Econometría I

Solución Workclass 03

Carlos A. Yanes Guerra

Universidad del Norte | Departamento de Economía

Contenido

Antes de empezar

Este tercer **workclass** puede ser entregado en grupo de máximo dos (2) estudiantes. En esta ocasión vamos a interactuar con una base de datos (ficticia). El tiempo para entrega es de *1 hora y 40 minutos*. El formato de recepción solo es en word o pdf en la carpeta del curso de **Brightspace**. No se admite por correo ni en formato distinto al solicitado.

Objetivo

Entender el manejo **interpretativo** de los datos. Familiarizarse con los paquetes de `tidyverse` y `moments` para hacer un análisis descriptivo de una base de datos.

Base de datos

Tiempo	Plan	Ingreso	Genero	Servicio
4312	43000	1840000	NB	Estudiante
5208	84000	1960000	NB	Premium
8311	42000	2320000	M	Estudiante
4712	42000	1810000	M	Estudiante
5645	84000	2890000	F	Premium
3765	62000	890000	F	Sencillo
2455	62000	920000	F	Sencillo
3664	62000	890000	NB	Sencillo
5535	91000	3490000	F	Premium
5212	92000	3650000	F	Premium
5245	94000	1290000	M	Premium
5141	92000	1790000	M	Premium
5425	92000	2890000	F	Premium
4312	42000	1850000	M	Estudiante
5711	42000	1810000	M	Estudiante
6642	44000	1980000	F	Estudiante
6842	44000	1000000	M	Estudiante
6712	45000	1850000	NB	Estudiante

La base de datos anterior contiene información un grupo de estudiantes de sus hábitos televisivos, se encuentran el número de horas que dedica en un mes a mirar cualquier contenido de entretenimiento, el valor del plan que paga, nivel de ingresos, género y por último el tipo de suscripción que tiene. Las últimas, son variables de tipo *binaria* y categórica que toman valores cualitativos.

Preguntas y solicitudes

De acuerdo a las respuestas situadas en la base, importe a R y/o Rstudio. *Si importa directamente y debe trabajar con este tipo de variables (cualitativas) la idea es convertirlas en un factor o etiqueta estipulada.* Esta etiqueta de cada variable cualitativa debe tomar un número o clasificación. Por ejemplo: (1=Premium, 2= Sencillo, 3= Estudiante). Las de genero puede codificarla con el orden que usted desee pero respetando la numeración de 0, 1 y 2 respectivamente para cada genero expuesto.

1. Piense por un segundo que desea agregar información de una persona de genero femenino, con un ingreso de dos millones seiscientos mil pesos m/lc, un plan sencillo, y que reporta utilizar en el mes alrededor de 6023 horas. Paga regularmente $54900. Utilice el operador pipe `%>%` y adicione la fila. Cree una nueva variable (para todxs) con la opción de `mutate`, que nos diga que porcentaje del ingreso se destina al pago del servicio. No olvide activar el paquete de `tidyverse` para eso. *Muestre todo el proceso*.

R./

- Primero se cargan los paquetes correspondientes a utilizar datos, algunos son `readxl` y `tidyverse` para inclusive transformar los datos que nos solicitan
- Esta vez lo que se hizo fue que manualmente se subieron los datos, esto es crear cada vector de ellos y finalmente con la opción de `add_row` se adiciona el nuevo valor solicitado. El resultado de esto nos da:

```
# Importamos la base de datos
library(readxl)
library(tidyverse)
library(moments)
library(ggplot2)

# Creación del data frame
data <- data.frame(
  Tiempo = c(4312, 5208, 8311, 4712, 5645, 3765, 2455,
             3664, 5535, 5212, 5245, 5141, 5425, 4312,
             5711, 6642, 6842, 6712),
  Plan = c(43000, 84000, 42000, 42000, 84000, 62000,
           62000, 62000, 91000, 92000, 94000, 92000,
           92000, 42000, 42000, 44000, 44000, 45000),
  Ingreso = c(1840000, 1960000, 2320000, 1810000, 2890000,
              890000, 920000, 890000, 3490000, 3650000,
              1290000, 1790000, 2890000, 1850000,
              1810000, 1980000, 1000000, 1850000),
  Genero = c("NB", "NB", "M", "M", "F", "F", "F",
             "NB", "F", "F", "M", "M", "F",
             "M", "M", "F", "M", "NB"),
  Servicio = c("Estudiante", "Premium", "Estudiante", "Estudiante", "Premium", "Sencillo", "Sencillo",
               "Sencillo", "Premium", "Premium",
               "Premium", "Premium", "Premium",
               "Estudiante", "Estudiante", "Estudiante", "Estudiante", "Estudiante")
)
# Añadimos los datos que hagan falta
data<-data %>%
  add_row(Tiempo=6023, Plan=54900, Ingreso= 2600000,
```

```
          Genero="F", Servicio= "Sencillo")
data
```

```
##    Tiempo  Plan Ingreso Genero   Servicio
## 1    4312 43000 1840000     NB Estudiante
## 2    5208 84000 1960000     NB    Premium
## 3    8311 42000 2320000      M Estudiante
## 4    4712 42000 1810000      M Estudiante
## 5    5645 84000 2890000      F    Premium
## 6    3765 62000  890000      F   Sencillo
## 7    2455 62000  920000      F   Sencillo
## 8    3664 62000  890000     NB   Sencillo
## 9    5535 91000 3490000      F    Premium
## 10   5212 92000 3650000      F    Premium
## 11   5245 94000 1290000      M    Premium
## 12   5141 92000 1790000      M    Premium
## 13   5425 92000 2890000      F    Premium
## 14   4312 42000 1850000      M Estudiante
## 15   5711 42000 1810000      M Estudiante
## 16   6642 44000 1980000      F Estudiante
## 17   6842 44000 1000000      M Estudiante
## 18   6712 45000 1850000     NB Estudiante
## 19   6023 54900 2600000      F   Sencillo
```

- Luego se procede a darle estructura a la base de datos de tal manera que se establezcan las etiquetas correspondiente a la orden de plan en el factor a utilizar. Eso se realiza de la siguiente manera:

```
dt<-data%>%
  transmute(Tiempo=as.numeric(Tiempo),
            Plan=as.numeric(Plan),
            Ingreso=as.numeric(Ingreso),
            Genero=as.factor(Genero),
            Servicio=factor(Servicio,labels = c("3=Estudiante","1=Premium","2=Sencillo")))
```

Para la parte de agregar una nueva variable que tiene el porcentaje de planes sobre el nivel de ingresos de los individuos, esto es:

```
dt<-dt %>%
  mutate(PartporIng=Plan/Ingreso*100)
dt
```

```
##    Tiempo  Plan Ingreso Genero     Servicio PartporIng
## 1    4312 43000 1840000     NB 3=Estudiante   2.336957
## 2    5208 84000 1960000     NB    1=Premium   4.285714
## 3    8311 42000 2320000      M 3=Estudiante   1.810345
## 4    4712 42000 1810000      M 3=Estudiante   2.320442
## 5    5645 84000 2890000      F    1=Premium   2.906574
## 6    3765 62000  890000      F   2=Sencillo   6.966292
## 7    2455 62000  920000      F   2=Sencillo   6.739130
## 8    3664 62000  890000     NB   2=Sencillo   6.966292
## 9    5535 91000 3490000      F    1=Premium   2.607450
## 10   5212 92000 3650000      F    1=Premium   2.520548
## 11   5245 94000 1290000      M    1=Premium   7.286822
## 12   5141 92000 1790000      M    1=Premium   5.139665
## 13   5425 92000 2890000      F    1=Premium   3.183391
## 14   4312 42000 1850000      M 3=Estudiante   2.270270
```

```
## 15   5711 42000 1810000      M 3=Estudiante   2.320442
## 16   6642 44000 1980000      F 3=Estudiante   2.222222
## 17   6842 44000 1000000      M 3=Estudiante   4.400000
## 18   6712 45000 1850000     NB 3=Estudiante   2.432432
## 19   6023 54900 2600000      F   2=Sencillo   2.111538
```

Ya con lo anterior tenemos parte de la míneria de datos ya resuelta para trabajar.

2. Presente una tabla con un resumen estadístico de las variables *Plan, Ingreso* y *Tiempo* en formato `tibble`. No solo plantee el código, interprete los resultados.

R./

- Una de las maneras de mostrar un resumen de datos es a partir de la función `summarise_each` seleccionando solo las variables cuantitativas o de formato double.

```
dt # es el nombre de la data que se trabaja
```

```
##    Tiempo  Plan Ingreso Genero     Servicio PartporIng
## 1    4312 43000 1840000     NB 3=Estudiante   2.336957
## 2    5208 84000 1960000     NB    1=Premium   4.285714
## 3    8311 42000 2320000      M 3=Estudiante   1.810345
## 4    4712 42000 1810000      M 3=Estudiante   2.320442
## 5    5645 84000 2890000      F    1=Premium   2.906574
## 6    3765 62000  890000      F   2=Sencillo   6.966292
## 7    2455 62000  920000      F   2=Sencillo   6.739130
## 8    3664 62000  890000     NB   2=Sencillo   6.966292
## 9    5535 91000 3490000      F    1=Premium   2.607450
## 10   5212 92000 3650000      F    1=Premium   2.520548
## 11   5245 94000 1290000      M    1=Premium   7.286822
## 12   5141 92000 1790000      M    1=Premium   5.139665
## 13   5425 92000 2890000      F    1=Premium   3.183391
## 14   4312 42000 1850000      M 3=Estudiante   2.270270
## 15   5711 42000 1810000      M 3=Estudiante   2.320442
## 16   6642 44000 1980000      F 3=Estudiante   2.222222
## 17   6842 44000 1000000      M 3=Estudiante   4.400000
## 18   6712 45000 1850000     NB 3=Estudiante   2.432432
## 19   6023 54900 2600000      F   2=Sencillo   2.111538
```

```
summarise_each(dt[1:3], funs(mean, var, min, max))
```

```
##   Tiempo_mean Plan_mean Ingreso_mean Tiempo_var  Plan_var  Ingreso_var
## 1    5309.053  63889.47      1985263    1754881 467420994 683815204678
##   Tiempo_min Plan_min Ingreso_min Tiempo_max Plan_max Ingreso_max
## 1       2455    42000      890000       8311    94000     3650000
```

- La interpretación data con los datos que nos brinda la salida:

La media de tiempo que se gastan en entretenimiento las personas ronda las 5309 horas en los últimos 5 años de tener o poseer una suscripción. eso es o se trata de alrededor unas 900 horas de entretenimiento al año siendo mensualmente un dato de 75 horas. La persona que menos ha gastado horas en ocio ha sido de unos 2455. en materia de ingresos, los individuos en promedio tienen casi dos salarios mínimos legales vigentes, quien menos se gana es de 890 mil pesos y se paga en promedio por canales de entretención 63 mil pesos con desviación de 21619 pesos por plan.

3. ¿Qué porcentaje de hombres, mujeres y no binarios tienen plan premium? ¿Considera que el nivel de ingreso tiene que ver con esa selección de plan?. ¿Qué tipo de servicio por ende es el mas usado entre todos los suscritos?. ¿Qué porcentaje de estudiantes de genero binario usan el tipo de servicio de estudiante? Pista

R./

Este lo resolvemos por partes

```
el1<-table(dt$Genero, dt$Servicio)
porcentaje_gen <- prop.table(el1)
porcentaje_gen
```

```
##
##      3=Estudiante  1=Premium 2=Sencillo
##   F    0.05263158 0.21052632 0.15789474
##   M    0.26315789 0.10526316 0.00000000
##   NB   0.10526316 0.05263158 0.05263158
```

El 21% de Mujeres o genero femenino tiene plan premium, tan solo el 10% de los masculinos y en un 5.2% los no-binarios

Solo un 10.5% de los No-binarios usan el servicio de estudiantes. El mayor uso con respecto a los demás servicios, puesto que todos los generos lo poseen en su hogar.

```
select(filter(dt, Ingreso > 2000000), Plan)
```

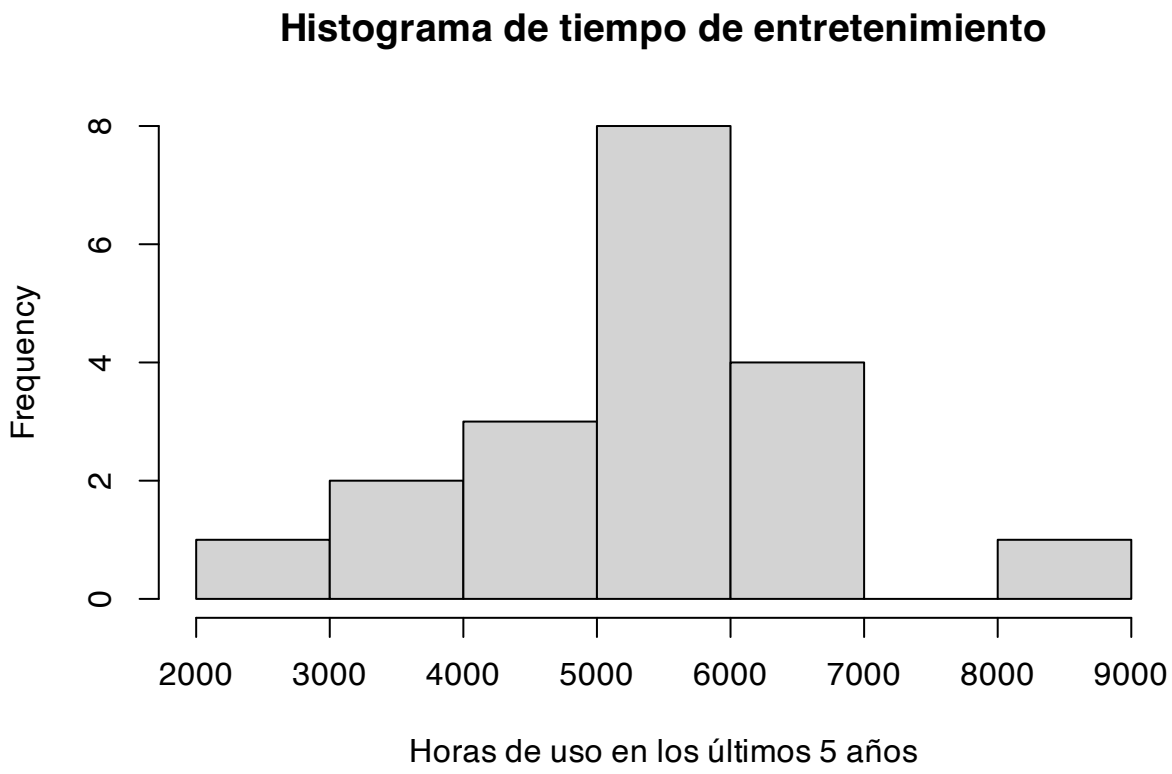
Nivel de ingreso y plan

```
##    Plan
## 1 42000
## 2 84000
## 3 91000
## 4 92000
## 5 92000
## 6 54900
```

Podemos decir que un mayor nivel de ingreso conlleva a que las personas demanden mas del servicio Premium que los demás niveles de ingreso.

4. ¿Qué tipo de distribución, curtosis y asimetría posee las horas de entretenimiento de las personas encuestadas? *Explique.*

```
hist(dt$Tiempo, main="Histograma de tiempo de entretenimiento", xlab="Horas de uso en los últimos 5 años")
```




```
###
library(moments)
kurtosis(dt$Tiempo) # La curtosis es
```

```
## [1] 3.342413
```

```
## Asímetria
skewness(dt$Tiempo)
```

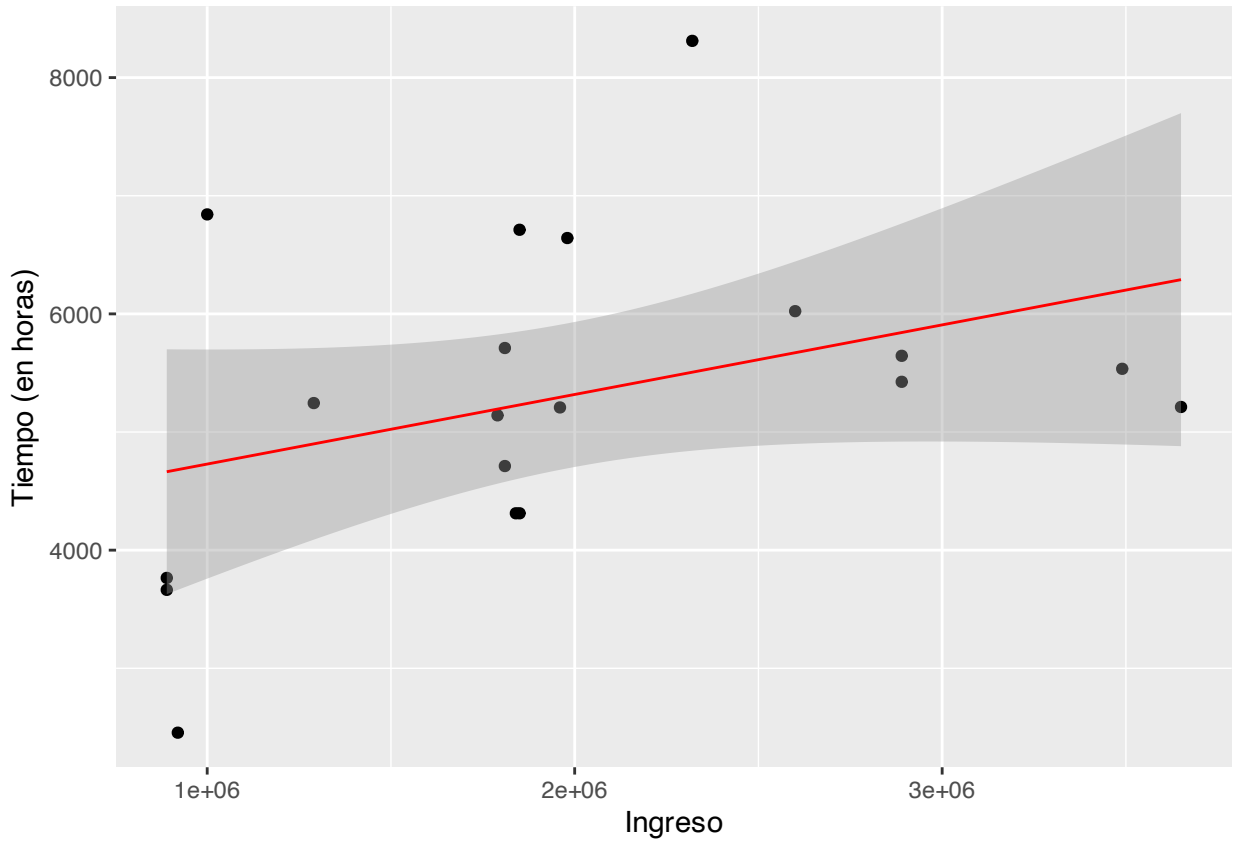
```
## [1] 0.06583239
```

La distribución es normal, mesocurtica y con un poco de simetria positiva. Al parecer los encuestados tienen uniformidad en el uso del tiempo de entretenimiento.

5. Desarrolle la gráfica de dispersión y linea de ajuste con el siguiente código base:

```
grafico<- dt %>%
 ggplot(aes(y = Tiempo, x = Ingreso))+
 geom_point() +
 ylab("Tiempo (en horas)")+xlab("Ingreso")+
 geom_smooth(method = 'lm', col = 'red', size = 0.5)

grafico # Salida de Gráfico
```



- Interprete el gráfico que observa entre las variables

R./

Esto nos quiere decir que existe una relación positiva entre el ingreso y las horas que se dedica al entretenimiento. Si bien hay dispersión por lo menos en linea con la teoría de demanda y los bienes normales, el entretenimiento aumenta cuando el ingreso tambien lo hace.

- Escriba la ecuación econométrica de regresión de lo anterior

R./

$$\text{tiempo} = \beta_0 + \beta_1 Ingreso_i + e_t$$

Si se estima, tenemos en regresión:

```
modelo<-lm(Tiempo ~ Ingreso, dt)
summary(modelo)
```

```
##
## Call:
## lm(formula = Tiempo ~ Ingreso, data = dt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2226.5  -905.2  -197.0   428.5  2804.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 4.140e+03  7.740e+02   5.348 5.32e-05 ***
## Ingreso     5.891e-04  3.613e-04   1.630    0.121
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1268 on 17 degrees of freedom
## Multiple R-squared:  0.1352, Adjusted R-squared:  0.08435
## F-statistic: 2.658 on 1 and 17 DF,  p-value: 0.1214
```