PEC6 Primavera 2023. Solución

UOC

La PEC se basará en el fichero "PIS_MUN.cvs" donde podemos encontrar información sobre el nombre de contratos de alquiler y el alquiler medio (media de los precios de alquileres por contrato) a los municipios de Catalunya y a los distritos de la ciudad de Barcelona de los años 2015, 2016, 2017 y 2018.

Contiene, entre otras, las variables

- *MUN* = Nombre del municipio
- *PROV* = Provincia del municipio
- *N201X* = Nombre de contratos hechos el año 201X, donde X puede ser 5, 6, 7 o 8.
- *M201X* = Media de los precios de los alquileres por contrato de contratos hechos el año 201X, donde X puede ser 5, 6, 7 o 8.

Tenéis que importar los datos y guardarlos con el nombre de dadespis. Por ejemplo con el comando

```
dadespis<-read.table("PIS_MUN.csv", header=TRUE,
          sep=";",na.strings="NA",
          fileEncoding = "UTF-8", quote = "\"")
```

Os puede resultar útil consultar el seguiente material:

1. Módulo 10 Regresión lineal simple de las notas de estudio
2. Tema 1 de Regresión lineal del módulo 5 de los Manuales de R
3. Actividades Resueltas del Reto 5 (Regresión lineal)

NOM: Solución

PAC6

Pregunta 1 (100%)

a) (10%) Utilizando la instrucción "subset" cread un subconjunto del conjunto "dadespis" que sólo contenga los datos de los municipios que todos los años han hecho 500 o más contratos. A este subconjunto ponedle el nombre "dadespis_10". Imprimid las tres primeras líneas.

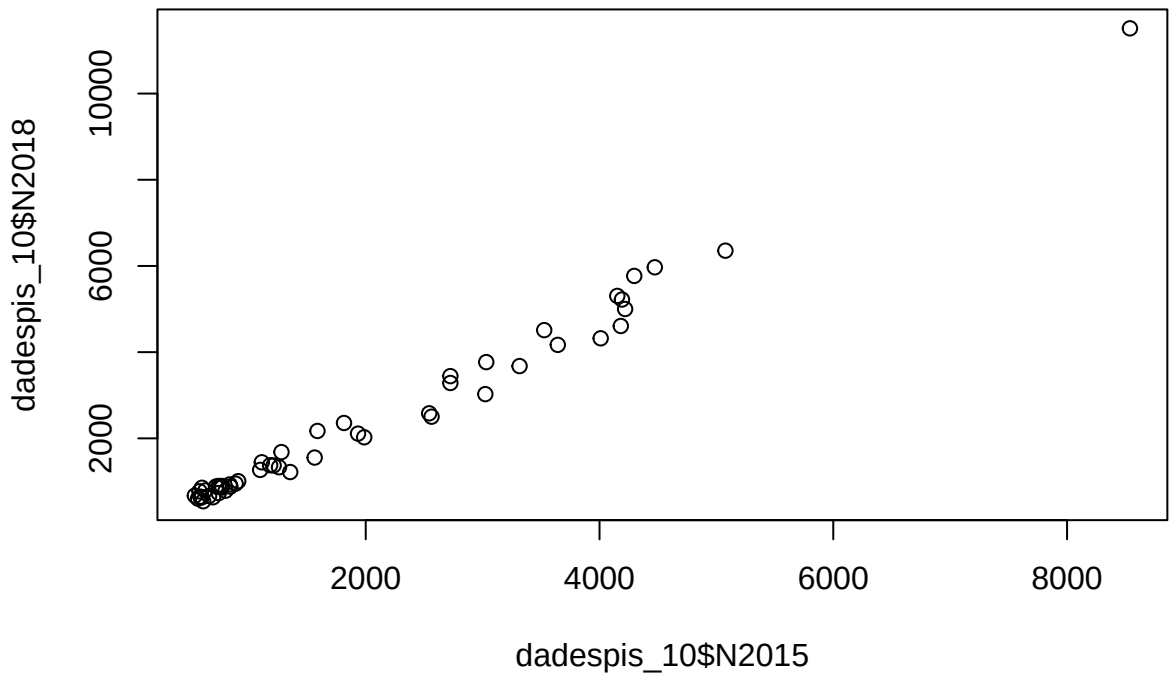
```
dadespis_10<-subset(dadespis, N2015>499 & N2016>499 & N2017>499  & N2018>499)
head(dadespis_10,3)
```

```
##       COD  MUN.DIS      MUN      PROV N2018 N2017 N2016 N2015  M2018  M2017
## 79   8015 Badalona Badalona Barcelona  3769  3352  3133  3031 688,56 638,24
## 124 17023   Blanes   Blanes    Girona   949   882   904   887  494,5 451,97
## 155 43037 Calafell Calafell Tarragona   671   560   563   538 541,31 492,32
##      M2016  M2015
## 79  580,08 549,16
## 124 430,78 424,56
## 155 465,35 449,17
```

A partir del conjunto de datos "dadespis_10", contestad las preguntas seguientes:

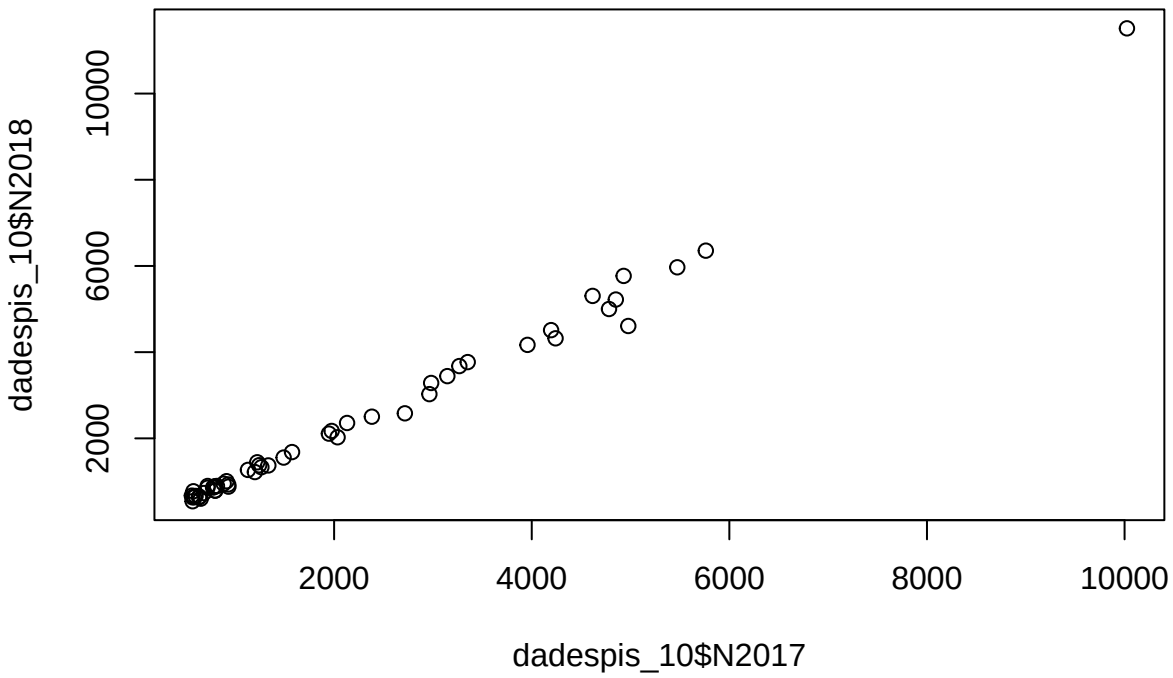
b) (20%) Realizad el diagrama de dispersión de las variables correspondientes a la variable "N2018" (en el eje de ordenadas) en función de la variable "N2017" (en el eje de abscisas). Realizad el diagrama de dispersión de las variables correspondientes a la variable "N2018" (en el eje de ordenadas) en función de la variable "N2015" (en el eje de abscisas). Comentad brevemente si hay diferencias.

```
plot(dadespis_10$N2015,dadespis_10$N2018)
```



Observamos que, en general, parece que a másss contratos a N2015, más contratos a N2018. Parece que hay una fuerte relación lineal.

```
plot(dadespis_10$N2017,dadespis_10$N2018)
```



Observamos que, en general, no se ven grandes diferencias.

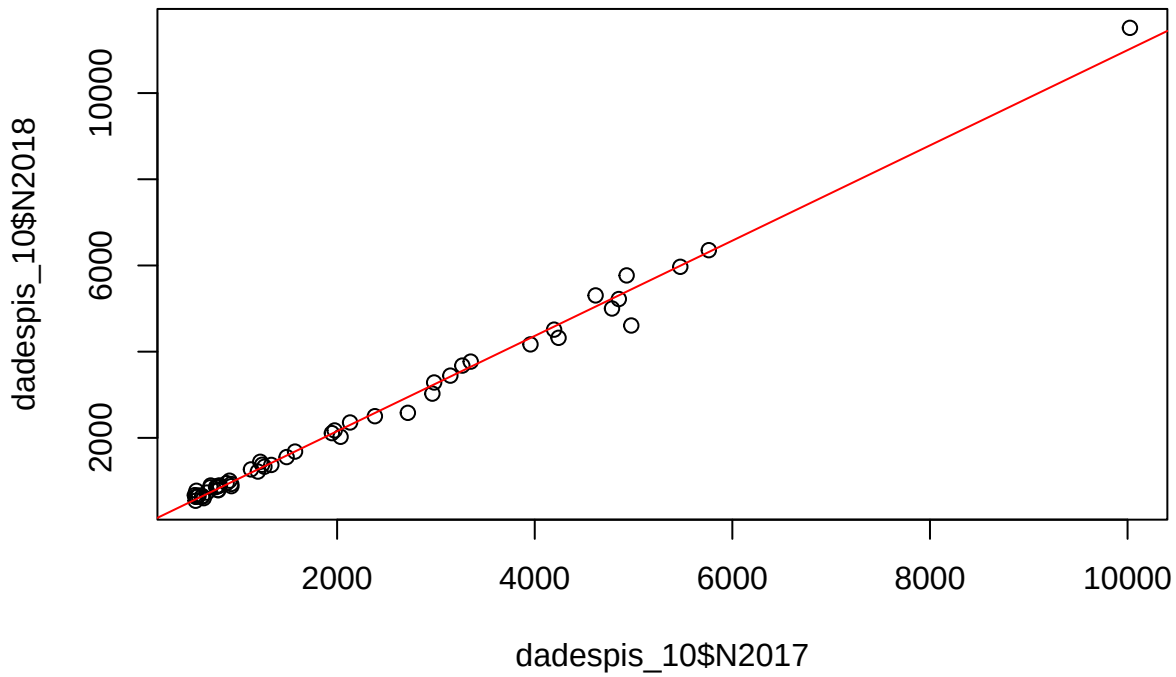
c) (30%) Calculad con R la recta de regresión de la variable "N2018" en función de la variable "N2017" y la recta de regresión de la variable "N2018" en función de la variable "N2015". Realizad los diagramas de dispersión, añadiendo las rectas de regresión. Dad explícitamente las rectas de regresión.

```
summary(lm(dadespis_10$N2018~dadespis_10$N2017))
```

```
##
## Call:
## lm(formula = dadespis_10$N2018 ~ dadespis_10$N2017)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -838.02  -47.66   26.22   61.27  489.98
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)       -55.61420   39.24308  -1.417    0.163
## dadespis_10$N2017   1.10521    0.01355  81.571   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 186.9 on 50 degrees of freedom
## Multiple R-squared:  0.9925, Adjusted R-squared:  0.9924
## F-statistic:  6654 on 1 and 50 DF,  p-value: < 2.2e-16
```

```
plot(dadespis_10$N2017,dadespis_10$N2018)
abline(lm(dadespis_10$N2018~dadespis_10$N2017),col="red")
```



Así, la recta de regresión es:

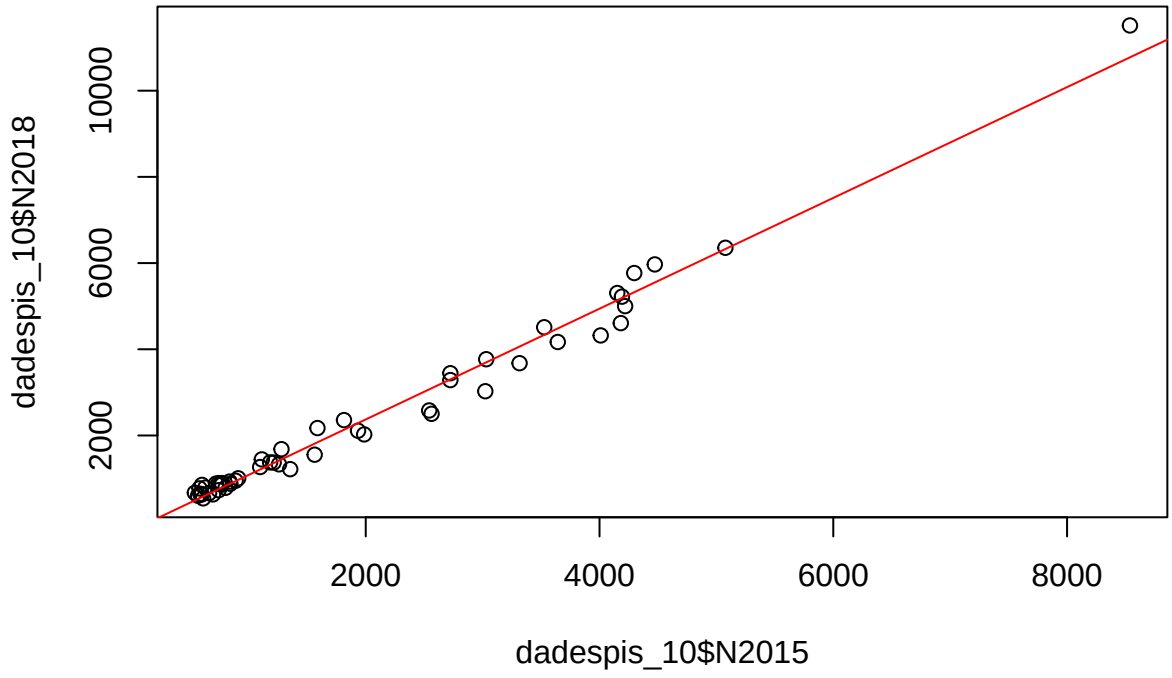
$$N2018 = -55.6142 + 1.1052 \cdot N2017$$

```
summary(lm(dadespis_10$N2018~dadespis_10$N2015))
```

```
##
## Call:
## lm(formula = dadespis_10$N2018 ~ dadespis_10$N2015)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -660.29  -66.28   55.54  160.14  739.38
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -198.56543   61.15928  -3.247  0.00209 **
## dadespis_10$N2015    1.28510    0.02379  54.013  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 281 on 50 degrees of freedom
## Multiple R-squared:  0.9832, Adjusted R-squared:  0.9828
## F-statistic:  2917 on 1 and 50 DF,  p-value: < 2.2e-16
```

```
plot(dadespis_10$N2015,dadespis_10$N2018)
abline(lm(dadespis_10$N2018~dadespis_10$N2015),col="red")
```



Así, la recta de regresión es:

$$N2018 = -198.5654 + 1.2851 \cdot N2015$$

d) (20 %) ¿Qué variabilidad de la "N2018" queda explicada por las variables "N2017" y "N2015" en los dos casos anteriores? ¿Qué podemos decir sobre la bondad del ajuste? ¿Qué recta es mejor?

Para la variable "N2017" el coeficiente de determinación es $R^2 = 0.9925$ que nos dice que el modelo explica un 99.25% de la variabilidad de "N2018". Es un buen ajuste.

Por la variable "N2015" el coeficiente de determinación es $R^2 = 0.9832$ que nos dice que el modelo explica un 98.32% de la variabilidad de "N2018". Es un buen ajuste.

Ambos ajustes son muy buenos.

e) (20 %) Queremos realizar contrastes de hipótesis con un nivel de significación del 0.01 sobre los coeficientes de las dos rectas de regresión que hemos obtenido. ¿Existe algún coeficiente no significativo? Razonad la respuesta.

En la regresión con "N2017" si consideramos

Hipótesis nula: $H_0 : \beta_0 = 0$, hipótesis alternativa: $H_1 : \beta_0 \neq 0$

tenemos un p-valor de 0.163, mayor que 0.01. Por tanto, aceptamos la hipótesis nula y podemos decir que β_0 no es significativo.

En la regressión con "N2015" todos los coeficientes son significativos.