- El tipo de correlación se clasifica de la siguiente manera:
  - Correlación Lineal Positiva ($0.5 \leq \rho \leq 1$): Si el conjunto de puntos en el plano presentan un comportamiento ascendente .
  - Correlación Lineal Negativa ($-1 \leq \rho \leq -0.5$): Si el conjunto de punto el plano presenta un comportamiento descendente.
  - Correlación Débil o Nula ($-0.5 < \rho < 0.5$): Si no es posible identificar un com- portamiento en el conjunto de puntos.
  - Correlación No Lineal: Si es posible identificar diferentes tendencias en varias secciones del plano.

$$\rho = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\sum_{i=1}^{n} y_i)}{n}$$

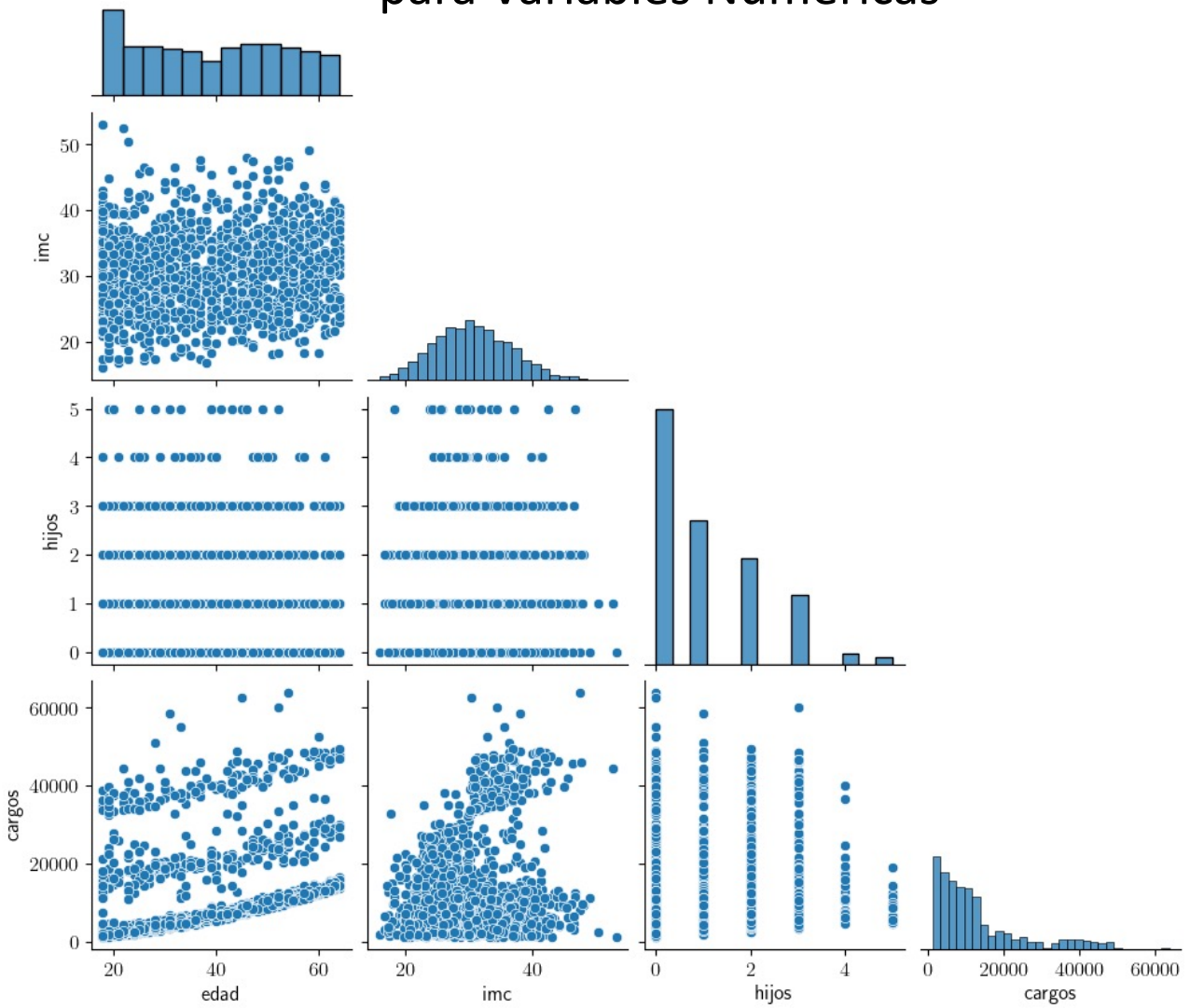


MATRIZ DE CORRELACIÓN

- En la tabla se presentan las correlaciones entre las variables numéricas del caso.
- Con la gráfica se pueden interpretar estas correlaciones.

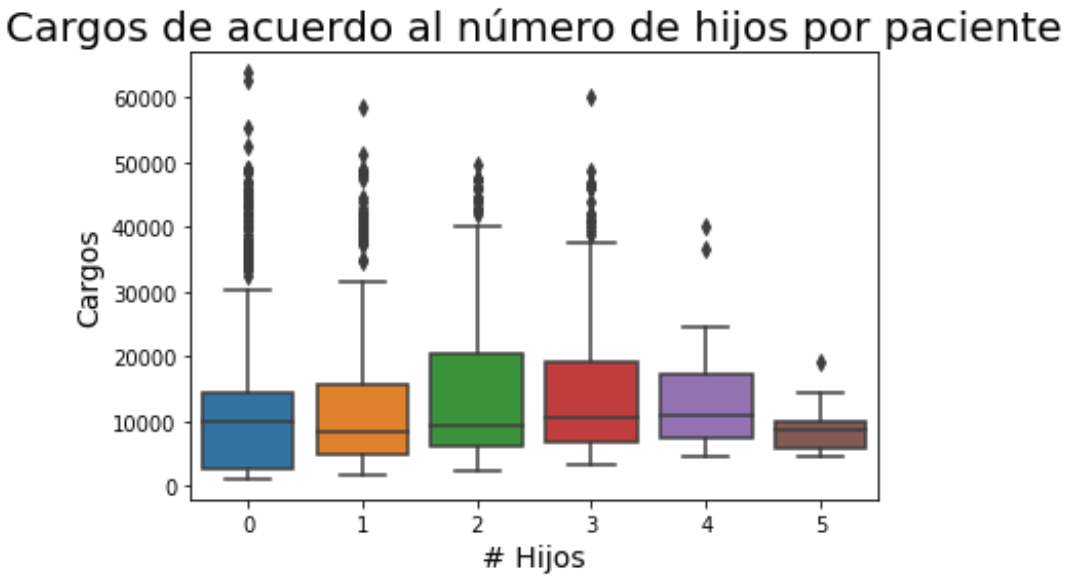
	Edad	IMC	Hijos	Cargos
Edad	1	0.109	0.042	0.299
IMC	0.109	1	0.013	0.198
Hijos	0.042	0.013	1	0.068
Cargos	0.299	0.198	0.068	1.000

Gráfica de Dispersión
para Variables Numéricas



GRÁFICOS COMPARATIVOS Y ESTADÍSTICAS POR GRUPO

- Si se quiere comparar la relación entre una variable X: categórica y una variable Y: numérica, se utilizan los gráficos de caja para comparar las distribuciones de la variable Y de acuerdo a cada categoría de la variable X.
- En el caso de los cargos médicos, se puede comparar los promedios de esta variable de acuerdo al número de hijos del paciente.
- A través de esta gráfica podemos analizar cómo cambian las tendencias centrales (mediana) y la dispersión (amplitud de las cajas y bigotes) de los cargos por número de hijos.
- Además, por utilizar las estadísticas por grupas para analizar la asociación entre dos variables (X: categórica, Y: Numérica)



# Hijos	Promedio	Mediana	Desviación Est.
0	12366.0	9857.0	12023.0
1	12731.0	8484.0	11824.0
2	15074.0	9265.0	12891.0
3	15355.0	10601.0	12331.0
4	13851.0	11034.0	9139.0
5	8786.0	8590.0	3808.0

ANOVA

- El ANOVA es un procedimiento estadístico que permite comparar las medias de diferentes grupos.
- El ejemplo de los cargos podemos confirmar estadística si existen diferencias significativas en los promedios de los cargos entre las diferentes regiones.
- Las hipótesis que se prueban son:

$$H_0: \mu_{grupo\ 1} = \mu_{grupo\ 2} = \cdots = \mu_k$$

$$H_1: \mu_{grupo\ i} = \mu_{grupo\ j}\ \forall\ i \neq j$$

Estimador de la varianza entre las medias de los grupos o categorías

Estadístico de Prueba F
Entre mas alto mayor diferencia entre las medias.

Es el valor P de la prueba. Cuando es menor a 0.05 la diferencia entre las medias es significativa.

	df	sum_sq	mean_sq	F	PR(>F)
C(hijos)	5.0	2.396917e+09	4.793834e+08	3.29692	0.005786
Residual	1332.0	1.936773e+11	1.454034e+08	NaN	NaN