Evaluación curso R

Barbod Aliaghazadeh

Entrega

- Recuerda completar tu **nombre en el apartado author**.
- Además del código, no olvides completar las respuestas a las preguntas indicadas en negrita.
- Si tienes dudas/dificultades, **puedes contactar con los profesores**.
- Deadline: **viernes 19 de junio, 23:59**.
- Puedes realizar la entrega en el mail: **constantino.garciama@ceu.es**.
- La entrega consistirá en el fichero que se genera al hacer `Knit` (un fichero `html` o `pdf`).

Apuestas de adolescentes en UK

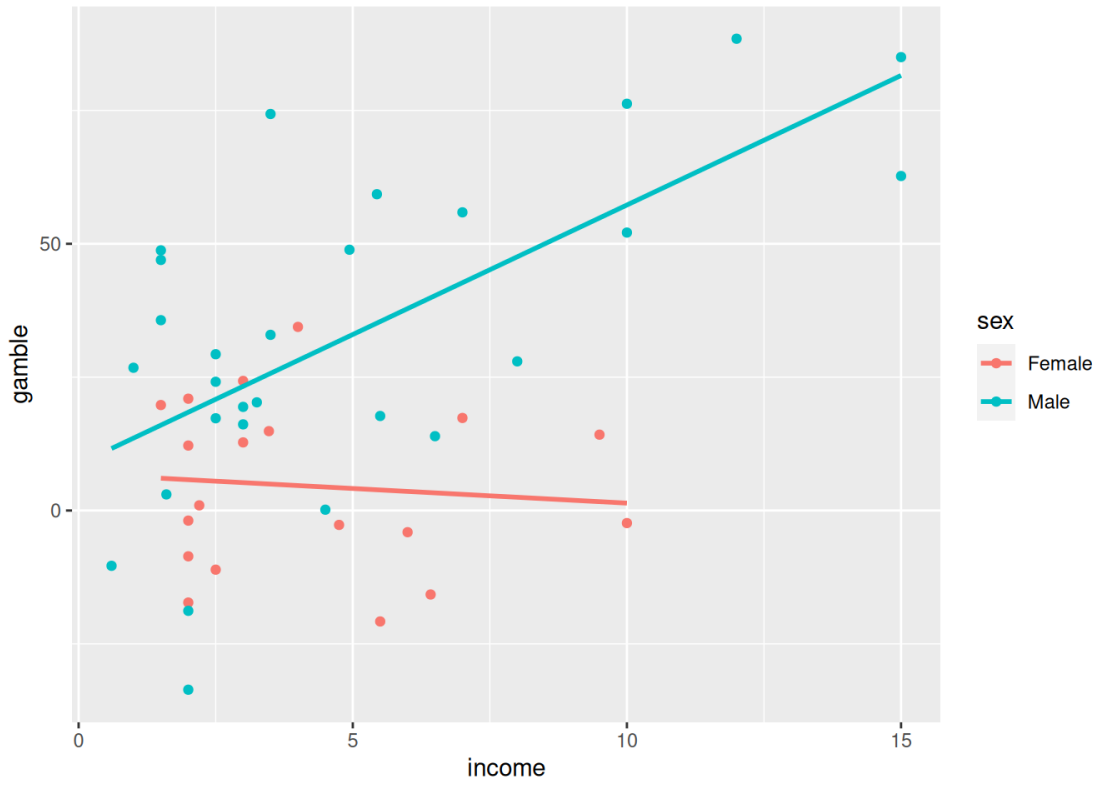
El conjunto de datos `teengamb.csv` contiene datos sobre las tasas de juego entre los adolescentes en Gran Bretaña, su género y estatus socioeconómico. Una pregunta que nos puede interesar es **si los ingresos del adolescente y su sexo influyen en la cantidad de dinero apostado (céntrate solo en las variables `income`, `sex` y `gamble`)** ... Sigue los siguientes pasos para crear un ANCOVA...

1a) Carga los datos...

```
# Carga los datos
teengamb <- read.csv("teengamb.csv")
```

1b) Visualiza los datos...

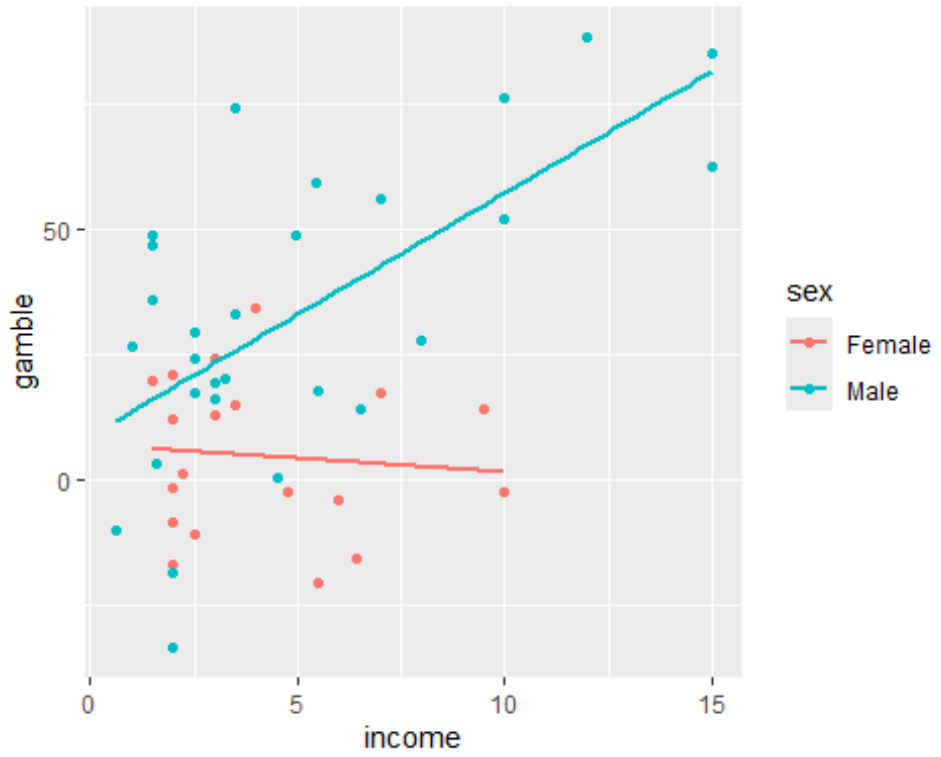
...para valorar si el modelo debe incluir interacciones. Para ello, escribe el código que genera una gráfica similar a la siguiente:




```r
library(ggplot2)


teengamb$sex <- as.factor(teengamb$sex)


ggplot(teengamb, aes(x = income, y = gamble, color = sex)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



La gráfica anterior apoya que hay interacciones entre las variables `sex` e `income`. **Explica los motivos brevemente** (Pista: ¿son las rectas paralelas?)

2) Crea el modelo...

... y obtén los intervalos de confianza para los coeficientes y su significación.

```
# Crea un modelo con interacciones en base a tu conclusión del apartado anterior
gamb_model <- lm(gamble ~ sex + income + sex:income, data = teengamb)


# Obtén p-valores e intervalos de confianza. Usa summary y confint
summary(gamb_model)

##
## Call:
## lm(formula = gamble ~ sex + income + sex:income, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.025 -15.479  -3.559  14.022  48.625
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.8743     9.2446   0.744   0.4612
## sexMale          1.8427    11.1946   0.165   0.8700
```

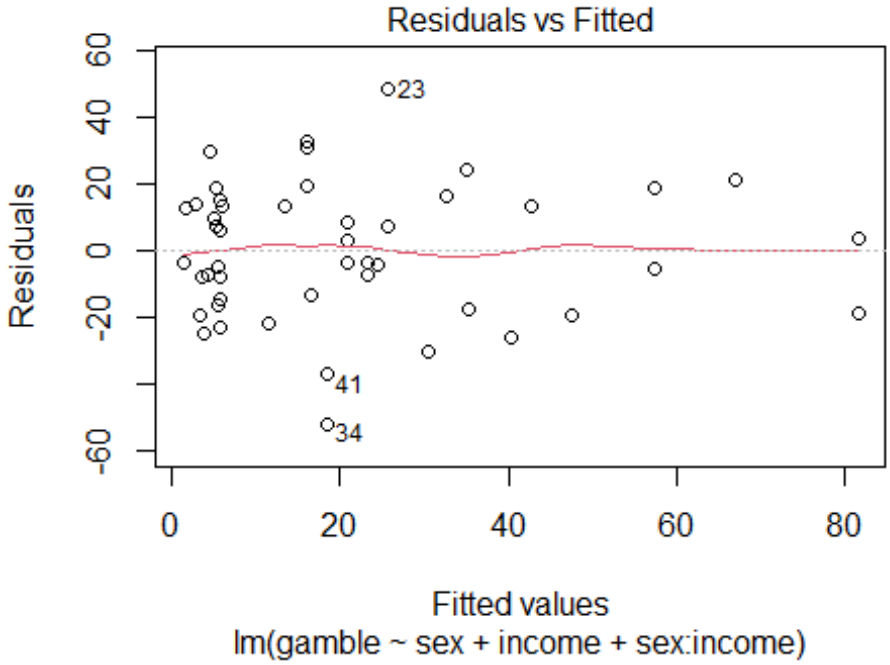
```
## income          -0.5489     1.9024  -0.289   0.7743
## sexMale:income   5.4050     2.1435   2.522   0.0155 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.97 on 43 degrees of freedom
## Multiple R-squared:  0.5105, Adjusted R-squared:  0.4764
## F-statistic: 14.95 on 3 and 43 DF,  p-value: 8.279e-07
```

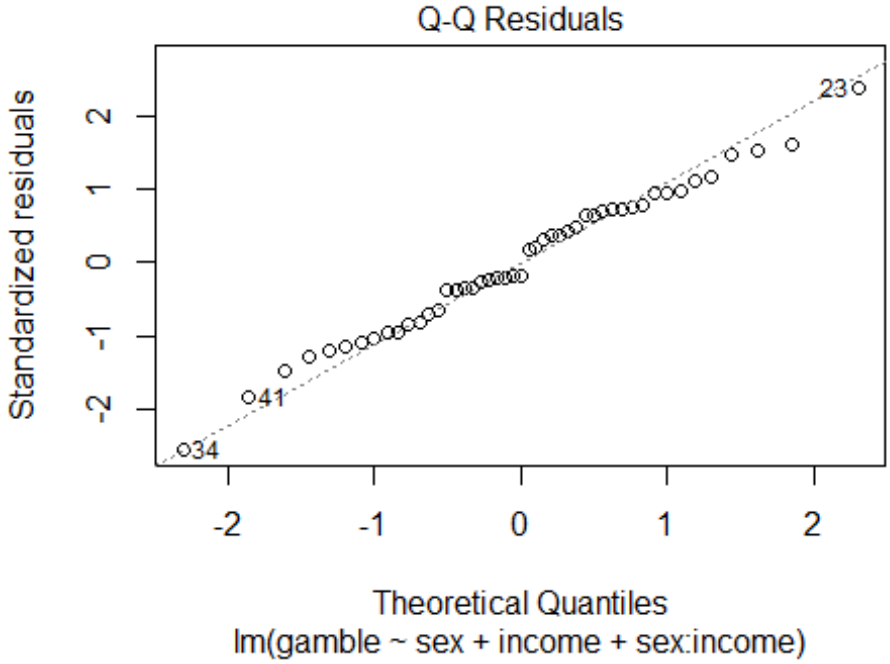
```
confint(gamb_model)
```

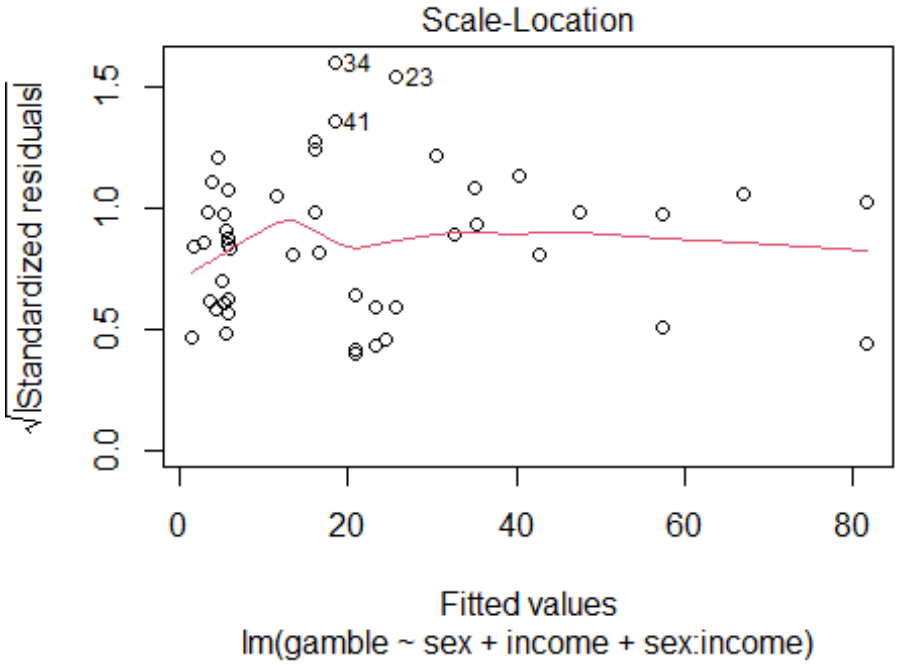
```
##                     2.5 %    97.5 %
## (Intercept)    -11.769152 25.517839
## sexMale        -20.733425 24.418773
## income          -4.385455  3.287728
## sexMale:income   1.082260  9.727751
```

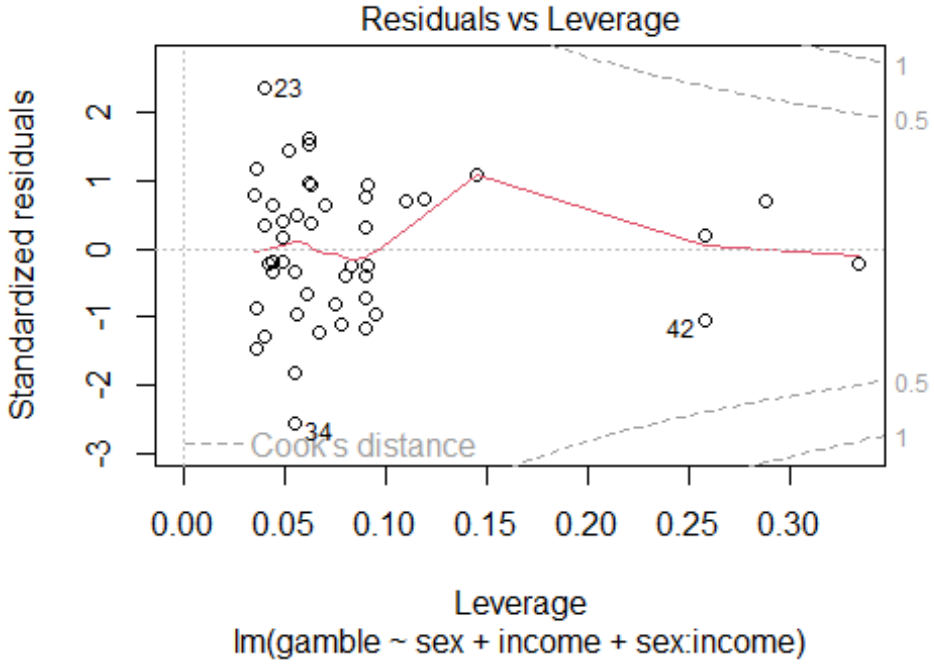
3) Valora si las asunciones del modelo se cumplen

```
# Interpretar las 4 gráficas siguientes para decidir si las asunciones del
modelo
# se cumplen
plot(gamb_model, ask = FALSE)
```


Residuals vs Fitted
Residuals
Fitted values
lm(gamble ~ sex + income + sex:income)
23
41
34


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(gamble ~ sex + income + sex:income)
23
41
34





¿Es el modelo correcto?:

#residuals vs fitted Pattern: The red line indicates a mostly flat line with some fluctuations at both extremes, which may indicate some non-linear behaviors or boundary conditions.

Scatter: The points do not seem to be spreading out more or less, which is desirable because it points towards homoscedasticity. However, the slight dip at the ends could be of interest and could require a closer look, such as checking the model or the variables for transformation. Outliers: Several of the points such as 23,34,41 seem to be somewhat off the main body of the data, which can be suggestive of outliers or influential observations.

#Normal q-q plot Center Alignment: The ‘‘good’’ alignment of the normal probability plot in the middle of the distribution indicates that the middle 50% of the data follow a normal distribution. Tail Behavior: The fluctuations at the upper and lower end of the line indicate with points 34,41 and 23 the presence of possible outliers or influential cases that may distort the model’s assumptions. These are distortions that can affect the credibility of regression outcomes especially the parameters and the tests of significance.

#Scale Location plot Line Trend: The red line in the Scale-Location plot, which should be a straight horizontal line to show homoscedasticity, has a slight tendency towards curvilinearity. This suggests that the spread of the residuals is not constant as the fitted values move across the range of values. Residual Spread: The points seem to be scattered irregularly and look like they are more spread out in the high fitted values than in the low fitted values. This indicates a positive relationship between the variance of the residual and the fitted values, a condition referred to as heteroscedasticity. Homoscedasticity Assumption: The first evidence of a violation of the homoscedasticity assumption is the observed trends in the red line and the increasing spread of residuals. This variance inhomogeneity may lead to inconsistency in the model standard errors, which in turn affects the accuracy and effectiveness of confidence intervals and significance tests that are obtained from the model.

#Residuals vs leverage Leverage Effect: Those with higher values of the leverage are easily observed in the plot indicating that they have the capability of exerting a greater influence on the coefficients of the regression model. These points are closer to the average of the response variable but are farther away from most of the other points in terms of the predictors. Cook’s Distance: The horizontal lines on the top of the plot represent Cook distances, which quantify the effect of each observation on the fitted values for the whole model. Observations that go beyond these lines are regarded as highly influential. Observations of Concern: Especially, the points as 42 and other points near the lines of leverage are the potential outliers. Their positions imply that they might be distorting the model’s outputs and the general figures to a large extent. Potential Impact on Model: These influential points may distort the model estimates and generate wrong results if not checked or corrected for. This could especially affect the performance of predictions and other logical inferences made within the data.

4) Interpreta los coeficientes y escribe tus conclusiones.

De la tabla de p-valores, podemos concluir que para las mujeres:

$$gambling = 6.87 - 0.5489 * income$$

Mientras que para hombres:

$$gambling = 8.72 + 4.8561 * income.$$

Fíjate que el coeficiente -0.5489 no es significativo, mientras que el salto en las pendientes entre mujeres y hombres es de 5.4050 y sí es significativo.

¿Cuáles de las siguientes conclusiones son correctas? (puede haber varias)

Los hombres y mujeres adolescentes apuestan de la misma forma. No evidencia suficiente de que las mujeres apuesten de forma diferente según sus ingresos. Para los hombres: a mayor nivel de ingresos, mayor cantidad apostada. La diferencia entre las pendientes de hombres y mujeres no es significativa.

a) Incorrect. The coefficients for men and women are significantly different in terms of size and signs, which mean that men and women are not similar in gambling.

b) Correct. In the equation for the women, the coefficient for income is not statistically significant, meaning that there is not enough evidence to suggest that the income influences how women gamble.

c) Correct. However, for men, the coefficient of income (+4. 8561) is positive and statistically significant, indicating that the amount gambled also rises with income.

d) Incorrect. The fact that the slopes increased by 5. 4050 between males and females indicates that the overall effect of income on gambling differs by gender with a significant difference between the groups.