# PEC 5

**Paula Corbatón Álvarez**

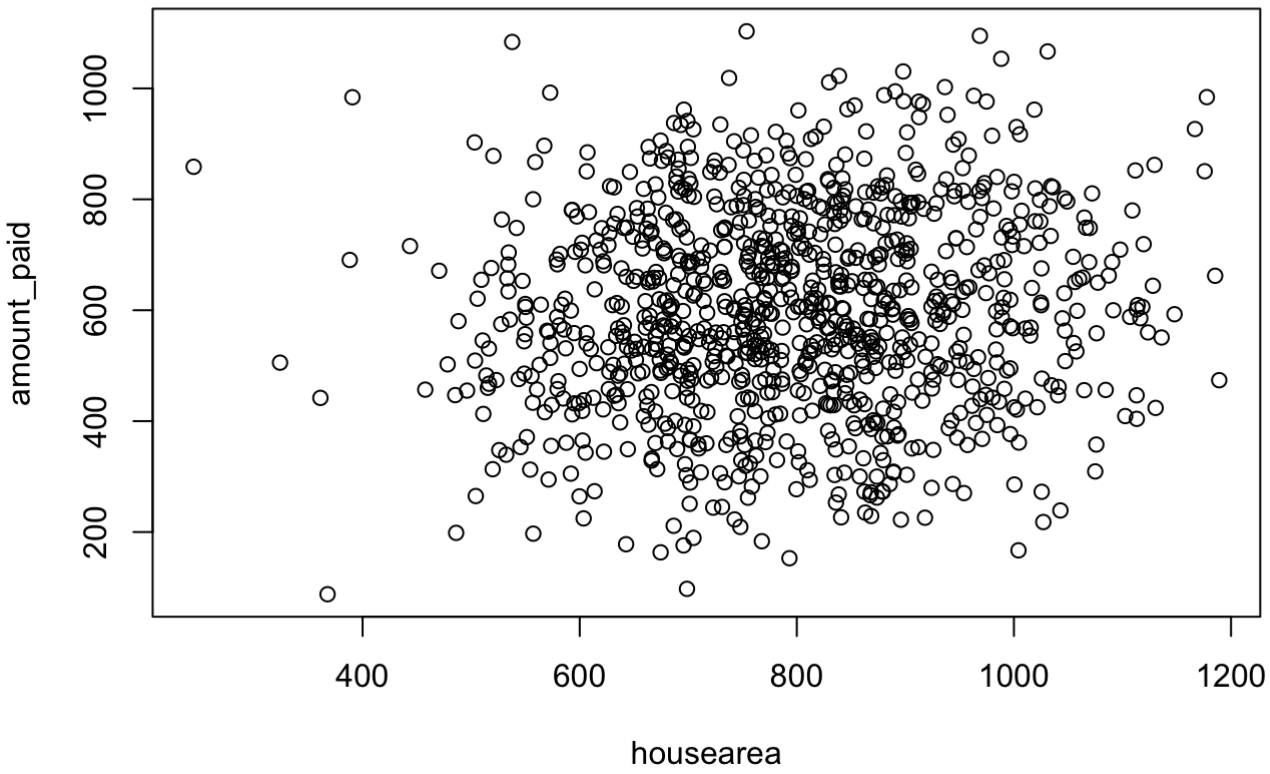
**2022-05-19**

# Pregunta 1

Mariluz está analizando cómo reducir su factura de la luz. Ha empezado a hacer algunas mejoras como cargar el móvil en el transporte público, pero no es suficiente. Se está planteando irse a vivir a una casa más pequeña, con menos habitaciones o fuera del centro urbano. Nos ha pedido ayuda para analizar como diferentes variables (como las comentadas) impactan en la factura de la luz.

a) Realiza un gráfico de dispersión entre la variable amount_paid y la variable housearea. A la vista de los resultados, ¿cree que existe una relación entre las variables?

```
plot(x=housearea, y =amount_paid)
```




A simple vista no parece que haya ninguna relación entre ambas variables.

b) Escribe la ecuación de la recta de regresión de la variable “amount_paid” en función de la variable “housearea”.

$$\begin{cases} X = housearea \\ Y = amountpaid \end{cases}$$

```
regresion_PaidArea <- lm(amount_paid~housearea)
summary(regresion_PaidArea)
```

```
##
## Call:
## lm(formula = amount_paid ~ housearea)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -491.54 -119.89   -5.72  130.50  513.66
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 506.58733   31.26493  16.203  < 2e-16 ***
## housearea     0.11804    0.03868   3.052  0.00233 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 180.7 on 998 degrees of freedom
## Multiple R-squared:  0.009246,   Adjusted R-squared:  0.008253
## F-statistic: 9.314 on 1 and 998 DF,  p-value: 0.002335
```

Por lo tanto la recta de regresion es y= 0.11804·x + 506.58733

c) ¿Cuál es el valor del coeficiente de determinación? ¿Se ajusta bien el modelo?
Como podemos observar en los cálculos anteriores, el coeficiente de determinación vale 0.009246 por lo tanto, el grado de ajuste de la recta de regresión a los valores de la muestra es muy pequeño. El modelo no se ajusta bien.

# Pregunta 2

Ahora Mariluz quiere determinar si hay diferencias significativas en los valores medios de la variable amount_paid según los niveles de la variable num_children.
Para realizar este ejercicio usa el siguiente comando antes, que permite transformar la variable num_children a una variable categórica de tipo factor:

```
data_pac5$num_children<- as.factor(data_pac5$num_children)
```

a) ¿Qué tipo de contraste de hipótesis (de los vistos en clase en este reto) se debería utilizar? Formula la hipótesis nula y alternativa que está pidiendo Mariluz.
Utilizaremos el análisis de la varianza (ANOVA)

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_1 : no\ todas\ las\ medias\ son\ iguales \end{cases}$$

b) Realiza el contraste en R y explica las conclusiones del mismo (para este ejercicio se debe tener en cuenta un nivel de significancia del 5%)

- Hipótesis nula: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$, es decir, todas las medias son iguales.
- Hipótesis alternativa: $H_1$: no todas las medias son iguales.

- Nivel de significancia: $\alpha = 0.05$

```
anova_PaidChildren <- aov(amount_paid~num_children)
summary(anova_PaidChildren)
```

```
##                Df   Sum Sq Mean Sq F value Pr(>F)
## num_children    1  6583848 6583848   249.9 <2e-16 ***
## Residuals     998 26291456   26344
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como podemos observar en la salida de R el p-valor obtenido es menor que 2e-16 (prácticamente 0) por lo que rechazamos $H_0$.
Puesto que rechazamos $H_0$ podemos afirmar que existen diferencias significativas entre las medias de la variable amount_paid según los niveles de la variable num_children.

c) Sólo con este análisis, ¿podríamos afirmar que hay diferencias significativas en la media de la variable amount_paid para los grupos num_children del nivel 2 y 3?
No. El análisis que acabamos de realizar nos indica que al menos un par de medias son significativamente distintas la una de la otra pero no podemos afirmar que concretamente haya diferencias en la media de la variable amount_paid para los grupos num_children del nivel 2 y 3.

# Pregunta 3

En este ejercicio vamos a desarrollar para Mariluz una estimación de la factura mensual (variable amount_paid) de la luz mediante un modelo de regresión múltiple con las siguiente variables:
• num_rooms
• ave_monthly_income
• is_urban
Para ello haga la siguiente transformación antes de empezar el ejercicio:

```
data_pac5$is_urban<- as.factor(data_pac5$is_urban)
```

a) Escribe la ecuación de la recta de regresión que explique la variable amount_paid según las variables: num_rooms, ave_monthly_income e is_urban.

- num_rooms = $x_1$
- ave_monthly_income = $x_2$
- is_urban = $x_3$

```
regresion_PaidRoomsIncomeUrban <- lm (amount_paid  ~ num_rooms + ave_monthly_income +
is_urban)
summary(regresion_PaidRoomsIncomeUrban)
```

```
##
## Call:
## lm(formula = amount_paid ~ num_rooms + ave_monthly_income + is_urban)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -355.82  -98.51   -2.76   94.11  422.79
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.342e+02  1.574e+01  27.592   <2e-16 ***
## num_rooms          -2.409e+00  4.232e+00  -0.569   0.5692
## ave_monthly_income  9.665e-04  4.508e-04   2.144   0.0323 *
## is_urban            2.419e+02  8.933e+00  27.081   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 137.8 on 996 degrees of freedom
## Multiple R-squared:  0.4248, Adjusted R-squared:  0.423
## F-statistic: 245.2 on 3 and 996 DF,  p-value: < 2.2e-16
```

$$y = 434.2 + -2.409 \cdot x_1 + 0.0009665 \cdot x_2 + 241.9 \cdot x_3$$

b) Analiza la significación individual de los parámetros del modelo estimado (considerando un nivel del 5%).

$$\begin{cases} H_0 : Bj = 0, por\ lo\ tanto,\ la\ variable\ x_j\ no\ es\ explicativa \\ H_1 : Bj \neq 0, por\ lo\ tanto,\ la\ variable\ x_j\ es\ explicativa \end{cases}$$

$\alpha = 0.05$

Utilizamos los resultados obtenidos en el apartado a):

• num_rooms:
- estadistico de contraste (t) = -0.569
- p-valor = 0.5692
Puesto que el p-valor es mayor que alfa, no rechazamos $H_0$, por lo tanto housearea no es una variable explicativa del modelo

• ave_monthly_income:
- estadistico de contraste (t) = 2.144
- p-valor = 0.0323
Puesto que el p-valor es menor que alfa rechazamos $H_0$, por lo tanto ave_monthly_income es una variable explicativa del modelo

• is_urban:
- estadistico de contraste (t) = 27.081
- p-valor: <2e-16 (parcticamente 0)
Puesto que el p-valor es menor que alfa rechazamos $H_0$, por lo tanto is_urban es una variable explicativa del modelo

c) Analiza el modelo estimado en conjunto (significación con un nivel del 5% y coeficiente de determinación).

$$\begin{cases} H_0 : B_1 = B_2 = B_3 = 0 \\ H_1 : B_1 \neq B_2 \neq B_3 \neq 0 \end{cases}$$

- $\alpha = 0.05$

Utilizamos los resultados obtenidos en el apartado a):

- estadístico F: 245.2
- p-value: < 2.2e-16

Puesto que el p-valor es prácticamente 0 podemos rechazar la hipótesis nula. Por lo tanto, podemos concluir que el modelo explica significativamente la variable y

d) Si Paco, un amigo de Mariluz, vive en un entorno rural con 7 habitaciones y tiene un salario de 2000€, ¿cuánto pagará de luz?
En el apartado a) concluimos que la ecuación de la recta de regresion para la variable amount_paid según num_rooms, ave_monthly_income y is_urban es:
$y = 434.2 - 2.409 \cdot numRooms + 0.0009665 \cdot aveMonthlyIncome + 241.9 \cdot isUrban$

Por lo tanto la estimación del precio que deberá pagar Paco es:
$y = 434.2 - 2.409 \cdot 7 + 0.0009665 \cdot 2000 + 241.9 \cdot 0 = 419.27$